# Incentive-Aware Dynamic Resource Allocation under Long-Term Cost Constraints

**Yan Dai**
Operations Research Center
MIT
yandai20@mit.edu

**Negin Golrezaei**
Sloan School of Management
MIT
golrezae@mit.edu

**Patrick Jaillet**
Department of EECS
MIT
jaillet@mit.edu

## Abstract

Motivated by applications such as cloud platforms allocating GPUs to users or governments deploying mobile health units across competing regions, we study the constrained dynamic allocation of a reusable resource to a group of strategic agents. Our objective is to simultaneously *(i)* maximize social welfare, *(ii)* satisfy multi-dimensional long-term cost constraints, and *(iii)* incentivize truthful reporting. We begin by numerically evaluating primal-dual methods widely used in constrained online optimization and find them to be highly fragile in strategic settings – agents can easily manipulate their reports to distort future dual updates for future gain. To address this vulnerability, we develop an incentive-aware framework that makes primal-dual methods robust to strategic behavior. Our primal-side design combines epoch-based lazy updates – discouraging agents from distorting dual updates – with dual-adjust pricing and randomized exploration techniques that extract approximately truthful signals for learning. On the dual side, we design a novel online learning subroutine to resolve a circular dependency between actions and predictions; this makes our mechanism achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ social welfare regret (where $T$ is the number of allocation rounds), satisfies all cost constraints, and ensures incentive alignment. This $\widetilde{\mathcal{O}}(\sqrt{T})$ performance matches that of non-strategic allocation approaches while additionally exhibiting robustness to strategic agents.

## 1 Introduction

Modern platforms and public agencies often face the challenge of allocating limited, reusable resources over time to self-interested agents, who may hide their true desire in sake of more favorable allocations. For example, cloud providers must decide how to distribute scarce GPUs to competing jobs under compute and energy constraints (Buyya et al., 2008; Nejad et al., 2014). Governments may deploy mobile health units or medical devices such as ventilators across regions, where needs vary over time and access is constrained by staffing or capacity (Stummer et al., 2004; Adan et al., 2009). In these settings, a same unit of resource is reallocated across time, and the allocation must respect some multi-dimensional long-term cost constraints – such as energy or staffing – while accounting for the strategic behavior of agents with private information.

A central challenge in these dynamic environments is being both *efficient* – *i.e.*, maximizing social welfare subject to constraints – and *robust* to agents' strategic manipulation. Focusing on efficiency, primal-dual methods serve as a powerful tool in online resource allocation, offering principled ways to handle constraints while adapting to changing demand (Devanur and Hayes, 2009; Golrezaei et al., 2014; Molinaro and Ravi, 2014; Balseiro et al., 2023). These methods maintain dual variables that act as shadow prices on resource usage, guiding allocations based on both values and costs. However, these approaches typically assume truthful agents and ignore any strategic responses agents make.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

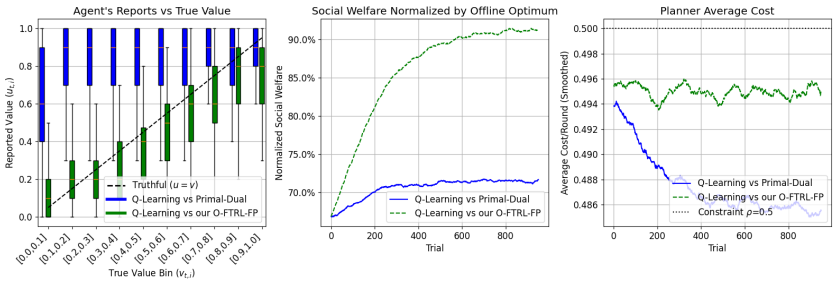


Figure 1: We simulate a $T$-round game for 1000 trials, during which agents use Q-learning to optimize their reporting strategy. Under the vanilla primal-dual algorithm of Balseiro et al. (2023), agents learn to frequently misreport their values, resulting in reduced social welfare and low budget utilization (blue). In contrast, under our incentive-aware mechanism, agents gradually learn to report truthfully, leading to significantly improved social welfare while adhering to cost constraints (green).

Indeed, as we illustrate numerically in Figure 1, classical primal-dual mechanisms are highly vulnerable to manipulation (see Section 5 for the detailed setup): Strategic agents game the learning process by distorting their current reports to influence future dual updates, thereby improving their individual utility but giving low budget utilization and less welfare. This fragility raises a natural question:

*With strategic agents, is it still possible to optimize social welfare subject to long-term constraints?*

To our knowledge, the only prior work addressing strategic agents in constrained online allocation is that of Yin et al. (2022). While their framework is a valuable step, it has two key limitations. First, it focuses on homogeneous agents with identical value distributions – an assumption critical for their equilibrium argument, but unrealistic in many applications. Second, their mechanism focuses on a specific type of cost constraint based on some "fair share" per agent, which requires knowing ideal allocation proportions in advance. These assumptions limit the practical applicability of their results. Due to space limitations, more discussions on related works are postponed to Appendix A.

This paper goes beyond these limitations and yields an *incentive-aware* primal-dual framework – one that is robust to strategic manipulation. To limit agents' influence on the future, we stabilize dual updates through epoch-based lazy updates (fixing dual variables within each epoch), which reduce the impact of any individual report on future duals and allocations. To further deter manipulation, we combine dual-adjusted pricing rounds with randomized exploration which imposes immediate utility loss on untruthful agents, thereby creating localized incentives for truthful reporting. We show that, when the dual variables are updated via the classical Follow-the-Regularized-Leader (FTRL) algorithm, our mechanism *(i)* achieves sublinear regret of $\widetilde{\mathcal{O}}(T^{2/3})$ *w.r.t.* offline optimal allocations, where $T$ is the number of rounds, *(ii)* satisfies all resource constraints exactly, and *(iii)* admits a Perfect Bayesian Equilibrium (PBE) in which agents have no incentives to misreport in most rounds.

We show that FTRL is unable to get $o(T^{2/3})$ regret; however, we observe that property *(iii)* allows the planner to treat historical reports as reliable estimates of true values, thus enabling optimistic predictions of future outcomes. Building on this, we introduce a novel online learning algorithm – Optimistic FTRL with Fixed Points (O-FTRL-FP) – which solves a small number of fixed-point problems across the time horizon to incorporate such predictive structure.[1] This gives an improved regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$, matching the $\Omega(\sqrt{T})$ lower bound for non-strategic constrained allocation (Arlotto and Gurvich, 2019). In doing so, we bridge the gap between online constrained optimization and dynamic mechanism design, enabling robust decision-making in complex, strategic environments.

## 2 Preliminaries

**Notations.** For an integer $n \geq 1$, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. For a set $\mathcal{X}$, the probabilistic simplex $\triangle(\mathcal{X})$ contains all probability distributions over $\mathcal{X}$. We use bold letters like $\boldsymbol{v}$ to denote

[1]Due to a circular dependency between actions and predictions, the O-FTRL framework (Rakhlin and Sridharan, 2013) is not applicable. Our newly proposed O-FTRL-FP framework resolves this issue, and we expect it to be of independent interest to the online learning literature; see Section 4.2.4 for more context.

**Protocol 1** Interaction Protocol for Repeated Resource Allocation

**Input:** Number of rounds $T$, number of agents $K$, value distributions $\{\mathcal{V}_i\}_{i\in[K]}$, cost distributions $\{\mathcal{C}_i\}_{i\in[K]}$, mechanism $\boldsymbol{M} = (M_t)_{t\in[T]}$, agents' strategy profile $\boldsymbol{\pi} = (\pi_{t,i})_{t\in[T],\, i\in[K]}$

Initialize public history: $\mathcal{H}_{1,0} \leftarrow \varnothing$
Initialize private history for each agent $i$: $\mathcal{H}_{1,i} \leftarrow \{(\mathcal{V}_j, \mathcal{C}_j)\}_{j\in[K]}$
**for** each round $t = 1, 2, \ldots, T$ **do**
  Each agent $i \in [K]$ observes:
  - Private value: $v_{t,i} \sim \mathcal{V}_i$
  - Public cost vector: $\boldsymbol{c}_{t,i} \sim \mathcal{C}_i$

  Agent $i$ submits report: $u_{t,i} \sim \pi_{t,i}(v_{t,i}, \boldsymbol{c}_t; \mathcal{H}_{t,i})$
  Planner applies mechanism: $(i_t, p_{t,i_t}) \sim M_t(\boldsymbol{u}_t, \boldsymbol{c}_t; \mathcal{H}_{t,0})$
  Update public history: add $(\boldsymbol{u}_t, \boldsymbol{c}_t, i_t, p_{t,i_t})$ to $\mathcal{H}_{t+1,0}$
  Each agent $i$ updates private history: add $v_{t,i}$ and $(\boldsymbol{u}_t, \boldsymbol{c}_t, i_t, p_{t,i_t})$ to $\mathcal{H}_{t+1,i}$

a vector, and use normal letters like $v_i$ for an element therein. For a random variable $X$, we use $\mathrm{PDF}(X)$ to denote its probability density function (PDF). We use $\mathcal{O}$ to hide all absolute constants, $\widetilde{\mathcal{O}}$ to additionally hide all logarithmic factors, and $\widetilde{\mathcal{O}}_T$ to focus on the polynomial dependency of $T$.

**Setup.** We consider the problem of allocating indivisible resources over $T$ rounds from a central planner to $K$ strategic agents, indexed by $1, 2, \ldots, K$. In each round, the planner allocates a single indivisible resource to one of the agents, aiming to maximize social welfare while satisfying $d$ long-term cost constraints simultaneously; more details on this objective can be found in Section 2.3.

### 2.1 Agents' Values and Costs, Planner's Allocation and Payment

In each round $t \in [T]$, agent $i \in [K]$ has a private scalar value $v_{t,i} \in [0, 1]$ and a public $d$-dimensional cost vector $\boldsymbol{c}_{t,i} \in [0, 1]^d$. Allocating the resource to agent $i$ in round $t$ yields a value of $v_{t,i}$ to the agent and incurs $c_{t,i,j}$ units of cost along dimension $j$ for all $j \in [d]$. We assume that values and costs are independent across agents and rounds; specifically, $v_{t,i}$ and $\boldsymbol{c}_{t,i}$ are *i.i.d.* samples from fixed but unknown distributions $\mathcal{V}_i \in \triangle([0, 1])$ and $\mathcal{C}_i \in \triangle([0, 1]^d)$, respectively, for all $t \in [T]$ and $i \in [K]$.

Every agent $i \in [K]$, after observing their own private value $v_{t,i}$, strategically generates a report $u_{t,i} \in [0, 1]$ which may differ from $v_{t,i}$. We defer the generation rule of such reports to Section 2.2. After observing agents' strategic reports $\boldsymbol{u}_t$ and cost vectors $\boldsymbol{c}_t$ (but without access to the true values $\boldsymbol{v}_t$), the planner either irrevocably allocates the resource to one of the agents $i_t \in [K]$ or forfeits it.

After the allocation, the planner decides a payment charged from the winner $i_t$, denoted by $p_{t,i_t}$. For all remaining agents $i \neq i_t$, the payment $p_{t,i} = 0$. The planner maximizes the $T$-round cumulative social welfare $\sum_{t=1}^{T} v_{t,i_t}$ subject to long-term constraints that the $T$-round average costs are no more than a pre-specified threshold $\boldsymbol{\rho} \in [0, 1]^d$, *i.e.*, $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{c}_{t,i_t} \leq \boldsymbol{\rho}$, where $\leq$ is compared element-wise.

### 2.2 History, Planner's Mechanism, and Agents' Strategies

At the beginning of round $t \in [T]$, the public history is given by $\mathcal{H}_{t,0} := \{(\boldsymbol{u}_\tau, \boldsymbol{c}_\tau, i_\tau, p_{\tau,i_\tau})\}_{\tau<t}$. Each agent $i \in [K]$ additionally has access to their own past values and all agents' value and cost distributions. Thus, the private history available to agent $i$ at the beginning of round $t$ is [2]

$$\mathcal{H}_{t,i} := \mathcal{H}_{t,0} \cup \{v_{\tau,i}\}_{\tau<t} \cup \{(\mathcal{V}_j, \mathcal{C}_j)\}_{j\in[K]}, \quad \forall t \in [T], i \in [K].$$

In each round $t \in [T]$, the planner determines the allocation and payment $(i_t, p_{t,i_t})$ based on agents' reports $\boldsymbol{u}_t$, cost vectors $\boldsymbol{c}_t$, public history $\mathcal{H}_{t,0}$, and possibly some internal randomness used to break ties or randomize decisions. We write $i_t = p_{t,i_t} = 0$ when the allocation is forfeited. Formally,

$$(i_t, p_{t,i_t}) \sim M_t(\boldsymbol{u}_t, \boldsymbol{c}_t; \mathcal{H}_{t,0}), \quad \text{where } M_t\colon (\boldsymbol{u}_t, \boldsymbol{c}_t, \mathcal{H}_{t,0}) \mapsto \mathrm{PDF}(i_t, p_{t,i_t}), \quad \forall t \in [T].$$

[2] We assume the distributional information is known across the agents since such information (or at least some prior) is necessary for the definition of Perfect Bayesian Equilibrium (PBE) in Definition 2. We adopt this mutually known setup as it is most challenging for the planner in terms of information asymmetry.

The collection of decision rules $\boldsymbol{M} = (M_t)_{t\in[T]}$ is referred to as the planner's *mechanism*. We emphasize that the planner does not know the agents' value or cost distributions, whereas the mechanism $\boldsymbol{M}$ is publicly known to the agents, as is standard in the literature.

For each agent $i \in [K]$, their report $u_{t,i}$ is determined based on their private value $v_{t,i}$, the cost vector $\boldsymbol{c}_t$, their private history $\mathcal{H}_{t,i}$, and potentially some internal randomness. Formally, we write

$$u_{t,i} \sim \pi_{t,i}(v_{t,i}, \boldsymbol{c}_t; \mathcal{H}_{t,i}), \quad \text{where } \pi_{t,i}\colon (v_{t,i}, \boldsymbol{c}_t, \mathcal{H}_{t,i}) \mapsto \mathrm{PDF}(u_{t,i}), \quad \forall t \in [T], i \in [K].$$

Agent $i$'s decision rules collectively form their strategy $\boldsymbol{\pi}_i := (\pi_{t,i})_{t\in[T]}$. The agents' strategies together constitute a *joint strategy* $\boldsymbol{\pi} := (\boldsymbol{\pi}_i)_{i\in[K]}$. We summarize the interaction as Protocol 1.

### 2.3 Agents' Behavior and Planner's Regret

To model agents' behavior in a dynamic environment, we adopt the $\gamma$-impatient agent framework introduced by Golrezaei et al. (2021a, 2023) which captures the idea that agents often prioritize immediate rewards over long-term gains – due to bounded rationality, uncertainty about future rounds, or limited planning horizons – while the planner is more patiently optimizing the long-run welfare.

**Assumption 1** ($\gamma$-Impatient Agents)**.** *For some fixed constant $\gamma \in (0,1)$ unknown to the planner, every agent $i \in [K]$ is $\gamma$-impatient in the sense that they maximize their $\gamma$-discounted $T$-round gain* [3]

$$V_i^{\gamma}(\boldsymbol{\pi}; \boldsymbol{M}) := \mathop{\mathbb{E}}_{\textit{Protocol 1}}\left[\sum_{t=1}^{T} \gamma^t (v_{t,i} - p_{t,i}) \mathbb{1}[i_t = i]\right], \quad \forall i \in [K], \boldsymbol{\pi} \in \Pi, \boldsymbol{M}. \tag{1}$$

In this paper, we study the equilibrium concept of Perfect Bayesian Equilibrium (PBE):

**Definition 2** (Perfect Bayesian Equilibrium)**.** *Fix a mechanism $\boldsymbol{M} = (M_1, M_2, \ldots, M_T)$. An agents' joint strategy $\boldsymbol{\pi}$ is a Perfect Bayesian Equilibrium (PBE) under $\boldsymbol{M}$, if any single agent's unilateral deviation from $\boldsymbol{\pi}$ does not increase their own gain. Formally, a joint strategy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ is a PBE if*

$$V_i^{\gamma}(\boldsymbol{\pi}_i \circ \boldsymbol{\pi}_{-i}; \boldsymbol{M}) \geq V_i^{\gamma}(\boldsymbol{\pi}_i' \circ \boldsymbol{\pi}_{-i}; \boldsymbol{M}), \quad \forall i \in [K], \forall \textit{agent } i\textit{'s strategy } \boldsymbol{\pi}_i'.$$

The planner aims to maximize social welfare, namely the expected total value yielded from allocations $\mathbb{E}[\sum_{t=1}^{T} v_{t,i_t}]$, while satisfying the $d$ long-term cost constraints $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{c}_{t,i_t} \leq \boldsymbol{\rho}$ simultaneously. To evaluate a mechanism's performance, we compare the allocations against the following *offline optimal benchmark*, which performs a hindsight optimization using agents' true values and costs:

$$\{i_t^*\}_{t\in[T]} := \mathop{\mathrm{argmax}}_{i_1^*,\ldots,i_T^* \in \{0\}\cup[K]} \sum_{t=1}^{T} v_{t,i_t^*} \quad \text{subject to} \quad \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{c}_{t,i_t^*} \leq \boldsymbol{\rho}. \tag{2}$$

We remark that benchmark in Eq. (2) depends on the full sequence of true values $\{\boldsymbol{v}_t\}_{t=1}^T$ and costs $\{\boldsymbol{c}_t\}_{t=1}^T$, which are not observable to the planner. This distinguishes it from typical online learning benchmarks, which fixes a policy before the game; see (Balseiro et al., 2023) for a related discussion. Since Eq. (2) relies on information unavailable at decision time, it cannot be matched exactly. Instead, we assess the mechanism's performance by measuring its *regret* relative to the offline optimal:

$$\mathfrak{R}_T(\boldsymbol{\pi}, \boldsymbol{M}) := \mathop{\mathbb{E}}_{\text{Protocol 1}}\left[\sum_{t=1}^{T} \left(v_{t,i_t^*} - v_{t,i_t}\right)\right], \tag{3}$$

where the expectation is over the randomness in the mechanism, agent strategies, and value/cost realizations. This regret notion generalizes the one studied in the non-strategic setting of Balseiro et al. (2023), where agents report truthfully (*i.e.*, $\boldsymbol{\pi} = \textsc{Truth}$ such that $u_{t,i} = v_{t,i}$ for all $t$ and $i$). In that setting, their mechanism $\boldsymbol{M}^0$ achieved regret $\mathfrak{R}_T(\textsc{Truth}, \boldsymbol{M}^0) = \widetilde{\mathcal{O}}(\sqrt{T})$. However, if agents are strategic, they may deviate from truthful reporting. In contrast, our mechanism in Algorithm 2 guarantees the existence of a Perfect Bayesian Equilibrium (PBE) strategy profile $\boldsymbol{\pi}$ such that no agent benefits from unilateral deviation, and under which the regret remains $\widetilde{\mathcal{O}}(\sqrt{T})$ (Theorem 3.2).

[3] Two special cases of Assumption 1 are 0-impatient agents, who only care about their gains in the current round (often referred to as myopic agent), and 1-impatient agents, who care about their total gains over the entire $T$-round game (as is typical in extensive-form games).

**Algorithm 2** Primal-Dual Mechanism Robust to Strategic Manipulations

**Input:** Number of rounds $T$, agents $K$, resources $d$, cost constraint $\boldsymbol{\rho} \in [0,1]^d$
Epochs $\{\mathcal{E}_\ell\}_{\ell=1}^L$, learning rates $\{\eta_\ell\}_{\ell=1}^L$, regularizer $\Psi : \mathbb{R}^d_{\geq 0} \to \mathbb{R}$, sub-routine $\mathcal{A}$
**Output:** Allocations $i_1, \ldots, i_T$, where $i_t = 0$ denotes no allocation

Define dual region $\boldsymbol{\Lambda} := \{\boldsymbol{\lambda} \in \mathbb{R}^d : \lambda_j \in [0, \rho_j^{-1}]\}$
**for** epoch $\ell = 1, 2, \ldots, L$ **do**
  Update dual variable $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$ via the sub-routine $\mathcal{A}$: $\boldsymbol{\lambda}_\ell \leftarrow \mathcal{A}(\ell, \eta_\ell, \Psi)$.
  **for** each round $t \in \mathcal{E}_\ell$ **do**
    Each agent $i \in [K]$ observes their value $v_{t,i} \sim \mathcal{V}_i$ and costs $\boldsymbol{c}_t \sim \mathcal{C}$
    Agent reports $u_{t,i} \in [0,1]$ according to Protocol 1; $\boldsymbol{u}_t$ and $\boldsymbol{c}_t$ become public.
    **if** round $t$ is selected for exploration (*w.p.* $1/|\mathcal{E}_\ell|$ independently) **then**
      Sample tentative agent $i \sim \text{Unif}([K])$ and payment $p \sim \text{Unif}([0,1])$
      If $u_{t,i} \geq p$, set $i_t = i$ and $p_{t,i_t} = p$. Otherwise, set $i_t = 0$
    **else**
      Compute adjusted cost $\widetilde{c}_{t,i} = \boldsymbol{\lambda}_\ell^\top \boldsymbol{c}_{t,i}$ and adjusted report $\widetilde{u}_{t,i} := u_{t,i} - \widetilde{c}_{t,i}, \forall i \in [K]$
      Allocate to agent $i_t := \arg\max_i \widetilde{u}_{t,i}$ with payment $p_{t,i_t} = \widetilde{c}_{t,i_t} + \max_{j \neq i_t} \widetilde{u}_{t,j}$
    **if** cost constraint is violated: $\sum_{s<t} \boldsymbol{c}_{s,i_s} + \boldsymbol{c}_{t,i_t} \not\leq T\boldsymbol{\rho}$ **then**
      Reject allocation by setting $i_t = 0$

In parallel, the planner also aims to minimize constraint violations, which we define as

$$\mathfrak{B}_T(\boldsymbol{\pi}, \boldsymbol{M}) := \mathop{\mathbb{E}}_{\text{Protocol 1}} \left[ \left\| \left( \sum_{t=1}^{T} (\boldsymbol{c}_{t,i_t} - \boldsymbol{\rho}) \right)_+ \right\|_1 \right], \tag{4}$$

where $(\cdot)_+$ is the coordinate-wise maximum with zero, *i.e.*, $\boldsymbol{x}_+ := [\max(x_i, 0)]_{i \in [d]}$. Our mechanism $\boldsymbol{M}$, presented in Algorithm 2, guarantees $\mathfrak{B}_T(\boldsymbol{\pi}, \boldsymbol{M}) = 0$; that is, cost constraints are satisfied.

We conclude this section with a smoothness assumption on cost distributions. The idea is to ensure that projected costs (linear combinations of the cost vector) do not place excessive probability mass on any single value. This smoothness condition prevents pathological behaviors where a small change in an agent's report could drastically alter outcomes due to spiky distributions. Such assumptions are common in strategic settings to ensure robustness to perturbations, including in bilateral trades (Cesa-Bianchi et al., 2024a), first-price auctions (Cesa-Bianchi et al., 2024b), second-price auctions with reserves (Golrezaei et al., 2021a), and smoothed revenue maximization (Durvasula et al., 2023).

**Assumption 3** (Smooth Costs)**.** *For any agent $i \in [K]$, the cost distribution $\mathcal{C}_i$ satisfies the following: for all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda} := \{\boldsymbol{\lambda} \in \mathbb{R}^d \mid \lambda_j \in [0, \rho_j^{-1}]\}$, the density of the projected cost $\text{PDF}_{\boldsymbol{c}_t \sim \mathcal{C}_i}(\boldsymbol{\lambda}^\top \boldsymbol{c}_i)$ is uniformly bounded above by some universal constant $\epsilon_c > 0$. We do not assume this $\epsilon_c$ to be known.*

## 3 Primal-Dual Mechanism Robust to Strategic Manipulation

We introduce our incentive-aware primal-dual mechanism, Algorithm 2, which overcomes the fragility of standard primal-dual methods to strategic agents. As demonstrated in Figure 1, vanilla primal-dual methods allow agents to manipulate future dual updates by misreporting, leading to misaligned incentives and degraded performance. Our mechanism addresses this challenge through three key innovations: epoch-based lazy dual updates, dual-adjusted allocation and payments, and randomized exploration rounds, which we describe below before presenting the algorithm and its guarantees.

**Epoch-Based Lazy Updates.** We divide the game horizon $[T]$ into $L$ epochs as $[T] = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_L$ and fix a single dual variable $\boldsymbol{\lambda}_\ell$ within each epoch $\mathcal{E}_\ell$. These dual variables act as implicit prices on resource consumption and adjust reported values accordingly. By holding $\boldsymbol{\lambda}_\ell$ constant within each epoch, we reduce agents' ability to manipulate future allocations through misreports. This "lazy update" scheme is a central ingredient in limiting intertemporal strategic behavior.

**Dual-Adjusted Allocation and Payments.** In each round $t \in [T]$, agents submit reports based on their private values and observed costs. With high probability, the mechanism enters a standard round, allocating the resource to the agent with the highest dual-adjusted report (reported value minus dual-weighted cost), and charging them their cost plus the second-highest dual-adjusted report.

**Algorithm 3** Dual Update Sub-Routine using Follow-the-Regularized-Leader (FTRL)

**Input:** Current epoch number $\ell$, learning rate $\eta_\ell > 0$, regularizer $\Psi\colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}$

1: Solve the following optimization problem for $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$ and return $\boldsymbol{\lambda}_\ell$.

$$\boldsymbol{\lambda}_\ell = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \sum_{\ell'<\ell} \sum_{\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i_\tau})^\intercal \boldsymbol{\lambda} + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda}). \tag{5}$$

**Algorithm 4** Dual Update Sub-Routine using Optimistic FTRL with Fixed Points (O-FTRL-FP)

**Input:** Current epoch number $\ell$, learning rate $\eta_\ell > 0$, regularizer $\Psi\colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}$

1: Solve the following fixed point problem for $(\boldsymbol{\lambda}_\ell, \widetilde{\boldsymbol{g}}_\ell) \in \boldsymbol{\Lambda} \times \mathbb{R}^d$ and return $\boldsymbol{\lambda}_\ell$.

$$\begin{aligned}
\boldsymbol{\lambda}_\ell &= \operatorname*{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \sum_{\ell'<\ell} \sum_{\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i_\tau})^\intercal \boldsymbol{\lambda} + \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\intercal \boldsymbol{\lambda} + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda}), \\
\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell) &:= |\mathcal{E}_\ell| \cdot \frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sum_{\ell'<\ell, \tau\in\mathcal{E}_{\ell'}} \left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)}\right), \\
\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell) &:= \operatorname*{argmax}_{i\in[K]} \left(u_{\tau,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{\tau,i}\right), \quad \forall \ell' < \ell, \tau \in \mathcal{E}_{\ell'}.
\end{aligned} \tag{6}$$

This payment rule, inspired by boosted second-price auctions (Golrezaei et al., 2021b), is incentive-compatible in static (one-shot) settings and encourages truthful reporting in our dynamic setup. If the allocation would violate the cumulative cost constraint, it is rejected to ensure feasibility.

**Randomized Exploration Rounds.** With a small probability, the mechanism initiates an *exploration round*, offering a random price to a randomly selected agent. This structure penalizes misreports by imposing a direct utility loss when the reported value deviates from the true value (see Theorem 4.2). These rounds act as incentive-compatible signal extractors and are essential for maintaining the accuracy of dual updates based on strategic reports. The idea of randomized pricing has been explored in repeated second-price auctions (Amin et al., 2013; Golrezaei et al., 2021a, 2023), but to our knowledge, our work is the first to leverage it for robust primal-dual learning.

**Dual Updates via Online Learning.** In the beginning of each epoch, the planner updates the dual variable $\boldsymbol{\lambda}_\ell$ via an online learning approach. Equipped with Follow-the-Regularized-Leader (FTRL), we prove $\widetilde{\mathcal{O}}(T^{2/3})$ is attainable; however, the epoch-based structure also poses an $\Omega(T^{2/3})$ online learning barrier (Theorem 4.4). To go beyond the $\widetilde{\mathcal{O}}(T^{2/3})$ regret of FTRL, we leverage the near-truthfulness induced by our mechanism to make predictions about future behavior. Our *Optimistic FTRL with Fixed Points (O-FTRL-FP)* augments classical FTRL with a forward-looking term that estimates how the current dual variable $\boldsymbol{\lambda}_\ell$ would perform if agent behavior remains consistent.

Specifically, O-FTRL-FP solves a fixed-point problem: it chooses $\boldsymbol{\lambda}_\ell$ to minimize a combination of past constraint violations $\sum_\tau (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i_\tau})^\intercal \boldsymbol{\lambda}$, a prediction term $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\intercal \boldsymbol{\lambda}$ based on simulated allocations using prior reports, and a regularization term $\frac{\Psi(\boldsymbol{\lambda})}{\eta_\ell}$. Since $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)$ itself depends on $\boldsymbol{\lambda}_\ell$, the optimization forms a self-consistent loop. We show that this method achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret while maintaining feasibility and incentive alignment.

**Main Results.** We now state our two main theoretical guarantees, corresponding to different dual update strategies. Full formal statements and proofs are provided in Appendix B.

**Theorem 3.1** (Algorithm 2 with FTRL)**.** *Under appropriate choice of epoch lengths and learning rates, Algorithm 2 using FTRL in Eq. (5) as the dual-update sub-routine $\mathcal{A}$ guarantees the existence of a PBE $\boldsymbol{\pi}^*$ such that $\mathfrak{R}_T(\boldsymbol{\pi}^*, \text{Algorithm 2}) = \widetilde{\mathcal{O}}(T^{2/3})$ and $\mathfrak{B}_T(\boldsymbol{\pi}^*, \text{Algorithm 2}) = 0$.*

**Theorem 3.2** (Algorithm 2 with O-FTRL-FP)**.** *Under appropriate choice of epoch lengths and learning rates, Algorithm 2 using O-FTRL-FP in Eq. (6) as the dual-update sub-routine $\mathcal{A}$ guarantees the existence of a PBE $\boldsymbol{\pi}^*$ such that $\mathfrak{R}_T(\boldsymbol{\pi}^*, \text{Algorithm 2}) = \widetilde{\mathcal{O}}(\sqrt{T})$ and $\mathfrak{B}_T(\boldsymbol{\pi}^*, \text{Algorithm 2}) = 0$.*

Remarkably, Arlotto and Gurvich (2019) proved that even when agents are non-strategic, with unknown value and cost distributions, it is unavoidable to suffer $\Omega(\sqrt{T})$ social welfare regret in the worst case. Therefore, our mechanism – while additionally being robust to strategic agents – matches this lower bound when focusing on poly$(T)$ dependencies.

## 4 Analysis Sketch of Theorems 3.1 and 3.2

The safety property of Algorithm 2, namely that $\mathfrak{B}_T = 0$, follows directly from Line 14. Thus, we focus on analyzing the regret $\mathfrak{R}_T = \mathbb{E}\big[\sum_{t=1}^{T}(v_{t,i_t^*} - v_{t,i_t})\big]$, which measures the expected difference between our allocation $\{i_t\}_{t\in[T]}$ and the offline optimal allocation $\{i_t^*\}_{t\in[T]}$ defined in Eq. (2). To help control this regret, we introduce an intermediate allocation that maximizes the dual-adjusted values (rather than the actual allocation $i_t$ maximizing dual-adjusted reports, which may be strategic):

$$\widetilde{i}_t^* := \operatorname*{argmax}_{i\in[K]} \left(v_{t,i} - \boldsymbol{\lambda}_\ell^\mathsf{T} \boldsymbol{c}_{t,i}\right), \quad \forall \ell \in [L], t \in \mathcal{E}_\ell.$$

Define a stopping time $\mathcal{T}_v := \min\{t \in [T] \mid \sum_{\tau=1}^{t} \boldsymbol{c}_{\tau,i_\tau} + \mathbf{1} \not\leq T\boldsymbol{\rho}\} \cup \{T+1\}$ as the last round where it is impossible for Line 14 to reject $i_t$. We decompose the regret $\mathfrak{R}_T$ as

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(v_{t,i_t^*} - v_{t,i_t})\right] \leq \mathbb{E}\Bigg[\underbrace{\sum_{t=1}^{\mathcal{T}_v}(v_{t,\widetilde{i}_t^*} - v_{t,i_t})}_{\textsc{PrimalAlloc}} + \underbrace{\sum_{t=1}^{T} v_{t,i_t^*} - \sum_{t=1}^{\mathcal{T}_v} v_{t,\widetilde{i}_t^*}}_{\textsc{DualVar}}\Bigg]. \tag{7}$$

### 4.1 Misallocations Due to Agents' Strategic Behavior (PRIMALALLOC)

As discussed in Section 3, agents' strategic reports in epoch $\ell$ can influence future dual variables $\boldsymbol{\lambda}_{\ell+1}$, which in turn affect subsequent allocations – potentially creating feedback loops that lead to misallocation. We give an example of this scenario:

**Example 4** (Agents are able to strategically affect $\boldsymbol{\lambda}_{\ell+1}$)**.** *Consider two agents with identical value distributions, i.e., $\mathcal{V}_1 = \mathcal{V}_2$. Suppose the cost vectors are fixed as $\boldsymbol{c}_{t,i} \equiv \boldsymbol{e}_i$ for all $t$ and $i$, and the cost budget is $\boldsymbol{\rho} = (1/2, 1/2)$. If both agents report truthfully during epoch $\mathcal{E}_1$, the resource is allocated approximately equally. As a result, the dual vector $\boldsymbol{\lambda}_2$, computed via Eq. (6), has similar values across its coordinates. However, suppose that agent 1 strategically under-reports their value throughout epoch 1, causing all allocations to go to agent 2 (i.e., $i_t = 2$ for all $t \in \mathcal{E}_1$). This skews the observed cost consumption toward the second type, causing $\lambda_{2,2} \gg \lambda_{2,1}$. Consequently, agent 2 – whose actions incur the second cost type – will face significantly higher penalties in future epochs.*

To understand this effect, we decompose the total inefficiency due to agents' strategic behavior, namely the PRIMALALLOC in Eq. (7), into two parts: *(i)* INTRAEPOCH measuring misallocations arisen due to agents' incentives for immediate or short-term gains, and *(ii)* INTEREPOCH measuring misallocations caused by agents influencing dual updates for future-epoch benefits. To isolate them, we introduce two behavioral models for agents, where Model 1 is exactly Assumption 1:

$$\text{Model 1: } \max_{\boldsymbol{u}} \mathbb{E}\left[\sum_{\tau=t}^{T} \gamma^\tau (v_{\tau,i} - p_{\tau,i}) \mathbb{1}[i_\tau = i]\right]; \text{ Model 2: } \max_{\boldsymbol{u}} \mathbb{E}\left[\sum_{\tau\in\mathcal{E}_\ell, \tau\geq t} \gamma^\tau (v_{\tau,i} - p_{\tau,i}) \mathbb{1}[i_\tau = i]\right]. \tag{8}$$

Model 2 essentially assumes agents only optimize over the current epoch $\mathcal{E}_\ell$, ignoring long-term impact. Let $\{i_t^h\}_{t\in[T]}$ be the allocations that would occur under Algorithm 2 if agents followed Model 2. Our goal is to first analyze this hypothetical setting to understand INTRAEPOCH, then examine the deviation introduced by agents following the realistic Model 1, which leads to INTEREPOCH effects.

#### 4.1.1 INTRAEPOCH: Misalignment Within Epochs

Under Model 2, each epoch $\ell \in [L]$ can be treated independently, which we call an "epoch-$\ell$ game". The planner selects allocations to maximize the total dual-adjusted value $\sum_{t\in\mathcal{E}_\ell} \widetilde{v}_{t,i_t^h}$, where $\widetilde{v}_{t,i} := v_{t,i} - \boldsymbol{\lambda}_\ell^\mathsf{T} \boldsymbol{c}_{t,i}$. In contrast, agents care about their true value $v_{t,i}$ rather than the dual-adjusted objective. This introduces a mismatch between the planner's and agents' optimization criteria. Despite this mismatch, we show that our mechanism's dual-adjusted allocation rule in Line 12 incentivizes truthful reporting within each epoch:

**Theorem 4.1** (INTRAEPOCH Guarantee; Informal Theorem C.2)**.** *In the epoch-$\ell$ game under Model 2, the allocation rule $i_t^h = \arg\max_i \widetilde{u}_{t,i}$ with payment $p_{t,i_t^h} = \boldsymbol{\lambda}_\ell^\mathsf{T} \boldsymbol{c}_{t,i_t^h}$ + 2nd-highest $\widetilde{u}_{t,i}$ ensures that truthful reporting is a PBE. Under this equilibrium, $\mathbb{E}[\textsc{IntraEpoch}] = 0$.*

#### 4.1.2 INTEREPOCH: Strategic Manipulation of Future Duals

We now return to Model 1, where agents optimize over the entire horizon $t \in [T]$. In this setting, an agent can misreport in the current epoch to influence the next epoch's dual variable $\boldsymbol{\lambda}_{\ell+1}$, thereby improving their chances of allocation in future rounds. This creates a new avenue for strategic behavior that was not captured under Model 2.

To mitigate this, our mechanism introduces randomized *exploration rounds*, which penalize deviations from truthful reporting through stochastic pricing: Reporting $u_{t,i} > v_{t,i}$ means paying a price higher than value when $p \in (v_{t,i}, u_{t,i})$; reporting $u_{t,i} < v_{t,i}$ means missing an opportunity to make profits when $p \in (v_{t,i}, u_{t,i})$. These rounds – by ensuring misreports carry immediate utility losses that do not outweigh the future gains – reduce agents' willingness to manipulate dual updates.

**Theorem 4.2** (PRIMALALLOC Guarantee; Informal Theorem C.5)**.** *There exists a PBE $\boldsymbol{\pi}^*$ such that for any epoch $\ell \in [L]$, the number of round-agent $(t,i)$-pairs where $|u_{t,i} - v_{t,i}| \geq \frac{1}{|\mathcal{E}_\ell|}$ is $\widetilde{\mathcal{O}}(1)$ with high probability. That is, agent reports under Models 1 and 2 rarely differ. Moreover, the resulting allocations $\{i_t\}_{t\in\mathcal{E}_\ell}$ and $\{i_t^h\}_{t\in\mathcal{E}_\ell}$ differ in at most $\widetilde{\mathcal{O}}(1)$ rounds with high probability. Consequently, this equilibrium ensures $\mathbb{E}[\text{PRIMALALLOC}] = \widetilde{\mathcal{O}}(L)$.*

### 4.2 Inaccurate Dual Variables Due to Incomplete Information (DUALVAR)

The second source of inefficiency comes from sub-optimal dual variables – or more precisely, the gap between the dual-adjusted allocation $\widetilde{i}_t^* := \arg\max_i \left(v_{t,i} - \boldsymbol{\lambda}_\ell^\mathsf{T} \boldsymbol{c}_{t,i}\right)$ that our mechanism maximizes and the offline optimal benchmark $i_t^*$ defined in Eq. (2) – which we call DUALVAR in Eq. (7).

#### 4.2.1 Translating DUALVAR to Online Learning Regret

Using primal-dual analysis similar to that of Balseiro et al. (2023), we relate this gap to the online learning regret over dual variables $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_L \in \boldsymbol{\Lambda}$ (where $\boldsymbol{\Lambda} := \{\boldsymbol{\lambda} \in \mathbb{R}^d : \lambda_j \in [0, \rho_j^{-1}]\}$ is chosen such that $\rho_j^{-1}\boldsymbol{e}_j \in \boldsymbol{\Lambda}$ for all $j \in [d]$) as follows, which is an informal version of Lemma D.1:

**Lemma 4.3** (DUALVAR as Online Learning Regret)**.** *Let $\widetilde{i}_t^* = \text{argmax}_{i\in[K]}(v_{t,i} - \boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,i})$ denote the best allocation under dual prices and $i_t^*$ denote the offline optimal benchmark. Then,*

$$\mathbb{E}[\text{DUALVAR}] := \mathbb{E}\left[\sum_{t=1}^{T} v_{t,i_t^*} - \sum_{t=1}^{\mathcal{T}_v} v_{t,\widetilde{i}_t^*}\right] \lesssim \underbrace{\mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \sum_{\ell=1}^{L}\sum_{t\in\mathcal{E}_\ell} (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_\ell - \boldsymbol{\lambda}^*)\right]}_{=:\mathfrak{R}_L^{\boldsymbol{\lambda}}} + \widetilde{\mathcal{O}}(L).$$

That is, DUALVAR reduces to an online learning problem where the planner selects a dual vector $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$ at the beginning of each epoch $\ell$ and incurs a linear loss based on constraint violations of $\{i_t\}_{t\in\mathcal{E}_\ell}$. Specifically, the loss function in epoch $\ell$ is given by $F_\ell(\boldsymbol{\lambda}) := \sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}$.

#### 4.2.2 Achievements and Limitations of FTRL

Using FTRL, the planner selects $\boldsymbol{\lambda}_\ell$ by minimizing a regularized sum of historical losses, namely $\boldsymbol{\lambda}_\ell = \text{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \sum_{\ell'<\ell} F_{\ell'}(\boldsymbol{\lambda}) + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda})$ where $\Psi$ is a strongly convex regularizer. This choice yields regret $\mathfrak{R}_L^{\boldsymbol{\lambda}} = \widetilde{\mathcal{O}}\left(\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2}\right)$. Choosing $L = T^{1/3}$ epochs each of size $T^{2/3}$ minimizes $\mathfrak{R}_L^{\boldsymbol{\lambda}} + L$, giving total regret $\widetilde{\mathcal{O}}(T^{2/3})$ as shown in Theorem 3.1.

However, vanilla FTRL is fundamentally limited due to epoch-based structures: Frequent updates to $\boldsymbol{\lambda}_\ell$ would enable faster learning, but break incentive compatibility. Conversely, fewer updates protect incentives, but slow learning. This tradeoff is formalized in the following hardness result:

**Theorem 4.4** (Hardness for Low-Switching Online Learning (Dekel et al., 2014))**.** *Consider an online learning algorithm that guarantees regret $\mathfrak{R}_L^{\boldsymbol{\lambda}} = \mathcal{O}(T^\alpha)$ for some $\alpha \in [1/2, 1)$. Then, it must switch decisions at least $\Omega(T^{2(1-\alpha)})$ times. In our setting, this means that the number of dual updates – i.e., the number of epochs $L$ – must satisfy $L = \Omega(T^{2(1-\alpha)})$. This infers that the $\mathfrak{R}_L^{\boldsymbol{\lambda}} + L$ term in Lemma 4.3 suffers a worst-case bound of $\Omega\left(\min_{\alpha\in[1/2,1)} T^\alpha + T^{2(1-\alpha)}\right) = \Omega(T^{2/3})$.*

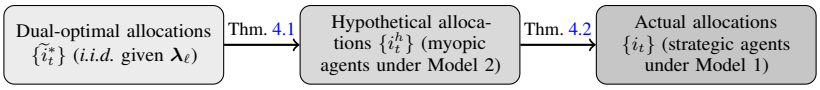


Figure 2: From dual-optimal to actual allocations: fixing dual $\boldsymbol{\lambda}_\ell$, our mechanism ensures that the actual allocations $\{i_t\}$ closely follow the dual-optimal $\{\widetilde{i}_t^*\}$ via an intermediate myopic model $\{i_t^h\}$.

#### 4.2.3 Exploiting Incentive Alignments for Boosted Regret

The hardness result in Theorem 4.4 applies to the worst-case senario, *i.e.*, when the loss functions $F_1, F_2, \ldots, F_L$ can be any arbitrary linear functions. But in our setting, due to the incentive-compatible primal allocations (Theorems 4.1 and 4.2), the loss $F_\ell(\boldsymbol{\lambda})$ has an almost-*i.i.d.* structure:

**Claim 4.5** (Loss Structure; Informal)**.** *Fix any $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$. Then for all but $\widetilde{\mathcal{O}}(1)$ rounds in epoch $\mathcal{E}_\ell$, the actual allocations $i_t$ match the dual-optimal choices $\widetilde{i}_t^* = \arg\max_i (v_{t,i} - \boldsymbol{\lambda}_\ell^\mathsf{T} \boldsymbol{c}_{t,i})$. Thus,*

$$\nabla_{\boldsymbol{\lambda}} F_\ell(\boldsymbol{\lambda}) = \sum_{t \in \mathcal{E}_\ell} (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t}) \approx \sum_{t \in \mathcal{E}_\ell} (\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*}),$$

*which behaves is the sum of $|\mathcal{E}_\ell|$ i.i.d. samples and thus has a variance of order $\widetilde{\mathcal{O}}(|\mathcal{E}_\ell|)$.*

*Proof Idea.* To understand how our mechanism enables accurate dual updates despite incomplete information and strategic behavior, we illustrate in Figure 2 the connection between three key allocation sequences within a fixed epoch. First, the sequence $\{\widetilde{i}_t^*\}$ represents the dual-optimal allocations computed in hindsight, assuming access to true valuations and a fixed dual vector $\boldsymbol{\lambda}_\ell$. Second, $\{i_t^h\}$ denotes the hypothetical allocations made under Model 2, where agents are myopic and only optimize within a single epoch. By Theorem 4.1, these allocations align with $\{\widetilde{i}_t^*\}$ under our incentive-compatible subroutine. Lastly, the actual sequence $\{i_t\}$, generated by strategic agents under Model 1, is shown to be close to $\{i_t^h\}$ via Theorem 4.2. Together, these approximations establish that $\{i_t\}$ behaves almost like an *i.i.d.* sample from the dual-adjusted best responses. □

This low-variance structure is essential for achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. In general, linear losses with gradients of norm $\widetilde{\mathcal{O}}(|\mathcal{E}_\ell|)$ can have variances as large as $\widetilde{\mathcal{O}}(|\mathcal{E}_\ell|^2)$, which explains the $T^{2/3}$ regret scaling of FTRL in Theorem 3.1 and the lower bound in Theorem 4.4. Even more importantly, because this near-*i.i.d.* structure holds across previous epochs as well, it enables us to accurately estimate losses associated with new dual choices using only historical data – without requiring access to agents' true values or distributions. These two insights provides predictability ahead of time, which enables the application of optimistic online learning algorithms, for example, Optimistic FTRL (O-FTRL) by Rakhlin and Sridharan (2013). O-FTRL ensures that if the actual loss $F_\ell$ is well-predicted by some predicted loss $\widehat{F}_\ell$, such that the expected squared error in gradients is of order $\widetilde{\mathcal{O}}(|\mathcal{E}_\ell|)$, then one can break the $\widetilde{\mathcal{O}}(T^{2/3})$ regret barrier and attain $\widetilde{\mathcal{O}}(\sqrt{T})$ performance (Lemma E.2).

#### 4.2.4 Resolving Circular Dependencies between Actions and Predictions

The final issue stopping us from obtaining Theorem 3.2 is that, O-FTRL framework requires us to construct the predicted loss $\widehat{F}_\ell(\boldsymbol{\lambda})$ before deciding action $\boldsymbol{\lambda}_\ell$; however, in our case, recall that

$$F_\ell(\boldsymbol{\lambda}) = \sum_{t \in \mathcal{E}_\ell} (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T} \boldsymbol{\lambda}, \quad \text{where } i_t \text{ is the primal allocation given dual } \boldsymbol{\lambda}_\ell.$$

In words, to construct a good $\widehat{F}_\ell(\boldsymbol{\lambda}) \approx F_\ell(\boldsymbol{\lambda})$ and decide $\boldsymbol{\lambda}_\ell$, we need to know $\boldsymbol{\lambda}_\ell$ first because $F_\ell(\boldsymbol{\lambda})$ depends on $\boldsymbol{\lambda}_\ell$. This *circular dependency* between action $\boldsymbol{\lambda}_\ell$ and prediction $\widehat{F}_\ell(\boldsymbol{\lambda})$ stops us from applying O-FTRL. To circumvent this issue, we instead allow the prediction to have a form $\widehat{F}_\ell(\boldsymbol{\lambda}; \boldsymbol{\lambda}_\ell)$ – such that $\widehat{F}_\ell(\cdot; \boldsymbol{\lambda}_\ell) \approx F_\ell(\cdot)$ if we really chose $\boldsymbol{\lambda}_\ell$ as our action for epoch $\ell$ – and decide the action via a fixed point problem (as in Eq. (6)). We call this novel online learning algorithm **O**ptimistic **FTRL** with **F**ixed **P**oints (O-FTRL-FP in short). In Lemma D.4, we prove that if a small perturbation in $\boldsymbol{\lambda}_\ell$ doesn't change $\nabla_{\boldsymbol{\lambda}} \widehat{F}_\ell(\boldsymbol{\lambda}; \boldsymbol{\lambda}_\ell)$ by a lot, O-FTRL-FP always admits an approximate fixed point.

To get the $\widetilde{\mathcal{O}}(\sqrt{T})$ social welfare regret claimed in Theorem 3.2, it only remains to *i)* verify our $\widehat{F}_\ell$ in Eq. (6) indeed makes Lemma D.4 applicable; *ii)* show that $\mathbb{E}[\|\nabla_\ell F_\ell(\boldsymbol{\lambda}) - \nabla_\ell \widehat{F}_\ell(\boldsymbol{\lambda}; \boldsymbol{\lambda}_\ell)\|_2^2] = \widetilde{\mathcal{O}}(|\mathcal{E}_\ell|)$ despite agents' misreports in the current epoch $\mathcal{E}_\ell$, the statistical barrier in reconstructing $\{(\mathcal{V}_i, \mathcal{C}_i)\}_i$, and agents' historical misreports; and *iii)* properly tune all hyper-parameters. Due to space limitations, these steps are deferred to Theorem D.3 in the appendix.

## 5 Numerical Study

**Basic Setup.** We simulate a game with $T = 1000$ rounds, $K = 3$ agents, a single resource dimension $d = 1$, and a discount factor $\gamma = 0.9$. Each agent's valuation is drawn from $\mathcal{V}_i = \text{Unif}[0, 1]$, and their cost is drawn from $\mathcal{C}_i = \text{Unif}[0.7\rho, 1.3\rho]$, for all $i = 1, 2, 3$.

**Agents' Model.** To reflect agents' strategic behaviors, we assume that every agent $i = 1, 2, 3$ models the game as a Markov Decision Process (MDP) with state defined as $(t, \boldsymbol{\lambda}_t, v_{t,i})$ – *i.e.*, round number $t$, current dual variable $\boldsymbol{\lambda}_t$,[4] and their own private value $v_{t,i}$ – and the action defined as $u_{t,i}$. The reward for agent $i$ after playing a report $u_{t,i}$ is $\gamma^t(v_{t,i} - p_{t,i})\mathbb{1}[i_t = i]$ as in Assumption 1. We discretize all the values, duals, and reports to the nearest multiple of 0.1. We repeat the same game for $N = 1000$ independent trails, where every agent keeps refining their strategy via Q-learning (Watkins and Dayan, 1992). In the $n$-th trial, every agent uses $\epsilon_n$-greedy with a geometrically decaying schedule of $\epsilon = 0.995^n$. We update Q-tables using $Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma(\max_{a'} Q(s', a')) - Q(s, a))$ where $\alpha = 0.1$, $r$ is the instantaneous reward, and $s'$ is the new state $(t + 1, \boldsymbol{\lambda}_{t+1}, v_{t+1,i})$.

**Mechanisms.** The vanilla primal-dual algorithm we use is Algorithm 1 of Balseiro et al. (2023), which ensures constraint satisfaction by rejecting any allocation that would violate the budget. When implementing the O-FTRL-FP update rule in Eq. (6), instead of solving the fixed-point problem $\boldsymbol{\lambda}_\ell = \text{argmin}_{\boldsymbol{\lambda}}(\cdots + \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\intercal \boldsymbol{\lambda} + \frac{\Psi(\boldsymbol{\lambda})}{\eta_\ell})$, we solve $\text{argmin}_{\boldsymbol{\lambda}}(\cdots + \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda})^\intercal \boldsymbol{\lambda} + \frac{\Psi(\boldsymbol{\lambda})}{\eta_\ell})$ for numerical simplicity. The validity of this approximation is due to Lemma D.5, which says $\forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda}$ s.t. $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_1 \leq \epsilon$, $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_1) \approx \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_2)$ *w.h.p.* Thus the two objectives agree locally around the true $\boldsymbol{\lambda}_\ell$.

**Results.** Figure 1 demonstrates that our mechanism significantly outperforms the standard primal-dual method in the presence of strategic agents, both adhering to cost constraints. Under the primal-dual approach of Balseiro et al. (2023, Algorithm 1) which is plotted in blue, agents systematically over-report their valuations (left), hurting the overall social welfare (middle), and resulting in lower budget utilization – when agents submit similar reports, the impact of cost minimization is amplified (right). In contrast, our mechanism equipped with O-FTRL-FP (in green) incentivizes near-truthful reporting, as shown by the alignment between reported and true values, resulting in a substantial increase in long-term social welfare. These findings highlight the fragility of standard primal-dual methods in strategic settings and the robustness of our proposed incentive-aware mechanism.

**Computational Resources.** Every illustration executes on a M1 MacBook Air 2020 in 10 minutes.

## 6 Conclusion and Future Directions

This paper investigates the dynamic allocation of reusable resources to strategic agents under multi-dimensional long-term cost constraints. We show that standard primal-dual methods, though effective in non-strategic settings, are vulnerable to manipulation when agents act strategically. To address this, we introduce a novel incentive-aware mechanism that stabilizes dual updates via an epoch-based structure and leverages randomized exploration rounds to extract truthful signals. Equipped with a computationally efficient FTRL dual update rule, our mechanism guarantees sublinear regret with respect to an offline benchmark, satisfies all cost constraints, and admits a PBE; further leveraging a novel O-FTRL-FP framework for dual updates, we boost the regret to $\widetilde{\mathcal{O}}(\sqrt{T})$ – which is near-optimal even in non-strategic constrained dynamic resource allocation settings. Looking ahead, several promising research directions remain open. For example, while our mechanism uses monetary transfers to ensure incentive compatibility, many real-world applications, *e.g.*, organ matching, school admissions, or vaccine distribution, operate under non-monetary constraints. Extending our framework to such settings is an important step toward more broadly applicable mechanism design.

[4] Since mechanism $\boldsymbol{M}$ is public and every agent knows every information the planner has, *i.e.*, $\mathcal{H}_{t,0} \subseteq \mathcal{H}_{t,i}$, agents are able to calculate $\boldsymbol{\lambda}_t$ on their end.

## Acknowledgements

P.J. and Y.D. are partially funded by the Office of Naval Research (ONR) under Award ID N00014-24-1-2470. N.G. and Y.D. are partially supported by the MIT Junior Faculty Research Assistance Grant and by the Office of Naval Research (ONR) under Award ID N00014-23-1-2584. The authors thank the anonymous reviewers for their constructive feedback.

## References

Ivo Adan, Jos Bekkers, Nico Dellaert, Jan Vissers, and Xiaoting Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health care management science*, 12:129–141, 2009.

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

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. *Advances in neural information processing systems*, 26, 2013.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Repeated contextual auctions with strategic buyers. *Advances in Neural Information Processing Systems*, 27, 2014.

Alessandro Arlotto and Itai Gurvich. Uniformly bounded regret in the multisecretary problem. *Stochastic Systems*, 9(3):231–260, 2019.

Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4): 328–346, 1950.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Santiago R Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.

Santiago R Balseiro, Huseyin Gurkan, and Peng Sun. Multiagent mechanism design without money. *Operations Research*, 67(5):1417–1436, 2019.

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

Siddhartha Banerjee, Giannis Fikioris, and Eva Tardos. Robust pseudo-markets for reusable public resources. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 241–241, 2023.

Damien Berriaud, Ezzat Elokda, Devansh Jalota, Emilio Frazzoli, Marco Pavone, and Florian Dörfler. To spend or to gain: Online learning in repeated karma auctions. *arXiv preprint arXiv:2403.04057*, 2024.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Moise Blanchard and Patrick Jaillet. Near-optimal mechanisms for resource allocation without monetary transfers. *arXiv preprint arXiv:2408.10066*, 2024.

Rajkumar Buyya, Chee Shin Yeo, and Srikumar Venugopal. Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In *2008 10th IEEE international conference on high performance computing and communications*, pages 5–13. Ieee, 2008.

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

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Regret analysis of bilateral trade with a smoothed adversary. *Journal of Machine Learning Research*, 25(234):1–36, 2024a.

Nicolò Cesa-Bianchi, Tommaso Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. The role of transparency in repeated first-price auctions with unknown valuations. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 225–236, 2024b.

Edward H Clarke. Multipart pricing of public goods. *Public choice*, pages 17–33, 1971.

Richard Cole, Vasilis Gkatzelis, and Gagan Goel. Positive results for mechanism design without money. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1165–1166, 2013.

Yan Dai, Moise Blanchard, and Patrick Jaillet. Non-monetary mechanism design without distributional information: Using scarce audits wisely. *arXiv preprint arXiv:2502.08412*, 2025.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.

Nikhil R Devanur and Thomas P Hayes. The adwords problem: online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 71–78, 2009.

Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *Journal of the ACM (JACM)*, 66(1):1–41, 2019.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Naveen Durvasula, Nika Haghtalab, and Manolis Zampetakis. Smoothed analysis of online non-parametric auctions. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 540–560, 2023.

Jon Feldman, Monika Henzinger, Nitish Korula, Vahab S Mirrokni, and Cliff Stein. Online stochastic packing applied to display ad allocation. In *European Symposium on Algorithms*, pages 182–194. Springer, 2010.

Giannis Fikioris, Siddhartha Banerjee, and Éva Tardos. Online resource sharing via dynamic max-min fairness: efficiency, robustness and non-stationarity. *arXiv preprint arXiv:2310.08881*, 2023.

Rigel Galgana and Negin Golrezaei. Learning in repeated multiunit pay-as-bid auctions. *Manufacturing & Service Operations Management*, 27(1):200–229, 2025.

A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

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

Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. No-regret learning in price competitions under consumer reference effects. *Advances in Neural Information Processing Systems*, 33:21416–21427, 2020.

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning:: Robust pricing in contextual auctions. *Operations Research*, 69(1):297–314, 2021a.

Negin Golrezaei, Max Lin, Vahab Mirrokni, and Hamid Nazerzadeh. Boosted second price auctions: Revenue optimization for heterogeneous bidders. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 447–457, 2021b.

Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. Incentive-aware contextual pricing with non-parametric market noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9331–9361. PMLR, 2023.

Artur Gorokh, Siddhartha Banerjee, and Krishnamurthy Iyer. The remarkable robustness of the repeated fisher market. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 562–562, 2021a.

Artur Gorokh, Siddhartha Banerjee, and Krishnamurthy Iyer. From monetary to nonmonetary mechanism design via artificial currencies. *Mathematics of Operations Research*, 46(3):835–855, 2021b.

Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

Mingyu Guo and Vincent Conitzer. Strategy-proof allocation of multiple items between two agents without payments or priors. In *AAMAS*, pages 881–888, 2010.

Anupam Gupta and Marco Molinaro. How the experts algorithm can help solve lps online. *Mathematics of Operations Research*, 41(4):1404–1431, 2016.

Li Han, Chunzhi Su, Linpeng Tang, and Hongyang Zhang. On strategy-proof allocation without payments or priors. In *International Workshop on Internet and Network Economics*, pages 182–193. Springer, 2011.

Devansh Jalota, Matthew Tsao, and Marco Pavone. Catch me if you can: Combatting fraud in artificial currency based government benefits programs. *arXiv preprint arXiv:2402.16162*, 2024.

Yash Kanoria and Hamid Nazerzadeh. Dynamic reserve prices for repeated auctions: Learning from bids. In *Web and Internet Economics: 10th International Conference*, volume 8877, page 232. Springer, 2014.

Thomas Kesselheim, Klaus Radke, Andreas Tonnis, and Berthold Vocking. Primal beats dual on online packing lps in the random-order model. *SIAM Journal on Computing*, 47(5):1939–1964, 2018.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017.

Christos Koufogiannakis and Neal E Young. A nearly linear-time ptas for explicit fractional packing and covering linear programs. *Algorithmica*, 70:648–674, 2014.

Xiaocheng Li, Chunlin Sun, and Yinyu Ye. Simple and fast algorithm for binary integer and online linear programming. *Mathematical Programming*, 200(2):831–875, 2023.

Antonio Miralles. Cardinal bayesian allocation mechanisms without transfers. *Journal of Economic Theory*, 147(1):179–206, 2012.

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

J.R. Munkres. *Topology*. Featured Titles for Topology. Prentice Hall, Incorporated, 2000. ISBN 9780131816299. URL `https://books.google.com/books?id=XjoZAQAAIAAJ`.

Mahyar Movahed Nejad, Lena Mashayekhy, and Daniel Grosu. Truthful greedy mechanisms for dynamic virtual machine provisioning and allocation in clouds. *IEEE transactions on parallel and distributed systems*, 26(2):594–603, 2014.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

N.A. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2): 187–217, 1975.

Christian Stummer, Karl Doerner, Axel Focke, and Kurt Heidenberger. Determining location and size of medical departments in a hospital network: A multiobjective decision support approach. *Health care management science*, 7:63–71, 2004.

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

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Zongjun Yang, Luofeng Liao, Yuan Gao, and Christian Kroer. Online fair allocation with best-of-many-worlds guarantees. *arXiv preprint arXiv:2408.02403*, 2024.

Steven Yin, Shipra Agrawal, and Assaf Zeevi. Online allocation and learning in the presence of strategic agents. *Advances in Neural Information Processing Systems*, 35:6333–6344, 2022.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Figure 1, we illustrated the fragility of primal-dual framework to agents' strategic behaviors. In Theorems 3.1 and 3.2 (with formal versions appearing as Theorems B.1 and B.2), we prove the claimed performance guarantee of our proposed mechanism in Algorithm 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All assumptions are clearly stated. These limitations are discussed in the Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Our main claims, Theorems 3.1 and 3.2, have their sketched proofs in Section 4 and their full proofs in Appendices B to D.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The numerical illustration setup is clearly explained in Section 5, with codes attached as supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code used for simulation is attached in the supplementary materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: They are described in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The claim regarding agents' truth-reporting, *e.g.*, the first plot in Figure 1, is plotted using repeated sampling and error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Work of purely theoretical nature.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Work of purely theoretical nature.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Work of purely theoretical nature.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In our numerical illustrations, we reproduced the algorithm designed by Balseiro et al. (2023) with proper citations.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets released.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing and research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No crowdsourcing and research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Work of purely theoretical nature without LLM involvement.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# Appendices



## A More Discussions on Related Work

**Dynamic Resource Allocation with Non-Strategic Agents.** With non-strategic agents, dynamic resource allocation, or more generally, online linear programming, was first studied under the random permutation model where an adversary selects a set of requests that are presented in a random order (Devanur and Hayes, 2009; Feldman et al., 2010; Gupta and Molinaro, 2016), which is more general than our *i.i.d.* model where all values and costs are identically distributed. For the *i.i.d.* model, various primal-dual-based algorithms were proposed with the main focus of refining computational efficiency (Agrawal et al., 2014; Devanur et al., 2019; Kesselheim et al., 2018; Li et al., 2023); specifically, Li et al. (2023) proposed a fast $\mathcal{O}(T^{1/2})$-regret algorithm for the online linear programming problem, which matches the $\Omega(T^{1/2})$ online resource allocation lower bound (Arlotto and Gurvich, 2019).

Recent progress on this problem includes $o(T^{1/2})$ regret with known distributions (Sun et al., 2020) or better robustness and adaptivity to adversarial corruptions (Balseiro et al., 2023; Yang et al., 2024). Another closely related problem is Bandits with Knapsacks (BwK), where the planner makes allocations without observing the values or costs in advance but only learns via post-decision feedback (Badanidiyuru et al., 2018; Castiglioni et al., 2022). Nevertheless, none of them considers strategic behaviors of the agents but assume the values and costs are fully truthful. In contrast, our main focus is to be robust to agents' strategic behaviors while remaining efficient and obeying all the constraints.

**Dynamic Resource Allocation with Strategic Agents.** When agents are strategic, ensuring both efficiency and incentive-compatibility provide a foundation for static truthful allocation with monetary transfers. In static (one-shot) allocations when money can be redistributed, the celebrated VCG mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973) provides a foundation way to achieve both. We study the repeated setup where money can only be charged but not redistributed, the so-called "money-burning" setup. A related problem is learning prices in repeated auctions (Amin et al., 2013, 2014; Kanoria and Nazerzadeh, 2014; Golrezaei et al., 2021a, 2023). These works focus on a seller learning prices to maximize revenue where the agents strategically react to these prices, sometimes

subject to buyer's budget constraints. In our work, we study the different social welfare maximization task and consider more general multi-dimensional cost constraints.

We also briefly discuss the settings where monetary transfers are completely disallowed. In static setups, incentive-compatibility is in general hard due to the Arrow's impossibility theorem (Arrow, 1950; Gibbard, 1973; Satterthwaite, 1975), though some positive results exist under restricted assumptions (Miralles, 2012; Guo and Conitzer, 2010; Han et al., 2011; Cole et al., 2013). Many recent efforts have been made to ensure efficiency and compatibility in repeated non-monetary allocations, which have very different setups from ours, for example when agents' value distributions are known (Balseiro et al., 2019; Gorokh et al., 2021b; Blanchard and Jaillet, 2024), when a pre-determined "fair share" is revealed to the planner (Gorokh et al., 2021a; Yin et al., 2022; Banerjee et al., 2023; Fikioris et al., 2023), or when the planner has extra power like audits (Jalota et al., 2024; Dai et al., 2025). We also remark that they either do not consider constraints or only have a specific "fair share" constraint that we discuss later. In contrast, we consider general multi-dimensional constraints.

**Multi-Agent Learning.** While our planner learns for better allocation mechanisms, the agents are also learning in reaction to it (*e.g.*, in our numerical illustration in Figure 1, we consider agents who use Q-learning to learn the best reporting strategies under different dual variables). There is a rich literature investigating this dynamics as well, for example Balseiro and Gur (2019); Golrezaei et al. (2020); Berriaud et al. (2024); Galgana and Golrezaei (2025) studying the convergence to equilibria when multiple agents deploy no-regret online learning algorithms in reaction to some mechanism at the same time. While such results are stronger than our existence of equilibrium results in the sense that agents find such an equilibrium on their own, we remark that the main focus of this work is designing robust mechanisms for the planner instead of designing learning algorithms for the agents.

**Comparison with (Yin et al., 2022).** Yin et al. (2022) also study the problem of ensuring efficiency and incentive-compatibility in the dynamic constrained resource allocation problem. The first critical difference is that they assume all the agents have identical value distributions, which is crucial for incentive-compatibility: By comparing one agents' reports to all the opponents', unilateral deviation from TRUTH is easily caught and thus TRUTH is a PBE even without using monetary transfers. In contrast, we need more delicate algorithmic components, including epoch-based lazy updates, random exploration rounds, and dual-adjusted allocation and payment plans, to ensure the near-truthfulness of agents. Another main difference is the type of constraints. They study a specific "fair share" type resource constraint, which says given some $\boldsymbol{p} \in \triangle([K])$, the number of allocations that agent $i$ receive should be roughly $Tp_i, \forall i \in [K]$. In contrast, our multi-dimensional long-term constraint is strictly more general than theirs, which can be written as $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{e}_{i_t} \leq \boldsymbol{p}$ in our language.

# B Proof of Main Theorems

## B.1 Main Theorem for Algorithm 2 with FTRL

**Theorem B.1** (Formal Version of Theorem 3.1: Algorithm 2 with FTRL)**.** *In Algorithm 2, let*

$$L = T^{1/3},\ \mathcal{E}_\ell = \left[(\ell-1)\frac{T}{L}+1, \ell\frac{T}{L}\right],\ \eta_\ell = \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_\ell|^2\right)^{-1/2},\ \Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2,$$

*and the sub-routine $\mathcal{A}$ chosen as FTRL (Eq. (5)). Then under Assumptions 1 and 3, there exists a PBE of agents' joint strategies under Algorithm 2, denoted by $\boldsymbol{\pi}^*$, such that*

$$\begin{aligned}&\mathfrak{R}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (5)})\\ &\leq T^{2/3}\left(4+4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log T^{1/3} + \frac{\log(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)KT^{4/3})}{\log\gamma^{-1}} + \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2\right) + \|\boldsymbol{\rho}^{-1}\|_1,\end{aligned}$$

*and* $\mathfrak{B}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (5)}) = 0.$

*Specifically, when only focusing on polynomial dependencies on $d, K$, and $T$, we have*

$$\begin{aligned}&\mathfrak{R}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (5)}) = \widetilde{\mathcal{O}}_{d,K,T}((K^2+\sqrt{d})T^{2/3}),\\ &\mathfrak{B}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (5)}) = 0.\end{aligned}$$

*Proof.* As mentioned in the main text, we introduce an intermediate allocation:

$$\widetilde{i}_t^* := \operatorname*{argmax}_{i\in[K]} \left(v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}\right), \quad \forall \ell \in [L], t \in \mathcal{E}_\ell. \tag{9}$$

Define stopping time $\mathcal{T}_v$ as the last round where no constraint violations can happen:

$$\mathcal{T}_v := \min\left\{ t \in [T] \;\middle|\; \sum_{\tau=1}^{t} \boldsymbol{c}_{\tau,i_\tau} + \mathbf{1} \not\leq T\boldsymbol{\rho} \right\} \cup \{T+1\}. \tag{10}$$

Thus in rounds $t \leq \mathcal{T}_v$, Line 14 never rejects our allocation $i_t$. Decompose the regret $\mathfrak{R}_T$ as

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(v_{t,i_t^*} - v_{t,i_t})\right] \leq \mathbb{E}\left[\underbrace{\sum_{t=1}^{\mathcal{T}_v}(v_{t,\widetilde{i}_t^*} - v_{t,i_t})}_{\textsc{PrimalAlloc}} + \underbrace{\sum_{t=1}^{\mathcal{T}_v}(v_{t,i_t^*} - v_{t,\widetilde{i}_t^*})}_{\textsc{DualVar}} + \underbrace{(T-\mathcal{T}_v)}_{\textsc{NoAct}}\right].$$

From Theorem C.5 presented later in Appendix C, there exists a PBE $\boldsymbol{\pi}^*$ such that

$$\begin{aligned}
&\mathbb{E}[\textsc{PrimalAlloc}] \leq \sum_{\ell=1}^{L}(N_\ell + 3),\\
&N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4).
\end{aligned}$$

This is the first main technical contribution of our paper, namely justifying that Algorithm 2 – equipped with epoch-based lazy updates, uniform exploration rounds, and dual-adjusted allocation and payment plans – is robust to agents' strategic manipulations.

From Theorem D.2 presented later in Appendix D, when setting

$$\Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2, \quad \eta_\ell = \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_\ell|^2\right)^{-1/2},$$

and under the same $\boldsymbol{\pi}^*$, the $\mathbb{E}[\textsc{DualVar}]$ term is bounded by

$$\mathbb{E}[\textsc{DualVar}] \leq \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2 \cdot \sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} + \sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + \|\boldsymbol{\rho}^{-1}\|_1.$$

Thus it only remains to balance the $\sum_{\ell=1}^{L} N_\ell = \widetilde{\mathcal{O}}(L)$ term and $\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2}$ term. Setting $L = T^{2/3}$ and $|\mathcal{E}_\ell| = T^{1/3}$ for all $\ell \in [L]$ ensures $\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} = \sqrt{T^{2/3}\times T^{2/3}} = T^{2/3}$ and thus

$$\begin{aligned}
&\mathfrak{R}_T(\boldsymbol{\pi}^*, \text{Algorithm 2 with Eq. (5)}) \leq \mathbb{E}[\textsc{PrimalAlloc}] + \mathbb{E}[\textsc{DualVar}]\\
&\leq \sum_{\ell=1}^{L}(N_\ell+3) + \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2 \cdot T^{2/3} + \sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + \|\boldsymbol{\rho}^{-1}\|_1\\
&= \sum_{\ell=1}^{L}(N_\ell+3)(1+\|\boldsymbol{\rho}^{-1}\|_1) + \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2 \cdot T^{2/3} + \|\boldsymbol{\rho}^{-1}\|_1.
\end{aligned}$$

Plugging our specific epoching rule that $L = T^{2/3}$ and $|\mathcal{E}_\ell| = T^{1/3}$ into $N_\ell$, we get

$$\sum_{\ell=1}^{L}(N_\ell+3) \leq T^{2/3} \times \left(4 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log T^{1/3} + \frac{\log(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)KT^{4/3})}{\log\gamma^{-1}}\right),$$

which gives our claimed bound after rearrangement. For $\mathfrak{B}_T$, since Line 14 rejects every infeasible allocation, we trivially have $\mathfrak{B}_T(\boldsymbol{\pi}^*, \text{Algorithm 2 with Eq. (5)}) = 0$. □

### B.2 Main Theorem for Algorithm 2 with O-FTRL-FP

**Theorem B.2** (Formal Version of Theorem 3.2: Algorithm 2 with O-FTRL-FP)**.** *In Algorithm 2, let*

$$L = \lceil \log T \rceil,\ \mathcal{E}_\ell = \left[2^{\ell-1}, \min(2^\ell - 1, T)\right],\ \eta_\ell = \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{112dK^2}} \left(\sum_{\ell'=1}^{\ell} |\mathcal{E}_\ell|\right)^{-1/2},\ \Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2,$$

*and the sub-routine $\mathcal{A}$ chosen as O-FTRL-FP (Eq. (6)). Then under Assumptions 1 and 3, there exists a PBE of agents' joint strategies under Algorithm 2, denoted by $\boldsymbol{\pi}^*$, such that*

$$\begin{aligned}
&\mathfrak{R}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (6)}) \\
&\le 2\sqrt{T}\left(\frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2} + 8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right) \\
&\quad + \sum_{\ell=1}^{L}\left(2N_\ell^2 + (N_\ell+3)(\|\boldsymbol{\rho}^{-1}\|_1 + 1) + 8dM_\ell^2\right) + 3\|\boldsymbol{\rho}^{-1}\|_1, \\
&\textit{and } \mathfrak{B}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (6)}) = 0,
\end{aligned}$$

*where $N_\ell$ and $M_\ell$ is defined as follows for all $\ell \in [L]$.*

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4),$$

$$M_\ell = \ell\log_{\gamma^{-1}}\left(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3 \cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log(dT) + \sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{\rho_j\epsilon}\right).$$

*Specifically, when only focusing on polynomial dependencies on $d$, $K$, and $T$, we have*

$$\begin{aligned}
&\mathfrak{R}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (6)}) = \widetilde{\mathcal{O}}_{d,K,T}(d^2\sqrt{T} + K^4), \\
&\mathfrak{B}_T(\boldsymbol{\pi}^*, \textit{Algorithm 2 with Eq. (6)}) = 0.
\end{aligned}$$

*Proof.* The proof follows the same structure as the previous one, but the treatment of the $\mathbb{E}[\textsc{DualVar}]$ term is extremely challenging and requests delicate analytical tools; we refer the readers to Theorem D.3 for more details. We still include a full proof of this theorem for completeness.

As mentioned in the main text, we introduce an intermediate allocation:

$$\widetilde{i}_t^* := \operatorname*{argmax}_{i\in[K]}\left(v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}\right), \quad \forall \ell \in [L], t \in \mathcal{E}_\ell.$$

Define stopping time $\mathcal{T}_v$ as the last round where no constraint violations can happen:

$$\mathcal{T}_v := \min\left\{t \in [T] \,\middle|\, \sum_{\tau=1}^{t} \boldsymbol{c}_{\tau,i_\tau} + \mathbf{1} \not\le T\boldsymbol{\rho}\right\} \cup \{T+1\}.$$

Thus in rounds $t \le \mathcal{T}_v$, Line 14 never rejects our allocation $i_t$. Decompose the regret $\mathfrak{R}_T$ as

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^{T}(v_{t,i_t^*} - v_{t,i_t})\right] \le \mathbb{E}\left[\underbrace{\sum_{t=1}^{\mathcal{T}_v}(v_{t,\widetilde{i}_t^*} - v_{t,i_t})}_{\textsc{PrimalAlloc}} + \underbrace{\sum_{t=1}^{\mathcal{T}_v}(v_{t,i_t^*} - v_{t,\widetilde{i}_t^*})}_{\textsc{DualVar}} + \underbrace{(T - \mathcal{T}_v)}_{\textsc{NoAct}}\right].$$

For the $\textsc{PrimalAlloc}$ term, we still use Theorem C.5 presented later in Appendix C (which holds regardless to the dual update rules). Theorem C.5 asserts that there exists a PBE $\boldsymbol{\pi}^*$ such that

$$\mathbb{E}[\textsc{PrimalAlloc}] \le \sum_{\ell=1}^{L}(N_\ell + 3),$$

where

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4).$$

For the DUALVAR term, the O-FTRL-FP analysis is substantially harder. This is the second main technical contribution of our paper. We first recall the O-FTRL-FP update rule from Eq. (6):

$$\boldsymbol{\lambda}_\ell = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \sum_{\ell'<\ell}\sum_{\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i_\tau})^\mathsf{T}\boldsymbol{\lambda} + \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda} + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda}),$$

$$\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell) = |\mathcal{E}_\ell| \cdot \frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)}\right),$$

$$\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell) = \operatorname*{argmax}_{i\in[K]} \left(u_{\tau,i} - \boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{\tau,i}\right), \quad \forall \ell' < \ell, \tau \in \mathcal{E}_{\ell'}.$$

We highlight the main technical challenges here and refer the readers to Theorem D.3 for more details:

1. Since $\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)$ is not continuous *w.r.t.* $\boldsymbol{\lambda}_\ell$ due to the argmax, the predicted loss function $\widetilde{F}_\ell(\boldsymbol{\lambda};\boldsymbol{\lambda}_\ell) := \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda}$ is non-continuous *w.r.t.* $\boldsymbol{\lambda}_\ell$ as well – which means an exact fixed point may not exist. In Lemma D.4, we prove that our O-FTRL-FP framework only requires an approximate continuity, which we show ensures the existence of an approximate fixed point.
2. In Lemma D.5, utilizing the smooth cost condition in Assumption 3 and an $\epsilon$-based uniform smoothness argument, we prove that our $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda}$ is indeed approximately continuous.
3. As agents can misreport in epoch $\mathcal{E}_\ell$, the actual epoch-$\ell$ loss function $F_\ell(\boldsymbol{\lambda}) := \sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}$ may differ from the predicted $\widetilde{F}_\ell(\boldsymbol{\lambda};\boldsymbol{\lambda}_\ell)$. We control this effect in Lemma D.6.
4. Due to the unknown distributions, namely $\{\mathcal{V}_i\}_{i\in[K]}$ and $\{\mathcal{C}_i\}_{i\in[K]}$, the planner can only use a finite number of samples (more preciously, $\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|$ ones) in $\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)$. We control the statistical error in Lemma D.7; however, since $\boldsymbol{\lambda}_\ell$ is not measurable in round $\tau$, we develop an $\epsilon$-net based uniform smoothness analysis to ensure the statistical error is small for every possible $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$.
5. Finally, since agents could also misreport in the past, namely epoch $\mathcal{E}_{\ell'}$ where $\ell' < \ell$, the reports used in $\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)$ can also be very different from the true values. In Lemma D.8, we analyze this type of error, again incorporating an $\epsilon$-net based uniform smoothness analysis.

Via careful investigation, Theorem D.3 proves that when configuring

$$\Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2,\ L = \log T,\ \mathcal{E}_\ell = [2^{\ell-1}, \min(2^\ell - 1, T)],\ \eta_\ell = \left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|\right)^{-1/2},$$

we can control the $\mathbb{E}[\text{DUALVAR}]$ as

$$\mathbb{E}[\text{DUALVAR}] \le 2\sqrt{T}\left(\frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2} + 8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right)$$
$$+ \sum_{\ell=1}^{L}\left(2N_\ell^2 + (N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + 8dM_\ell^2\right) + 3\|\boldsymbol{\rho}^{-1}\|_1.$$

where (the definition of $N_\ell$ is the same as that in PRIMALALLOC)

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4),$$

$$M_\ell = \ell\log_{\gamma^{-1}}\left(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3\cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log(dT) + \sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{\rho_j\epsilon}\right).$$

Putting two parts together, we get

$$\mathfrak{R}_T(\boldsymbol{\pi}^*, \text{Algorithm 2 with Eq. (6)}) \le \mathbb{E}[\text{PRIMALALLOC}] + \mathbb{E}[\text{DUALVAR}]$$
$$\le 2\sqrt{T}\left(\frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2} + 8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right)$$

$$+\sum_{\ell=1}^{L}\left(2N_\ell^2+(N_\ell+3)(\|\boldsymbol{\rho}^{-1}\|_1+1)+8dM_\ell^2\right)+3\|\boldsymbol{\rho}^{-1}\|_1.$$

The bound that $\mathfrak{B}_T(\boldsymbol{\pi}^*, \text{Algorithm 2 with Eq. (6)})=0$ directly follows from Line 14 of Algorithm 2.

For the $\widetilde{\mathcal{O}}_{d,K,T}$ version, since $L\le\lceil\log_2 T\rceil=\widetilde{\mathcal{O}}_{d,K,T}(1)$ and $|\mathcal{E}_\ell|\le T$ for all $\ell$, we know $N_\ell=\widetilde{\mathcal{O}}_{d,K,T}(K^2)$ and $M_\ell=\widetilde{\mathcal{O}}_{d,K,T}(\ell)=\widetilde{\mathcal{O}}_{d,K,T}(1)$. This gives $\mathfrak{R}_T=\widetilde{\mathcal{O}}_{d,K,T}(d^2\sqrt{T}+K^4)$. □

# C PrimalAlloc: Regret due to Agents' Strategic Reports

## C.1 IntraEpoch: Agents Lie to Affect Current-Epoch Allocations

**Lemma C.1** (Truthfulness of a Cost-Adjusted Second-Price Auction)**.** *Consider a one-shot monetary allocation setting with $K$ agents. Each agent $i\in[K]$ privately observes their value $v_i\sim\mathcal{V}_i$ and publicly incurs a known cost $c_i\sim\mathcal{C}_i$. Agents submit scalar reports $u_i\in[0,1]$, and the planner allocates the item to one agent $i_t\in[K]$ and charges payment $p_{i_t}$. The utility of the selected agent is $v_{i_t}-p_{i_t}$, while all others receive zero utility.*

*Suppose the planner implements the following mechanism:*

$$i_t=\operatorname*{argmax}_{i\in[K]}(u_i-c_i),\quad p_{i_t}=c_{i_t}+\max_{j\neq i_t}(u_j-c_j).$$

*Then:*

- *(i) For any agent $i\in[K]$ aiming to maximize their expected utility $(v_i-p_i)\cdot\mathbb{1}[i_t=i]$, truthful reporting $u_i=v_i$ is a weakly dominant strategy.*
- *(ii) When all agents report truthfully ($u_i=v_i$), the allocation $i_t=\operatorname{argmax}_{i\in[K]}(v_i-c_i)$, i.e., it maximizes the value-minus-cost $v_i-c_i$ across all the agents.*

*Proof.* Define each agent's pseudo-report as $\widetilde{u}_i:=u_i-c_i$, and let $\widetilde{u}_i^*:=v_i-c_i$ be the truthful pseudo-report. Fix any agent $i\in[K]$ and suppose all other agents' reports $\{u_j\}_{j\neq i}$ are fixed. We evaluate the utility $\mathcal{U}_i$ that agent $i$ obtains under various reporting strategies.

**Case 1: Truthfully report** $u_i=v_i$ so that $\widetilde{u}_i=\widetilde{u}_i^*$:

- If $\widetilde{u}_i^*>\max_{j\neq i}\widetilde{u}_j$, then $i_t=i$, and $\mathcal{U}_i=v_i-c_i-\max_{j\neq i}\widetilde{u}_j=\widetilde{u}_i^*-\max_{j\neq i}\widetilde{u}_j$.
- If $\widetilde{u}_i^*<\max_{j\neq i}\widetilde{u}_j$, then $i_t\neq i$, and $\mathcal{U}_i=0$.

**Case 2: Over-reporting** $u_i>v_i$ so $\widetilde{u}_i>\widetilde{u}_i^*$:

- If $\widetilde{u}_i>\widetilde{u}_i^*>\max_{j\neq i}\widetilde{u}_j, \mathcal{U}_i=\widetilde{u}_i^*-\max_{j\neq i}\widetilde{u}_j$ (same as truthful).
- If $\max_{j\neq i}\widetilde{u}_j>\widetilde{u}_i>\widetilde{u}_i^*, \mathcal{U}_i=0$ (same as truthful).
- Otherwise if $\widetilde{u}_i>\max_{j\neq i}\widetilde{u}_j>\widetilde{u}_i^*, \mathcal{U}_i=\widetilde{u}_i^*-\max_{j\neq i}\widetilde{u}_j<0$ (worse than truthful).

**Case 3: Under-reporting** $u_i<v_i$ so $\widetilde{u}_i<\widetilde{u}_i^*$:

- If $\widetilde{u}_i^*>\widetilde{u}_i>\max_{j\neq i}\widetilde{u}_j, \mathcal{U}_i=\widetilde{u}_i^*-\max_{j\neq i}\widetilde{u}_j$ (same as truthful).
- If $\max_{j\neq i}\widetilde{u}_j>\widetilde{u}_i^*>\widetilde{u}_i, \mathcal{U}_i=0$ (same as truthful).
- Otherwise, if $\widetilde{u}_i^*>\max_{j\neq i}\widetilde{u}_j>\widetilde{u}_i, \mathcal{U}_i=0$. But they originally get $\widetilde{u}_i^*-\max_{j\neq i}\widetilde{u}_j\ge 0$.

In all cases, deviating from truth-telling does not improve agent $i$'s utility, and may strictly reduce it. Hence, truthful reporting is a weakly dominant strategy, proving claim *(i)*.

For claim *(ii)*, when all agents report truthfully ($u_i=v_i$), the planner allocates the resource to $i_t=\arg\max_i(v_i-c_i)$, which maximizes net social value. □

**Theorem C.2** (Intra-Epoch Truthfulness; Formal Theorem 4.1)**.** *Fix any epoch $\mathcal{E}_\ell\subseteq[T]$ and dual variable $\boldsymbol{\lambda}_\ell\in\boldsymbol{\Lambda}$. Suppose that all the agents, when crafting their reports only consider their discounted gains within the current epoch (Model 2 in Eq. (8)), namely Eq. (11). To highlight the*

*different reports, allocations, and payments due to the different agent model, we add a superscript $h$.*

$$\max_{\{u^h_{t,i}\}_{t\in\mathcal{E}_\ell}} \mathbb{E}\left[\sum_{t\in\mathcal{E}_\ell}\gamma^t(v_{t,i}-p^h_{t,i})\cdot\mathbb{1}[i^h_t=i]\right], \quad \forall i\in[K]. \tag{11}$$

*The planner, on the other hand, uses the allocation and payment rule in Line 12. Formally, in round $t\in\mathcal{E}_\ell$ the planner allocates to the agent with maximal dual-adjusted report – which we denote by $i^h_t$ to highlight its difference with $i_t$ due to the different agent model – and sets payments accordingly:*

$$i^h_t=\operatorname*{argmax}_{i\in[K]}(u^h_{t,i}-\boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,i}), \quad p^h_{t,i^h_t}=\boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,i^h_t}+\max_{j\neq i^h_t}(u^h_{t,j}-\boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,j}), \quad \forall t\in\mathcal{E}_\ell.$$

*Then, truthful reporting $u_{t,i}=v_{t,i}$ for all $t\in\mathcal{E}_\ell$ and all $i\in[K]$ – a joint strategy denoted as* TRUTH *– constitutes a Perfect Bayesian Equilibrium (PBE). Furthermore, under this PBE, the planner always chooses the optimal agent according to dual-adjusted value:*

$$i^h_t=\widetilde{i}^*_t:=\arg\max_{i\in[K]}\left(v_{t,i}-\boldsymbol{\lambda}_\ell^\top\boldsymbol{c}_{t,i}\right),$$

*and the regret due to intra-epoch misallocations is zero, i.e.,* $\mathbb{E}[\textsc{IntraEpoch}]=0$.

*Proof.* For any round $t\in\mathcal{E}_\ell$, we observe that the planner's allocation and payment plan $(i^h_t,\boldsymbol{p}^h_t)$ depends only on the current reports $\boldsymbol{u}^h_t$, current costs $\boldsymbol{c}_t$, and the fixed dual variable $\boldsymbol{\lambda}_\ell$. Specifically, it does not depend on historical reports $\{\boldsymbol{u}^h_\tau\}_{\tau<t}$, past allocations $\{i^h_\tau\}_{\tau<t}$, or payments $\{\boldsymbol{p}^h_\tau\}_{\tau<t}$.

Likewise, for any fixed agent $i\in[K]$, suppose that all the opponents follow truthful reporting $\textsc{Truth}_{-i}$, *i.e.*, ($u_{t,j}=v_{t,j}$ for all $t\in\mathcal{E}_\ell$ and $j\neq i$). In this case, for any round $t\in\mathcal{E}_\ell$ whose value is $v_{t,i}\in[0,1]$, the expected gain of any tentative report $u_{t,i}\in[0,1]$ does not depend on the history but only on $\boldsymbol{u}^h_t$, $\boldsymbol{c}_t$, $\boldsymbol{\lambda}_\ell$, and $v_{t,i}$. Therefore, agent $i$ does not benefit from unilaterally deviating to a "history-dependent" strategy that depends on previous public or private information, namely $\{\boldsymbol{u}^h_\tau\}_{\tau<t}$, $\{i^h_\tau\}_{\tau<t}$, $\{\boldsymbol{p}^h_\tau\}_{\tau<t}$, and $\{v_{\tau,i}\}_{\tau<t}$.

Therefore we only need to consider agent $i$'s potential unilateral deviation to history-independent policies, which means we can isolate each round $t\in\mathcal{E}_\ell$. From Lemma C.1, we know that in any such round $t$, given fixed costs and fixed dual $\boldsymbol{\lambda}_\ell$, truthful reporting maximizes an agent's expected utility regardless of opponents' actions. Hence, no agent can benefit from deviating – whether using a history-dependent strategy or a history-independent one – and thus TRUTH is a PBE.

Finally, under PBE TRUTH, the planner allocates to the agent with maximal $v_{t,i}-\boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,i}$, *i.e.*, $i^h_t=\widetilde{i}^*_t$. Therefore, there is no misallocation in the epoch and thus $\textsc{IntraEpoch}=0$. □

### C.2 INTEREPOCH: Agents Lie to Affect Next-Epoch Dual Variables

The key ideas of Lemmas C.3 and C.4 are largely motivated by Golrezaei et al. (2021a, 2023). The main difference is due to the costs $\{\boldsymbol{c}_t\}_{t\in[T]}$, which forbids us from using their results as black-boxes.

For epoch $\ell\in[L]$, consider the "epoch-$\ell$ game with exploration rounds" induced by Lines 4 to 12 in Algorithm 2. Theorem C.2 proves that under Model 2 in Eq. (8) (*i.e.*, agents only care about current-epoch gains) and when there are no exploration rounds, TRUTH is a PBE. Now, we claim that under Model 1 in Eq. (8) (*i.e.*, agents optimize over the whole future) and the actual mechanism with exploration rounds (Lines 4 to 12 in Algorithm 2), there exists a PBE of agents' joint strategies $\boldsymbol{\pi}$ that is not too far from TRUTH. Formally, we present Lemma C.3.

**Lemma C.3** (Large Misreport Happens Rarely)**.** *For an epoch $\ell\in[L]$, consider the agent Model 1 in Eq. (8) and the "epoch-$\ell$ game with exploration rounds" specified by Lines 4 to 12 in Algorithm 2. There exists a PBE of agents' joint strategies $\boldsymbol{\pi}$, such that for all $i\in[K]$,*

$$\Pr\left\{\sum_{t\in\mathcal{E}_\ell}\mathbb{1}\left[|u_{t,i}-v_{t,i}|\geq\frac{1}{|\mathcal{E}_\ell|}\right]\leq\log_{\gamma^{-1}}(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4)\right\}\geq 1-\frac{1}{|\mathcal{E}_\ell|},$$

*where $\{u_{t,i}\}_{t\in\mathcal{E}_\ell}$ are the reports made by agent $i$ under $\boldsymbol{\pi}$.*

*Proof.* Consider the history-independent auxiliary game defined in the proof of Theorem C.2, where for every round $t \in \mathcal{E}_\ell$, agent $i \in [K]$ is only allowed to craft their reports $u_{t,i}$ based on current-epoch dual variable $\boldsymbol{\lambda}_\ell$, round number $t$, current-round private value $v_{t,i}$, and current-round public costs $\boldsymbol{c}_t$.

Let $\boldsymbol{\pi}$ be a PBE in this history-independent auxiliary game. Using same arguments from Theorem C.2, unilaterally deviating to a history-dependent strategy is not beneficial as all opponents' strategies $\boldsymbol{\pi}_{-i}$ and the mechanism (Lines 4 to 12 in Algorithm 2) are history-independent. Hence, $\boldsymbol{\pi}$ remains a PBE in the actual "epoch-$\ell$ game with exploration rounds" where history dependency is allowed.

To prove the claim for this PBE $\boldsymbol{\pi}$, consider the unilateral deviation of any agent $i \in [K]$ to the truth-telling policy, *i.e.*, $\boldsymbol{\pi}^i := \textsc{Truth}_i \circ \boldsymbol{\pi}_{-i}$. Since $\boldsymbol{\pi}$ is a PBE, $\boldsymbol{\pi}^i$ is no better than $\boldsymbol{\pi}$ under Model 1 (the actual model). That is, for $s_\ell := \min\{t \mid t \in \mathcal{E}_\ell\}$ and any history $\mathcal{H}_{s_\ell}$, we always have

$$
\begin{aligned}
0 &\le \mathop{\mathbb{E}}_{\boldsymbol{\pi}}\left[\sum_{\tau=t}^{T}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\,\middle|\,\mathcal{H}_{s_\ell}\right] - \mathop{\mathbb{E}}_{\boldsymbol{\pi}^i}\left[\sum_{\tau=t}^{T}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\,\middle|\,\mathcal{H}_{s_\ell}\right] \\
&= \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\pi}}\left[\sum_{\tau\in\mathcal{E}_\ell}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\,\middle|\,\mathcal{H}_{s_\ell}\right] - \mathop{\mathbb{E}}_{\boldsymbol{\pi}^i}\left[\sum_{\tau\in\mathcal{E}_\ell}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\,\middle|\,\mathcal{H}_{s_\ell}\right]}_{\text{Current-Epoch Difference}} + \\
&\quad \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\pi}}\left[\sum_{\ell'>\ell,\tau\in\mathcal{E}_{\ell'}}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\,\middle|\,\mathcal{H}_{s_\ell}\right] - \mathop{\mathbb{E}}_{\boldsymbol{\pi}^i}\left[\sum_{\ell'>\ell,\tau\in\mathcal{E}_{\ell'}}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\,\middle|\,\mathcal{H}_{s_\ell}\right]}_{\text{Future-Epoch Difference}}.
\end{aligned}
\tag{12}
$$

For the second term, fix any $\ell' > \ell$ and $\tau \in \mathcal{E}_{\ell'}$. Since the values $\boldsymbol{v}_\tau$, reports $\boldsymbol{u}_\tau$, and costs $\boldsymbol{c}_\tau$ are all bounded by $[0,1]$, and that $\boldsymbol{\lambda}_{\ell'} \in \boldsymbol{\Lambda} = \bigotimes_{j=1}^{d}[0,\rho_j^{-1}]$ (which infers $\|\boldsymbol{\lambda}_{\ell'}\|_1 \le \|\boldsymbol{\rho}^{-1}\|_1$), we have

$$
p_{\tau,i_\tau} = \boldsymbol{\lambda}_{\ell'}^\intercal \boldsymbol{c}_{\tau,i_\tau} + \max_{j\neq i_\tau}(u_{\tau,j} - \boldsymbol{\lambda}_{\ell'}^\intercal \boldsymbol{c}_{\tau,j}) \in \left[-2\|\boldsymbol{\lambda}_{\ell'}\|_1 \cdot \max_{j\in[K]}\|\boldsymbol{c}_{\tau,j}\|_\infty,\, 1+2\|\boldsymbol{\lambda}_{\ell'}\|_1 \cdot \max_{j\in[K]}\|\boldsymbol{c}_{\tau,j}\|_\infty\right],
$$

for all $\ell' > \ell, \tau \in \mathcal{E}_{\ell'}$. Since $v_{t,i} \in [0,1]$, this suggests that $v_{\tau,i} - p_{\tau,i} \in [-1-2\|\boldsymbol{\rho}^{-1}\|_1, 1+2\|\boldsymbol{\rho}^{-1}\|_1]$ for all $i \in [K]$, and thus

$$
\text{Future-Epoch Difference} \le \sum_{\ell'=\ell+1}^{L}\sum_{\tau\in\mathcal{E}_{\ell'}}\gamma^\tau \cdot 2(1+2\|\boldsymbol{\rho}^{-1}\|_1) \le 2(1+2\|\boldsymbol{\rho}^{-1}\|_1)\frac{\gamma^{s_{\ell+1}}}{1-\gamma},
$$

where the second inequality uses the fact that $\sum_{\tau=s_{\ell+1}}^{T}\gamma^\tau \le \frac{\gamma^{s_{\ell+1}}}{1-\gamma}$.

We now focus on the current-epoch difference part. Since both $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^i$ are history independent (recall that $\boldsymbol{\pi}$ is a joint strategy from the history-independent auxiliary game and $\boldsymbol{\pi}^i = \textsc{Truth}_i \circ \boldsymbol{\pi}_{-i}$), the conditioning on $\mathcal{H}_{s_\ell}$ is redundant and consequently

$$
\text{Current-Epoch Difference} = \mathop{\mathbb{E}}_{\boldsymbol{\pi}}\left[\sum_{\tau\in\mathcal{E}_\ell}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\right] - \mathop{\mathbb{E}}_{\boldsymbol{\pi}^i}\left[\sum_{\tau\in\mathcal{E}_\ell}\gamma^\tau(v_{\tau,i}-p_{\tau,i})\mathbb{1}[i_\tau=i]\right].
$$

For any round $t \in \mathcal{E}_\ell$, we control this difference utilizing the exploration rounds:

- Suppose that round $t \in \mathcal{E}_\ell$ is an exploration round for agent $i$ (which happens with probability $\frac{1}{|\mathcal{E}_\ell|}\cdot\frac{1}{K}$), the expected gain of reporting $u_{t,i}$ rather than $v_{t,i}$ (*i.e.*, under $\boldsymbol{\pi}$ versus under $\boldsymbol{\pi}^i$) is
$$
\mathop{\mathbb{E}}_{p\in\text{Unif}([0,1])}[(v_{t,i}-p)\mathbb{1}[u_{t,i}\ge p] - (v_{t,i}-p)\mathbb{1}[v_{t,i}\ge p]] = -\frac{1}{2}(u_{t,i}-v_{t,i})^2.
$$
- If round $t$ is an exploration round but not for agent $i$, then reporting $u_{t,i}$ and $v_{t,i}$ both give 0 gain.
- Finally, suppose that round $t \in \mathcal{E}_\ell$ is not an exploration round. Notice that *i)* both $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_{-i}$ are history-independent, and *ii)* if round $t \in \mathcal{E}_\ell$ is not an exploration round, then the epoch-$\ell$ game in Lines 4 to 12 in Algorithm 2 coincides with Line 12. Therefore, via Lemma C.1, the gain of reporting $u_{t,i}$ is no larger than that of reporting $v_{t,i}$, *i.e.*, the expectation difference is non-positive.

Let $\mathcal{M}_{\ell,i} := \left\{ t \in \mathcal{E}_\ell \mid |u_{t,i} - v_{t,i}| \geq \frac{1}{|\mathcal{E}_\ell|} \right\}$ be the set of large misreports from agent $i$ (regardless of whether $t$ turns out to be an exploration round, since reports happen before it). Let the last round in epoch $\mathcal{E}_\ell$ be $e_\ell := \max\{t \mid t \in \mathcal{E}_\ell\}$. Since $\sum_{t \in \mathcal{M}_{\ell,i}} \gamma^t \geq \sum_{t=e_\ell - |\mathcal{M}_{\ell,i}|+1}^{e_\ell} \gamma^t$, we have

$$\begin{aligned}
\text{Current-Epoch Difference} &\leq -\mathbb{E}\left[ \sum_{t \in \mathcal{M}_{\ell,i}} \gamma^t \frac{1}{|\mathcal{E}_\ell| \cdot K} \frac{1}{2} (u_{t,i} - v_{t,i})^2 \right] \\
&\leq -\mathbb{E}\left[ \frac{\gamma^{e_\ell - |\mathcal{M}_{\ell,i}|+1}(1 - \gamma^{|\mathcal{M}_{\ell,i}|})}{1-\gamma} \frac{1}{|\mathcal{E}_\ell| \cdot K} \frac{1}{2|\mathcal{E}_\ell|^2} \right].
\end{aligned}$$

In order for $\boldsymbol{\pi}^i$ to be inferior when compared with $\boldsymbol{\pi}$, *i.e.*, Eq. (12) holds, we therefore must have

$$\begin{aligned}
2(1 + 2\|\boldsymbol{\rho}^{-1}\|_1) \frac{\gamma^{s_{\ell+1}}}{1-\gamma} &\geq \mathbb{E}\left[ \frac{\gamma^{e_\ell - |\mathcal{M}_{\ell,i}|+1}(1 - \gamma^{|\mathcal{M}_{\ell,i}|})}{1-\gamma} \frac{1}{|\mathcal{E}_\ell| \cdot K} \frac{1}{2|\mathcal{E}_\ell|^2} \right] \\
&\geq \Pr\{|\mathcal{M}_{\ell,i}| \geq c\} \frac{\gamma^{s_{\ell+1}}}{1-\gamma} \frac{\gamma^{-c} - 1}{2K|\mathcal{E}_\ell|^3}, \quad \forall c > 0, \qquad (13)
\end{aligned}$$

where the last step uses the fact that $s_{\ell+1} = e_\ell + 1$.

Picking $c$ such that $2(1 + 2\|\boldsymbol{\rho}^{-1}\|_1) \frac{2K|\mathcal{E}_\ell|^3}{\gamma^{-c}-1} = \frac{1}{|\mathcal{E}_\ell|}$, we reach our conclusion that

$$\Pr\left\{ |\mathcal{M}_{\ell,i}| \geq \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1) K |\mathcal{E}_\ell|^4) \right\} \leq \frac{1}{|\mathcal{E}_\ell|}.$$

This completes the proof. □

Still focusing on a fixed epoch $\ell \in [L]$, Lemma C.3 bounds the number of rounds with large misreports. We now turn to the remaining rounds, *i.e.*, those $t \in \mathcal{E}_\ell$ such that $|u_{t,i} - v_{t,i}| \leq \frac{1}{|\mathcal{E}_\ell|}$ for all $i \in [K]$. We claim that misallocations are rare among these rounds, formalized in Lemma C.4.

**Lemma C.4** (Misallocation with Small Misreports Happens Rarely)**.** *Consider an epoch $\ell \in [L]$ with dual variable $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$. We have*

$$\begin{aligned}
\Pr\Bigg\{ &\sum_{t \in \mathcal{E}_\ell} \mathbb{1}\left[ \operatorname*{argmax}_{i \in [K]} (u_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \neq \operatorname*{argmax}_{i \in [K]} (v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \right] \mathbb{1}\left[ |u_{t,i} - v_{t,i}| \leq \frac{1}{|\mathcal{E}_\ell|}, \forall i \in [K] \right] \\
&\leq 4K^2 \epsilon_c \|\boldsymbol{\lambda}_\ell\|_1 + 4\log|\mathcal{E}_\ell| \Bigg\} \geq 1 - \frac{2}{|\mathcal{E}_\ell|}.
\end{aligned}$$

*Proof.* For any round $t \in \mathcal{E}_\ell$, we bound the probability that the allocation based on reported utilities differs from that based on true values, even when reports are close to truthful. Specifically, consider the event

$$\operatorname*{argmax}_{i \in [K]} (u_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \neq \operatorname*{argmax}_{i \in [K]} (v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \quad \text{and} \quad |u_{t,i} - v_{t,i}| \leq \frac{1}{|\mathcal{E}_\ell|}, \ \forall i \in [K].$$

Such a mismatch can only happen if there exists a pair of indices $i \neq j$ whose true dual-adjusted values are very close – within $\frac{2}{|\mathcal{E}_\ell|}$ – so that small deviations in reported utilities (bounded by $\frac{1}{|\mathcal{E}_\ell|}$) are able to flip the argmax decision. We apply a union bound over all such pairs to upper bound this probability:

$$\begin{aligned}
&\Pr\left\{ \operatorname*{argmax}_{i \in [K]} (u_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \neq \operatorname*{argmax}_{i \in [K]} (v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \wedge |u_{t,i} - v_{t,i}| \leq \frac{1}{|\mathcal{E}_\ell|} \right\} \\
\leq &\sum_{1 \leq i < j \leq K} \Pr\left\{ |(v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) - (v_{t,j} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,j})| \leq \frac{2}{|\mathcal{E}_\ell|} \right\} \leq \frac{2K^2}{|\mathcal{E}_\ell|} \epsilon_c \|\boldsymbol{\lambda}_\ell\|_1, \qquad (14)
\end{aligned}$$

where the second inequality comes from Assumption 3: Since it assumes that $\mathrm{PDF}(\boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i})$ is uniformly upper bounded by $\epsilon_c$ where $\boldsymbol{\lambda}_\ell \in \boldsymbol{\Lambda}$, we know

$$\Pr_{\boldsymbol{c}_{t,i}\sim\mathcal{C}_i}\left\{-\frac{2}{|\mathcal{E}_\ell|} \le \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i} - (v_{t,i} - v_{t,j} + \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,j}) \le \frac{2}{|\mathcal{E}_\ell|}\right\} \le \frac{4}{|\mathcal{E}_\ell|}\epsilon_c.$$

Now consider the martingale difference sequence $\{X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]\}_{t\in\mathcal{E}_\ell}$ where

$$X_t := \mathbb{1}\left[\operatorname*{argmax}_{i\in[K]}(u_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i}) \neq \operatorname*{argmax}_{i\in[K]}(v_{t,i} - \boldsymbol{\lambda}_\ell^\intercal \boldsymbol{c}_{t,i})\right] \mathbb{1}\left[|u_{t,i} - v_{t,i}| \le \frac{1}{|\mathcal{E}_\ell|}\right],$$

and $(\mathcal{F}_t)_t$ is the natural filtration defined as $\mathcal{F}_t = \sigma(X_1, X_2, \ldots, X_t)$.

We apply the multiplicative Azuma-Hoeffding inequality restated as Lemma E.4 (Koufogiannakis and Young, 2014, Lemma 10) with $Y_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$, $\epsilon = \frac{1}{2}$, and $A = 2\log|\mathcal{E}_\ell|$. Since $X_t \in [0,1]$ *a.s.* and $\mathbb{E}[X_t - Y_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}] = 0$, the two conditions in Lemma E.4 hold and thus

$$\Pr\left\{\frac{1}{2}\sum_{t\in\mathcal{E}_\ell} X_t \ge \sum_{t\in\mathcal{E}_\ell}\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] + 2\log|\mathcal{E}_\ell|\right\} \le \exp(-\log|\mathcal{E}_\ell|).$$

From Eq. (14), we know $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \Pr\{X_t \mid \mathcal{F}_{t-1}\} \le \frac{2K^2}{|\mathcal{E}_\ell|}\epsilon_c\|\boldsymbol{\lambda}_\ell\|_1$. Therefore, rearranging the above concentration result gives

$$\Pr\left\{\sum_{t\in\mathcal{E}_\ell} X_t \le 4K^2\epsilon_c\|\boldsymbol{\lambda}_\ell\|_1 + 4\log|\mathcal{E}_\ell|\right\} \ge 1 - \frac{1}{|\mathcal{E}_\ell|}.$$

Plugging back the definition of $X_t$ completes the proof. □

Putting the previous two parts together for all $\ell \in [L]$ gives the following theorem.

**Theorem C.5** (INTEREPOCH Guarantee; Formal Theorem 4.2)**.** *Under the mechanism specified in Algorithm 2, there exists a PBE of agents' strategies $\boldsymbol{\pi}^*$, such that the* PRIMALALLOC *term (which is the sum of* INTRAEPOCH *and* INTEREPOCH *terms) is bounded as*

$$\mathbb{E}[\textsc{PrimalAlloc}] = \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v}(v_{t,\widetilde{i}_t^*} - v_{t,i_t})\right] \le \sum_{\ell=1}^{L}(N_\ell + 3),$$

*where* $N_\ell := 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4), \forall \ell \in [L]$.

*Proof.* Applying Lemma C.3 to every epoch $\ell \in [L]$, we get a PBE $\boldsymbol{\pi}_\ell$ for every "epoch-$\ell$ game with exploration rounds" (Lines 4 to 12 in Algorithm 2). By definition of $\mathcal{T}_v$, the safety constraint is never violated before that and thus Line 14 has no effect. Furthermore, since Algorithm 2's allocations and payments within every epoch $\ell \in [L]$ only directly depend on the current dual variable $\boldsymbol{\lambda}_\ell$ but not anything else from the past, using the same auxiliary game arguments as in Theorem C.2, there is a PBE $\boldsymbol{\pi}^*$ for the whole game under mechanism Algorithm 2 that matches $(\boldsymbol{\pi}_\ell)_{\ell\in[L]}$ up to $\mathcal{T}_v$.

For every epoch $t \in \mathcal{E}_\ell$, it can be either *i)* an exploration round, which happens *w.p.* $\frac{1}{|\mathcal{E}_\ell|}$ independently, *ii)* a standard round with large misreports: $\exists i \in [K]$ such that $|u_{t,i} - v_{t,i} \ge \frac{1}{|\mathcal{E}_\ell|}$, or *iii)* a standard round with only small misreports. For *ii)*, we apply Lemma C.3; for *iii)*, we apply Lemma C.4. For *i)*, applying Chernoff inequality,

$$\Pr\left\{\sum_{t\in\mathcal{E}_\ell}\mathbb{1}[t \text{ exploration round}] > 1 + c\right\} \le \exp\left(-\frac{c^2}{2 + 2c/3}\right), \quad \forall c > 0.$$

Setting $c = \log|\mathcal{E}_\ell|$ so that the RHS is no more than $\frac{1}{|\mathcal{E}_\ell|}$, we get

$$\Pr\left\{\sum_{t\in\mathcal{E}_\ell}\mathbb{1}[t \text{ exploration round}] \le 1 + \log|\mathcal{E}_\ell|\right\} \ge 1 - \frac{1}{|\mathcal{E}_\ell|}. \tag{15}$$

Now we put the aforementioned three cases together:

$$\begin{aligned}
\mathbb{E}[\textsc{PrimalAlloc}] &\overset{(a)}{\leq} \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \mathbb{1}[\widetilde{i}_t^* \neq i_t]\right] \\
&\overset{(b)}{\leq} \sum_{\ell=1}^{L} \mathbb{E}\left[\sum_{t\in\mathcal{E}_\ell}\left(\mathbb{1}[t \text{ exploration round}] + \mathbb{1}\left[\exists i\in[K], |u_{t,i}-v_{t,i}| \geq \frac{1}{|\mathcal{E}_\ell|}\right]\right.\right. \\
&\qquad \left.\left. + \mathbb{1}\left[\operatorname*{argmax}_{i\in[K]}(u_{t,i}-\boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,i}) \neq \operatorname*{argmax}_{i\in[K]}(v_{t,i}-\boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{t,i}) \wedge |u_{t,i}-v_{t,i}| \leq \frac{1}{|\mathcal{E}_\ell|}, \forall i\in[K]\right]\right)\right] \\
&\overset{(c)}{\leq} \sum_{\ell=1}^{L}\left(1+\log|\mathcal{E}_\ell| + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4) + \frac{3|\mathcal{E}_\ell|}{|\mathcal{E}_\ell|}\right) \qquad (16) \\
&= \sum_{\ell=1}^{L}(N_\ell+3), \quad N_\ell := 1+4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4),
\end{aligned}$$

where (a) uses the fact that $v_{t,i}\in[0,1]$ for all $t$ and $i$, (b) uses the above discussions of *(i)*, *(ii)*, and *(iii)*, and (c) applied Line 9 to *(i)*, Lemma C.3 to *(ii)*, Lemma C.4 to *(iii)*, and the trivial bound that $\sum_{t\in\mathcal{E}_\ell}\mathbb{1}[\widetilde{i}_t^* \neq i_t] \leq |\mathcal{E}_\ell|$ if any of the conclusions in Line 9 or Lemmas C.3 and C.4 do not hold (every conclusion holds with probability $1-\frac{1}{|\mathcal{E}_\ell|}$, and thus a Union Bound controls the overall failure probability by $\frac{3}{|\mathcal{E}_\ell|}$). This completes the proof. □

# D DualVar: Regret due to Primal-Dual Framework

## D.1 DualVar and Online Learning Regret

**Lemma D.1** (DualVar and Online Learning Regret; Formal Lemma 4.3)**.** *Under Algorithm 2, the* DualVar *term can be controlled as follows:*

$$\begin{aligned}
\mathbb{E}[\textsc{DualVar}] &= \mathbb{E}\left[\sum_{t=1}^{T} v_{t,i_t^*} - \sum_{t=1}^{\mathcal{T}_v} v_{t,\widetilde{i}_t^*}\right] \\
&\leq \mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^*)\right] + \sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + \|\boldsymbol{\rho}^{-1}\|_1,
\end{aligned}$$

*where*

$$N_\ell = 1+4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4).$$

*Proof.* This lemma largely adopts Theorem 1 of Balseiro et al. (2023), but we incorporate some arguments of Castiglioni et al. (2022) to fix a measurability issue in the original proof.

Slightly abusing the notations, for any round $t$ belonging to an epoch $\mathcal{E}_\ell$, we define $\boldsymbol{\lambda}_t := \boldsymbol{\lambda}_\ell$. Let $(\mathcal{F}_t)_{t\geq 0}$ be the filtration specified as $\mathcal{F}_t = \sigma(\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_t,\boldsymbol{v}_1,\ldots,\boldsymbol{v}_t,\boldsymbol{c}_1,\ldots,\boldsymbol{c}_t)$.

For any $t\in[T]$, let $v_t^*$ be the convex conjugate (Fenchel dual) function of $\boldsymbol{v}_t$, namely

$$v_t^*(\boldsymbol{\lambda}) = \max_{i\in[K]}(v_{t,i}-\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{t,i}), \quad \forall\boldsymbol{\lambda}\in\boldsymbol{\Lambda},$$

which is convex in $\boldsymbol{\lambda}$ since it's the maximum of convex functions. We further define $v^*(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{v}\sim\mathcal{V},\boldsymbol{c}\sim\mathcal{C}}[\max_{i\in[K]}(v_i-\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_i)]$, which is an expectation of a convex function and thus also convex.

**Step 1: Lower bound the values collected by $\{\widetilde{i}_t^*\}_{t\in[\mathcal{T}_v]}$.** By definition of $\widetilde{i}_t^*$ from Eq. (9) that $\widetilde{i}_t^* = \operatorname{argmax}_{i\in[K]}(v_{t,i}-\boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,i})$, we have

$$v_{t,\widetilde{i}_t^*} = v_t^*(\boldsymbol{\lambda}_t) + \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,\widetilde{i}_t^*}, \quad \forall t\leq\mathcal{T}_v.$$

Since $\boldsymbol{\lambda}_t$ is $\mathcal{F}_{t-1}$-measurable but $\boldsymbol{v}_t$ and $\boldsymbol{c}_t$ are sampled from $\mathcal{V}$ and $\mathcal{C}$ independently to $\mathcal{F}_{t-1}$,

$$\mathbb{E}\left[v_{t,\widetilde{i}_t^*} \,\middle|\, \mathcal{F}_{t-1}\right] = \mathbb{E}\left[v_t^*(\boldsymbol{\lambda}_t) + \boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*} \,\middle|\, \mathcal{F}_{t-1}\right] = v^*(\boldsymbol{\lambda}_t) + \mathbb{E}\left[\boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*} \,\middle|\, \mathcal{F}_{t-1}\right].$$

Put this equality in another way, the stochastic process $(X_t)_{t\geq 1}$ adapted to $(\mathcal{F}_t)_{t\geq 0}$ defined as

$$X_t := v_{t,\widetilde{i}_t^*} - \boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*} - v^*(\boldsymbol{\lambda}_t), \quad \forall t \leq \mathcal{T}_v,$$

ensures $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$ and is thus a martingale difference sequence. Since $\mathcal{T}_v \leq T+1$ *a.s.* by definition from Eq. (10), Optional Stopping Time theorem (Williams, 1991, Theorem 10.10) gives

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} X_t\right] = 0 \Longrightarrow \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} v_{t,\widetilde{i}_t^*}\right] = \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} v^*(\boldsymbol{\lambda}_t) + \sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*}\right].$$

We now utilize the convexity of $v^*$, which is the expectation of a convex function, and conclude that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} v_{t,\widetilde{i}_t^*}\right] &= \mathbb{E}\left[\mathcal{T}_v \cdot \frac{1}{\mathcal{T}_v}\sum_{t=1}^{\mathcal{T}_v} v^*(\boldsymbol{\lambda}_t) + \sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*}\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[\mathcal{T}_v \cdot v^*\left(\frac{1}{\mathcal{T}_v}\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t\right) + \sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*}\right] \\
&\overset{(b)}{\geq} \mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} v^*(\boldsymbol{\lambda}) + \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\intercal \boldsymbol{c}_{t,\widetilde{i}_t^*}\right],
\end{aligned} \tag{17}$$

where (a) uses Jensen's inequality and (b) uses the fact that $\boldsymbol{\lambda} = \frac{1}{\mathcal{T}_v}\sum_{\tau=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_\tau \in \boldsymbol{\Lambda}$ because $\boldsymbol{\lambda}_\tau \in \boldsymbol{\Lambda}$ for all $\tau$ and the fact that $\boldsymbol{\Lambda} = \bigotimes_{j=1}^d [0, \rho_j^{-1}]$ is convex.

**Step 2: Upper bound the offline optimal social welfare $\sum_{t=1}^T v_{t,i_t^*}$.** We now work on the other term in $\mathbb{E}[\textsc{DualVar}]$, namely $\mathbb{E}[\sum_{t=1}^T v_{t,i_t^*}]$. For any fixed $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, since $v_t^*(\boldsymbol{\lambda}) = \max_{i\in[K]}(v_{t,i} - \boldsymbol{\lambda}^\intercal \boldsymbol{c}_{t,i}) \geq v_{t,i_t^*} - \boldsymbol{\lambda}^\intercal \boldsymbol{c}_{t,i_t^*}$ for all $t \in [T]$, we have

$$\sum_{t=1}^T v_{t,i_t^*} \leq \sum_{t=1}^T v_t^*(\boldsymbol{\lambda}) + \sum_{t=1}^T \boldsymbol{\lambda}^\intercal \boldsymbol{c}_{t,i_t^*}, \quad \forall \boldsymbol{\lambda} \in \boldsymbol{\Lambda}. \tag{18}$$

Further recall from Eq. (2) that $\{i_t^*\}_{t\in[T]}$ is the optimum of the offline optimization problem

$$\max \sum_{t=1}^T v_{t,i_t^*} \quad \text{s.t.} \ \frac{1}{T}\sum_{t=1}^T \boldsymbol{c}_{t,i_t^*} \leq \boldsymbol{\rho},$$

it ensures $\frac{1}{T}\sum_{t=1}^T \boldsymbol{c}_{t,i_t^*} \leq \boldsymbol{\rho}$. Therefore, for any fixed $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, we further know

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathcal{T}_v}{T}\sum_{t=1}^T v_{t,i_t^*}\right] &\overset{(a)}{\leq} \mathbb{E}\left[\frac{\mathcal{T}_v}{T}\left(\sum_{t=1}^T v_t^*(\boldsymbol{\lambda}) + \sum_{t=1}^T \boldsymbol{\lambda}^\intercal \boldsymbol{c}_{t,i_t^*}\right)\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{\mathcal{T}_v}{T}\sum_{t=1}^T v_t^*(\boldsymbol{\lambda}) + \mathcal{T}_v \boldsymbol{\lambda}^\intercal \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{c}_{t,i_t^*}\right)\right] \\
&\overset{(c)}{\leq} \mathbb{E}\left[\mathcal{T}_v \cdot \frac{1}{T}\sum_{t=1}^T v_t^*(\boldsymbol{\lambda}) + \mathcal{T}_v \boldsymbol{\lambda}^\intercal \boldsymbol{\rho}\right], \\
&\overset{(d)}{=} \mathbb{E}\left[\mathcal{T}_v\right](v^*(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^\intercal \boldsymbol{\rho}), \quad \forall \boldsymbol{\lambda} \in \boldsymbol{\Lambda},
\end{aligned}$$

where (a) uses Eq. (18), (b) rearranges the terms, (c) uses $\frac{1}{T}\sum_{t=1}^T \boldsymbol{c}_{t,i_t^*} \leq \boldsymbol{\rho}$, (d) uses the definition that $v^*(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{v}\sim\mathcal{V},\boldsymbol{c}\sim\mathcal{C}}[\max_{i\in[K]}(v_i - \boldsymbol{\lambda}^\intercal \boldsymbol{c}_i)]$ (and thus $\mathbb{E}[v_t(\boldsymbol{\lambda})] = v^*(\boldsymbol{\lambda})$ for any fixed $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$).

Taking infimum of $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and further recalling that all the values $v_{t,i_t^*}$ are $[0,1]$-bounded, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} v_{t,i_t^*}\right] &\le \mathbb{E}\left[\frac{\mathcal{T}_v}{T}\sum_{t=1}^{T} v_{t,i_t^*} + (T-\mathcal{T}_v)\right] \\
&\le \mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} (v^*(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho}) + \mathbb{E}[T-\mathcal{T}_v] \\
&\le \mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} v^*(\boldsymbol{\lambda}) + \mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho} + \mathbb{E}[T-\mathcal{T}_v]. \qquad (19)
\end{aligned}$$

**Step 3: Combine Steps 1 and 2.** Putting Eqs. (17) and (19) together, we get

$$\begin{aligned}
\mathbb{E}[\textsc{DualVar}] &= \mathbb{E}\left[\sum_{t=1}^{T} v_{t,i_t^*} - \sum_{t=1}^{\mathcal{T}_v} v_{t,\widetilde{i}_t^*}\right] \\
&\le \left(\mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} v^*(\boldsymbol{\lambda}) + \mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho} + \mathbb{E}[T-\mathcal{T}_v]\right) \\
&\quad - \left(\mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} v^*(\boldsymbol{\lambda}) + \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,\widetilde{i}_t^*}\right]\right) \\
&= \left(\mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho} - \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,\widetilde{i}_t^*}\right]\right) + \mathbb{E}[T-\mathcal{T}_v].
\end{aligned}$$

Consider the following stochastic process $(Y_t)_{t\ge 1}$ adapted to $(\mathcal{F}_t)_{t\ge 0}$:

$$Y_t := \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho} - \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{\rho}, \quad \forall t \le \mathcal{T}_v,$$

we must have $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \le 0$ since $\boldsymbol{\lambda}_t \in \boldsymbol{\Lambda}$, which means $(Y_t)_{t\ge 1}$ is a super-martingale difference sequence. Again utilizing the fact that $\mathcal{T}_v \le (T+1)$ *a.s.* and the Optional Stopping Time theorem (Williams, 1991, Theorem 10.10), we know

$$\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} Y_t\right] \le 0 \Longrightarrow \mathbb{E}[\mathcal{T}_v] \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho} = \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \inf_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\rho}\right] \le \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{\rho}\right].$$

This reveals that

$$\begin{aligned}
\mathbb{E}[\textsc{DualVar}] &\le \left(\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{\rho}\right] - \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,\widetilde{i}_t^*}\right]\right) + \mathbb{E}[T-\mathcal{T}_v] \\
&= \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*})\right] + \mathbb{E}[T-\mathcal{T}_v]. \qquad (20)
\end{aligned}$$

Comparing Eq. (20) to our conclusion, it only remains to associate $\boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,\widetilde{i}_t^*}$ with $\boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{c}_{t,i_t}$ and control $\mathbb{E}[T-\mathcal{T}_v]$. We first focus on the former objective.

**Step 4: Relate $\boldsymbol{\lambda}_t^\mathsf{T}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*})$ to $\boldsymbol{\lambda}_t^\mathsf{T}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})$.** Using Eq. (16) from Theorem C.5, with probability $1 - \frac{3}{|\mathcal{E}_\ell|}$, the sequence $\{\widetilde{i}_t^*\}_{t\in\mathcal{E}_\ell}$ and $\{i_t\}_{t\in\mathcal{E}_\ell}$ only differs by no more than $N_\ell$, where

$$N_\ell := 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4), \quad \forall \ell \in [L].$$

That is, we have shown than $\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \mathbb{1}[\widetilde{i}_t^* \ne i_t]\right] \le \sum_{\ell=1}^{L}(N_\ell + 3)$ where the 3 comes from the $\frac{3}{|\mathcal{E}_\ell|}$ failure probability and the fact that $\sum_{t\in\mathcal{E}_\ell} \mathbb{1}[\widetilde{i}_t^* \ne i_t] \le |\mathcal{E}_\ell|$. Combining it with the observation that

$$|\boldsymbol{c}_{t,i}^\mathsf{T}\boldsymbol{\lambda}_t - \boldsymbol{c}_{t,j}^\mathsf{T}\boldsymbol{\lambda}_t| \le \|\boldsymbol{\lambda}_t\|_1 \cdot \|\boldsymbol{c}_{t,i} - \boldsymbol{c}_{t,i}\|_1 \le \|\boldsymbol{\rho}^{-1}\|_1, \quad \forall i \ne j,$$

where we shall recall that $\boldsymbol{c}_{t,i}, \boldsymbol{c}_{t,j} \in [0,1]^d$ and that $\boldsymbol{\lambda}_t \in \boldsymbol{\Lambda} = \bigotimes_{j=1}^{d}[0, \rho_j^{-1}]$, Eq. (20) gives

$$\mathbb{E}[\textsc{DualVar}] \le \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v} \boldsymbol{\lambda}_t^\mathsf{T}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*})\right] + \mathbb{E}[T-\mathcal{T}_v]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\lambda}_t^\mathsf{T}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})+\|\boldsymbol{\rho}^{-1}\|_1\cdot\mathbb{1}[\widetilde{i}_t^*\neq i_t])\right]+\mathbb{E}[T-\mathcal{T}_v]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v}\boldsymbol{\lambda}_t^\mathsf{T}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})\right]+\|\boldsymbol{\rho}^{-1}\|_1\sum_{\ell=1}^{L}(N_\ell+3)+\mathbb{E}[T-\mathcal{T}_v]. \tag{21}$$

**Step 5: Control $\mathbb{E}[T-\mathcal{T}_v]$.** We recall the definition of $\mathcal{T}_v$ from Eq. (10):

$$\mathcal{T}_v := \min\left\{t\in[T]\;\middle|\;\sum_{\tau=1}^{t}\boldsymbol{c}_{\tau,i_\tau}+\mathbf{1}\not\leq T\boldsymbol{\rho}\right\}\cup\{T+1\}.$$

If $\mathcal{T}_v = T+1$, then $(T-\mathcal{T}_v)$ is trivially bounded. Otherwise, suppose that $\sum_{\tau=1}^{\mathcal{T}_v}\boldsymbol{c}_{\tau,i_\tau}+\mathbf{1}\not\leq T\boldsymbol{\rho}$ is violated for the $j\in[d]$-th coordinate (if there are multiple $j$'s, pick one arbitrarily). We have

$$\sum_{t=1}^{\mathcal{T}_v}c_{t,i_t,j}+1>T\rho_j \Longrightarrow \sum_{t=1}^{\mathcal{T}_v}(\rho_j-c_{t,i_t,j})<1-(T-\mathcal{T}_v)\rho_j. \tag{22}$$

Let $\boldsymbol{\lambda}^*=\frac{1}{\rho_j}\boldsymbol{e}_j$ where $\boldsymbol{e}_j$ is the one-hot vector over coordinate $j$, we know $\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}$ and that

$$\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}^*=\sum_{t=1}^{\mathcal{T}_v}\frac{\rho_j-c_{t,i_t,j}}{\rho_j}\overset{\text{Eq. (22)}}{<}\frac{1-(T-\mathcal{T}_v)\rho_j}{\rho_j}=\rho_j^{-1}-(T-\mathcal{T}_v).$$

Rearranging gives

$$(T-\mathcal{T}_v)\leq\max_{j\in[d]}\rho_j^{-1}+\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\left(\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}^*\right).$$

**Final Bound.** Taking expectation and plugging it back to Eq. (21), we yield

$$\mathbb{E}[\textsc{DualVar}]\leq\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}_t\right]+\sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1+\mathbb{E}[T-\mathcal{T}_v]$$
$$\leq\mathbb{E}\left[\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}_t\right]+\sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1+\mathbb{E}\left[\max_{j\in[d]}\rho_j^{-1}+\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\left(\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}^*\right)\right]$$
$$\leq\mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^*)\right]+\sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1+\|\boldsymbol{\rho}^{-1}\|_\infty.$$

This completes the proof. □

### D.2 DUALVAR Guarantee for FTRL in Eq. (5)

**Theorem D.2** (DUALVAR Guarantee with FTRL)**.** *When using Follow-the-Regularized-Leader (FTRL) in Eq. (5) to decide $\{\boldsymbol{\lambda}_\ell\}_{\ell\in[L]}$, the online learning regret is no more than*

$$\mathfrak{R}_L^{\boldsymbol{\lambda}}:=\mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^*)\right]\leq\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L}+d\sum_{\ell=1}^{L}\eta_\ell|\mathcal{E}_\ell|^2.$$

*Specifically, when setting*

$$\Psi(\boldsymbol{\lambda})=\frac{1}{2}\|\boldsymbol{\lambda}\|_2^2,\quad \eta_\ell=\frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|^2\right)^{-1/2},$$

*the* $\mathbb{E}[\textsc{DualVar}]$ *term is bounded by*

$$\mathbb{E}[\textsc{DualVar}] \le \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2 \cdot \sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} + \sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + \|\boldsymbol{\rho}^{-1}\|_1,$$

*where*

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4).$$

*Proof.* We apply the FTRL guarantee stated as Lemma E.1 with their decision region $\mathcal{X}$ as our dual decision region $\boldsymbol{\Lambda} = \bigotimes_{j=1}^{d}[0,\rho_j^{-1}]$, their norm $\|\cdot\|$ as $\ell_2$-norm, their round number $R$ as our epoch number $L$, their round-$\ell$ loss function $f_\ell$ as our observed loss $F_\ell(\boldsymbol{\lambda}) := \sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}$, the FTRL decisions $\{\boldsymbol{\lambda}_\ell\}_{\ell\in[L]}$ suggested by Lemma E.1 recover our dual-decision rule in Eq. (5), *i.e.*,

$$\boldsymbol{\lambda}_\ell = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}} \sum_{\ell'<\ell}\sum_{\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho}-\boldsymbol{c}_{\tau,i_\tau})^\mathsf{T}\boldsymbol{\lambda} + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda}), \quad \forall \ell\in[L].$$

Since $\nabla_{\boldsymbol{\lambda}}F_\ell(\boldsymbol{\lambda}) = \sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})$ and the dual norm of $\ell_2$-norm is still $\ell_2$-norm, Lemma E.1 gives

$$\mathfrak{R}_L^{\boldsymbol{\lambda}} = \mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^*)\right] \le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L} + \sum_{\ell=1}^{L}\eta_\ell\,\mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})\right\|_2^2\right].$$

Recalling that $\boldsymbol{\rho}, \boldsymbol{c}_{t,i} \in [0,1]^d$, we have $\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})\|_2^2 \le d|\mathcal{E}_\ell|^2$ and hence

$$\mathfrak{R}_L^{\boldsymbol{\lambda}} \le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L} + d\sum_{\ell=1}^{L}\eta_\ell|\mathcal{E}_\ell|^2.$$

This gives the first conclusion in this theorem.

We now move on to the second conclusion in this theorem, namely the online learning regret under the given configuration of $\Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2$ and $\eta_\ell = \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|^2\right)^{-1/2}$. First of all,

$$\Psi(\boldsymbol{\lambda}^*) = \frac{1}{2}\|\boldsymbol{\lambda}^*\|_2^2 \le \frac{1}{2}\|\boldsymbol{\rho}^{-1}\|_2^2, \quad \forall\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda} = \bigotimes_{j=1}^{d}[0,\rho_j^{-1}].$$

Plugging in the specific choice that $\eta_\ell = \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|^2\right)^{-1/2}$, we therefore have

$$\begin{aligned}
\mathfrak{R}_L^{\boldsymbol{\lambda}} &\le \frac{\sqrt{2d}}{\|\boldsymbol{\rho}^{-1}\|_2}\cdot\frac{1}{2}\|\boldsymbol{\rho}^{-1}\|_2^2\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} + \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\cdot d\sum_{\ell=1}^{L}\frac{|\mathcal{E}_\ell|^2}{\sqrt{\sum_{\ell'\le\ell}|\mathcal{E}_{\ell'}|^2}} \\
&\overset{(a)}{\le} \sqrt{2d}\cdot\frac{1}{2}\|\boldsymbol{\rho}^{-1}\|_2\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} + \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{2d}}\cdot\frac{2d}{2}\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} \\
&\overset{(b)}{=} \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2},
\end{aligned}$$

where (a) uses the folklore summation lemma that $\sum_{t=1}^{T}\frac{x_t}{\sqrt{\sum_{s=1}^{t}x_s}} \le 2\sqrt{\sum_{t=1}^{T}x_t}$ for all $x_1, x_2, \ldots, x_T \in \mathbb{R}_{\ge 0}$ (Duchi et al., 2011, Lemma 4) and (b) follows from rearranging the terms.

Plugging the online learning regret $\mathfrak{R}_L^{\boldsymbol{\lambda}}$ into Lemma D.1, we therefore get

$$\mathbb{E}[\textsc{DualVar}] \le \sqrt{2d}\|\boldsymbol{\rho}^{-1}\|_2\sqrt{\sum_{\ell=1}^{L}|\mathcal{E}_\ell|^2} + \sum_{\ell=1}^{L}(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + \|\boldsymbol{\rho}^{-1}\|_1,$$

where

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4).$$

This finishes the proof. □

### D.3 DUALVAR Guarantee for O-FTRL-FP in Eq. (6)

**Theorem D.3** (DUALVAR Guarantee with O-FTRL-FP)**.** *When using Optimistic Follow-the-Regularized-Leader with Fixed-Points (O-FTRL-FP) in Eq. (6) to decide $\{\boldsymbol{\lambda}_\ell\}_{\ell\in[L]}$, the online learning regret is no more than* [5]

$$\begin{aligned}
\mathfrak{R}_L^{\boldsymbol{\lambda}} &:= \mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \sum_{t=1}^{\mathcal{T}_v} (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*)\right] \\
&\le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L} + \sum_{\ell=1}^{L} \eta_\ell \frac{|\mathcal{E}_\ell|^2}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\left(4d^2 + 24d\log(dTL) + 24d\sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{\rho_j}\right) \\
&\quad + \sum_{\ell=1}^{L}\eta_\ell|\mathcal{E}_\ell|(16d+10) + \sum_{\ell=1}^{L}\eta_\ell\left(2N_\ell^2 + \frac{2d|\mathcal{E}_\ell|^2M_\ell^2}{(\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|)^2}\right) + 2\|\boldsymbol{\rho}^{-1}\|_1,
\end{aligned}$$

*where for any $\ell \in [L]$, $N_\ell$ and $M_\ell$ are defined as follows:*

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4),$$

$$M_\ell = \ell\log_{\gamma^{-1}}\left(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3\cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log(dT) + \sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{\rho_j\epsilon}\right).$$

*Specifically, when setting*

$$\Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2,\ L = \log T,\ \mathcal{E}_\ell = [2^{\ell-1}, \min(2^\ell - 1, T)],\ \eta_\ell = \frac{\|\boldsymbol{\rho}^{-1}\|_2}{\sqrt{112d}K^2}\left(\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|\right)^{-1/2},$$

*the* $\mathbb{E}$[DUALVAR] *term is bounded by*

$$\begin{aligned}
\mathbb{E}[\text{DUALVAR}] &\le 2\sqrt{T}\left(\frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2} + 8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^{d}\log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right) \\
&\quad + \sum_{\ell=1}^{L}\left(2N_\ell^2 + (N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + 8dM_\ell^2\right) + 3\|\boldsymbol{\rho}^{-1}\|_1.
\end{aligned}$$

*Proof.* We would like to apply the O-FTRL-FP guarantee stated as Lemma D.4. One main challenge we face is the discontinuity of our predictions; recall the O-FTRL-FP dual update rule from Eq. (6):

$$\begin{aligned}
\boldsymbol{\lambda}_\ell &= \operatorname*{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}} \sum_{\ell'<\ell}\sum_{\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i_\tau})^\mathsf{T}\boldsymbol{\lambda} + \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda} + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda}), \\
\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell) &= |\mathcal{E}_\ell|\cdot\frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)}\right), \\
\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell) &= \operatorname*{argmax}_{i\in[K]}\left(u_{\tau,i} - \boldsymbol{\lambda}_\ell^\mathsf{T}\boldsymbol{c}_{\tau,i}\right), \quad \forall\ell'<\ell, \tau\in\mathcal{E}_{\ell'}.
\end{aligned}$$

**Step 1: Make sure Lemma D.4 is applicable.** Since predicted loss term $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda}$ – where $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)$ is the estimated gradient from historical reports and costs – is not continuous *w.r.t.* $\boldsymbol{\lambda}_\ell$ due to the argmax in $\widetilde{i}_\tau(\boldsymbol{\lambda}_\ell)$, we cannot directly apply the Brouwer's Fixed Point theorem (Munkres, 2000,

[5] For the readers to better interpret this long inequality, we provide an $\widetilde{\mathcal{O}}_T$ version as Eq. (29) in the proof.

Theorem 55.6) to conclude the existence of an exact fixed point $\boldsymbol{\lambda}_\ell$. Fortunately, in Lemma D.4, we prove the existence of an $(\eta_\ell L_\ell \epsilon_\ell)$-approximate fixed point given $(\epsilon_\ell, L_\ell)$-approximate continuity, which requires the existence of two constants $(\epsilon_\ell, L_\ell)$ such that

$$\|\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}^1) - \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}^2)\|_2 \le L_\ell, \quad \forall \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2 \in \boldsymbol{\Lambda} \text{ s.t. } \|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\| \le \epsilon_\ell.$$

We refer the readers to Lemma D.4 for a more general version of the $(\epsilon_\ell, L_\ell)$-approximate continuity condition that we propose, which generalizes to non-linear predicted losses. In Lemma D.5, we prove

$$\begin{aligned} \|\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_1) - \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_2)\|_2 &\le 4|\mathcal{E}_\ell|K^2\epsilon_\ell\epsilon_c d + \frac{4|\mathcal{E}_\ell|(\log\frac{1}{\delta_\ell} + \sum_{j=1}^d \log\frac{\sqrt{d}}{\rho_j\epsilon_\ell})}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sqrt{d}, \\ \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda} \text{ s.t. } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \le \epsilon_\ell, &\quad \textit{w.p. } 1-\delta_\ell, \end{aligned} \tag{23}$$

for any fixed constant $\epsilon_\ell > 0$ and $\delta_\ell \in (0,1)$. Roughly speaking, Lemma D.5 first picks an $\frac{\epsilon_\ell}{2}$-net of $\boldsymbol{\Lambda}$ and then utilizes the smooth cost condition (Assumption 3) to conclude that close $\boldsymbol{\lambda}$'s give similar $\widetilde{i}_\tau(\boldsymbol{\lambda})$'s for all $\tau$, which consequently give similar $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda})$'s within every small ball in the net.

Properly tuning $\epsilon_\ell$ to minimize the RHS in Eq. (23), it translates to the $(\epsilon_\ell, L_\ell)$-approximate continuity of prediction $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda}$ *w.r.t.* $\boldsymbol{\lambda}_\ell$ where

$$\epsilon_\ell = \frac{\sqrt{d}}{K^2\epsilon_c\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}, \quad L_\ell = \frac{4|\mathcal{E}_\ell|\sqrt{d}}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\left(d + \log\frac{1}{\delta_\ell} + \sum_{j=1}^d \log\frac{\sqrt{d}}{\rho_j\epsilon_\ell}\right), \quad \forall \ell \in [L].$$

Under this specific configuration of $\epsilon_\ell$ and $L_\ell$, we call the event defined in in Eq. (23) $\mathcal{G}_\ell$. Conditioning on $\mathcal{G}_1, \dots, \mathcal{G}_L$ which happens with probability at least $1 - \sum_{\ell=1}^L \delta_\ell$ (the conditioning is valid because $\mathcal{G}_\ell$ only depends on a fixed $\frac{\epsilon_\ell}{2}$-net of $\boldsymbol{\Lambda}$ and historical reports and costs, and is thus measurable before the start of epoch $\mathcal{E}_\ell$), we can apply the O-FTRL-FP guarantee stated as Lemma D.4.

Specifically, set their decision region $\mathcal{X}$ as our dual decision region $\boldsymbol{\Lambda} = \bigotimes_{j=1}^d [0, \rho_j^{-1}]$, their norm $\|\cdot\|$ as $\ell_2$-norm, their round number $R$ as our epoch number $L$, their round-$\ell$ loss function $f_\ell$ as our observed loss $F_\ell(\boldsymbol{\lambda}) := \sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}\boldsymbol{\lambda}$, their prediction $\widetilde{f}_\ell(\boldsymbol{\lambda}_\ell, \boldsymbol{\lambda})$ as our $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda}$. The O-FTRL-FP decisions $\{\boldsymbol{\lambda}_\ell\}_{\ell\in[L]}$ suggested by Lemma D.4 recover our dual-decision rule in Eq. (5):

$$\boldsymbol{\lambda}_\ell \approx \operatorname*{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\lambda}} \sum_{\ell'<\ell}\sum_{\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i_\tau})^\mathsf{T}\boldsymbol{\lambda} + \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda} + \frac{1}{\eta_\ell}\Psi(\boldsymbol{\lambda}), \quad \forall \ell \in [L],$$

where the $\approx$ means the $(\eta_\ell L_\ell)$-approximate fixed point suggested by Lemma D.4.

Further noticing that $\nabla_{\boldsymbol{\lambda}} F_\ell(\boldsymbol{\lambda}) = \sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})$, $\nabla_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)^\mathsf{T}\boldsymbol{\lambda}) = \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)$, the loss $F_\ell(\boldsymbol{\lambda})$ is $\sqrt{d}|\mathcal{E}_\ell|$-Lipschitz since $(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t}) \in [0,1]^d$, the loss and predictions are 0-smooth *w.r.t.* $\boldsymbol{\lambda}$ since they are linear, and the dual norm of $\ell_2$-norm is still $\ell_2$-norm, Lemma D.4 gives

$$\begin{aligned} \mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \sum_{t=1}^{\mathcal{T}_v} (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*)\right] \le \underbrace{\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L}}_{\texttt{Diameter}} + \sum_{\ell=1}^L \eta_\ell \underbrace{\mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t}) - \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_\ell)\right\|_2^2\right]}_{\texttt{Stability}_\ell} \\ + \underbrace{\sum_{\ell=1}^L \sqrt{d}|\mathcal{E}_\ell|\eta_\ell L_\ell + \left(\sum_{\ell=1}^L \delta_\ell\right) 2T\|\boldsymbol{\rho}^{-1}\|_1}_{\texttt{Fixed Point Error}}, \end{aligned} \tag{24}$$

where the last term considers the failure probability of $\mathcal{G}_1, \dots, \mathcal{G}_L$, in which case we use the trivial bound that $\sum_{t=1}^T (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*) \le T \cdot (\max_{t,i}\|\boldsymbol{\rho} - \boldsymbol{c}_{t,i}\|_\infty)(2\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}}\|\boldsymbol{\lambda}\|_1) \le 2T\|\boldsymbol{\rho}^{-1}\|_1$, where the last inequality is due to $\boldsymbol{\rho} - \boldsymbol{c}_{t,i} \in [-1,1]^2$ and $\boldsymbol{\lambda} \in \bigotimes_{j=1}^d [0, \rho_j^-]$.

**Step 2: Control the `Stability` terms.** For analytical convenience, we add a superscript $u$ to the notation $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda})$ to highlight it is yielded from reports $\{\boldsymbol{u}_\tau\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$. Analogous to $\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda})$, which

is computed from agents' strategic reports $\{\boldsymbol{u}_\tau\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$, we define $\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})$ using the true values $\{\boldsymbol{v}_\tau\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$, and $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})$ using the underlying true distributions $\mathcal{V}=\{\mathcal{V}_i\}_{i\in[K]}$ and $\mathcal{C}=\{\mathcal{C}_i\}_{i\in[K]}$:

$$\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda}) = |\mathcal{E}_\ell|\cdot\frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\left(\boldsymbol{\rho}-\boldsymbol{c}_{\tau,\widetilde{i}^u_\tau(\boldsymbol{\lambda})}\right),\quad \widetilde{i}^u_\tau(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]}\left(u_{\tau,i}-\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right);$$

$$\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}) = |\mathcal{E}_\ell|\cdot\frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\left(\boldsymbol{\rho}-\boldsymbol{c}_{\tau,\widetilde{i}^v_\tau(\boldsymbol{\lambda})}\right),\quad \widetilde{i}^v_\tau(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]}\left(v_{\tau,i}-\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right);$$

$$\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}) = |\mathcal{E}_\ell|\cdot\mathop{\mathbb{E}}_{\boldsymbol{v}_*\sim\mathcal{V},\boldsymbol{c}_*\sim\mathcal{C}}\left[\boldsymbol{\rho}-\boldsymbol{c}_{*,\widetilde{i}^*(\boldsymbol{\lambda})}\right],\qquad \widetilde{i}^*(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]}\left(v_{*,i}-\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{*,i}\right). \quad (25)$$

We now decompose the $\texttt{Stability}_\ell$ term above as

$$\texttt{Stability}_\ell = \mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})-\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda}_\ell)\right\|_2^2\right]$$

$$\le \underbrace{2\,\mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})-\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell)\right\|_2^2\right]}_{\text{Difference between }\{i_t\}_{t\in\mathcal{E}_\ell}\text{ and }\{\widetilde{i}^*_t\}_{t\in\mathcal{E}_\ell}} + \underbrace{2\,\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell)-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\ell)\|_2^2\right]}_{\text{Empirical Estimation}} + \underbrace{2\,\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\ell)-\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda}_\ell)\|_2^2\right]}_{\text{Untruthful Reports}}.$$

In Lemmas D.6 to D.8 presented immediately after this theorem, we control these three terms one by one. Specifically, Lemma D.6 relates $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell)$ first to $\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,\widetilde{i}^*_t})$ and then to $\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})$, which gives

$$\mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})-\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell)\right\|_2^2\right] \le (d+3)|\mathcal{E}_\ell|+N_\ell^2, \quad (26)$$

where $N_\ell = 1+4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1+5\log|\mathcal{E}_\ell|+\log_{\gamma^{-1}}(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4)$.

Meanwhile, in Lemmas D.7 and D.8, by an $\epsilon$-net argument, we first prove $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})\approx\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})$ (*resp.* $\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})\approx\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda})$) for all $\boldsymbol{\lambda}$'s in the $\epsilon$-net and then extend this similarity to all $\boldsymbol{\lambda}\in\boldsymbol{\Lambda}$ using the smooth cost condition; we refer the readers to corresponding proofs for more details. Summing up their conclusions and taking Union Bound, they together ensure that for any $\epsilon>0$ and $\delta\in(0,1)$, with probability $1-6\delta$, we have

$$\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}}\left(\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})\|_2^2+\|\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda})\|_2^2\right)$$

$$\le 7d|\mathcal{E}_\ell|^2\cdot(K^2\epsilon\epsilon_c)^2+10d|\mathcal{E}_\ell|^2\cdot\frac{\log\frac{1}{\delta}+\sum_{j=1}^d\log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}+\frac{d|\mathcal{E}_\ell|^2}{(\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|)^2}M_\ell^2, \quad (27)$$

where

$$M_\ell := \ell\log_{\gamma^{-1}}\left(1+4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3\cdot 4\ell\delta^{-1}\right)+4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1+4\left(\log\frac{1}{\delta}+\sum_{j=1}^d\log\frac{d}{\rho_j\epsilon}\right).$$

**Step 3: Plug $\texttt{Stability}$ back to O-FTRL-FP guarantee.** Now we plug the $\texttt{Stability}$ bounds from Eqs. (26) and (27) into the online learning regret bound derived in Eq. (24). For every $\ell\in[L]$, Eq. (27) happen with probability $1-6\delta$; in case it does not hold, we use the trivial bound that $\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})\|_2^2+\|\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda})\|_2^2\le 2d|\mathcal{E}_\ell|^2$. Therefore, Eq. (24) translates to:

$$\mathfrak{R}_L^{\boldsymbol{\lambda}} = \mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\mathsf{T}(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^*)\right]$$

$$\le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L}+2\sum_{\ell=1}^L\eta_\ell\left((d+3)|\mathcal{E}_\ell|+N_\ell^2+7d|\mathcal{E}_\ell|^2\cdot(K^2\epsilon\epsilon_c)^2\right.$$

$$+ 10d|\mathcal{E}_\ell|^2 \cdot \frac{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} + \frac{d|\mathcal{E}_\ell|^2}{(\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|)^2}M_\ell^2 + 6\delta\cdot 2d|\mathcal{E}_\ell|^2\Bigg)$$
$$+ \sum_{\ell=1}^L \sqrt{d}|\mathcal{E}_\ell|\eta_\ell \frac{4|\mathcal{E}_\ell|\sqrt{d}}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\left(d + \log\frac{1}{\delta_\ell} + \sum_{j=1}^d \log\frac{\sqrt{d}}{\rho_j\epsilon_\ell}\right) + \left(\sum_{\ell=1}^L \delta_\ell\right) 2T\|\boldsymbol{\rho}^{-1}\|_1,$$

where $\epsilon_\ell = \frac{\sqrt{d}}{K^2\epsilon_c\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}$ and $\{\delta_\ell\}_{\ell\in[L]}$, $\epsilon$, and $\delta$ are some parameters that we can tune. We also recall the definitions of $N_\ell$ and $M_\ell$:

$$N_\ell = 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4),$$
$$M_\ell = \ell\log_{\gamma^{-1}}\left(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3\cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right).$$

**Step 4: Deriving first half of this theorem.** We now configure $\epsilon$, $\delta$, and $\delta_\ell$'s as follows:

$$\epsilon = \frac{1}{\sqrt{T}K^2\epsilon_c}, \quad \delta = \frac{1}{6dT}, \quad \delta_\ell = \frac{1}{TL},$$

We remark that we did not make every effort to make the overall online learning regret as small as possible. Instead, the above tuning mainly focuses on polynomial dependencies on $T$ and $L$.

Under this specific tuning and simplifying using the fact that $\sum_{\ell=1}^L|\mathcal{E}_\ell| \le T$, we get

$$\begin{aligned}
\mathfrak{R}_L^{\boldsymbol{\lambda}} &\le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L} \\
&\quad + 2\sum_{\ell=1}^L \eta_\ell\left((8d+5)|\mathcal{E}_\ell| + N_\ell^2 + 10d|\mathcal{E}_\ell|^2\frac{\log(dT) + \sum_{j=1}^d\log\frac{dK^2\epsilon_cT}{\rho_j}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} + \frac{d|\mathcal{E}_\ell|^2M_\ell^2}{(\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|)^2}\right) \\
&\quad + \sum_{\ell=1}^L \eta_\ell\frac{4d|\mathcal{E}_\ell|^2}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\left(d + \log(TL) + \sum_{j=1}^d\log\frac{K^2\epsilon_c\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}{\rho_j}\right) + 2\|\boldsymbol{\rho}^{-1}\|_1 \\
&\le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L} + \sum_{\ell=1}^L \eta_\ell\frac{|\mathcal{E}_\ell|^2}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\left(4d^2 + 24d\log(dTL) + 24d\sum_{j=1}^d\log\frac{dK^2\epsilon_cT}{\rho_j}\right) \\
&\quad + \sum_{\ell=1}^L \eta_\ell|\mathcal{E}_\ell|(16d+10) + \sum_{\ell=1}^L\eta_\ell\left(2N_\ell^2 + \frac{2d|\mathcal{E}_\ell|^2M_\ell^2}{(\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|)^2}\right) + 2\|\boldsymbol{\rho}^{-1}\|_1. \qquad (28)
\end{aligned}$$

For the readers to better interpret, we annotate the order of every term in terms of $\widetilde{\mathcal{O}}_T$, which only highlights the polynomial dependency on $T$, and consequently, also $L$, $\{|\mathcal{E}_\ell|\}_{\ell\in[L]}$, and $\{\eta_\ell\}_{\ell\in[L]}$:

$$\mathfrak{R}_L^{\boldsymbol{\lambda}} = \widetilde{\mathcal{O}}_T\left(\eta_L^{-1} + \sum_{\ell=1}^L\eta_\ell\frac{|\mathcal{E}_\ell|^2}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} + \sum_{\ell=1}^L\eta_\ell|\mathcal{E}_\ell| + \sum_{\ell=1}^L\eta_\ell + 1\right). \qquad (29)$$

**Step 5: Plug in specific tuning of $\Psi, L, \{\mathcal{E}_\ell\}_{\ell\in[L]}$, and $\{\eta_\ell\}_{\ell\in[L]}$.** We now plug in the following specific configuration:

$$\Psi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2,\ L = \log T,\ \mathcal{E}_\ell = [2^{\ell-1}, \min(2^\ell-1, T)],\ \eta_\ell = \left(\sum_{\ell'=1}^\ell|\mathcal{E}_{\ell'}|\right)^{-1/2}.$$

Due to the doubling epoch length structure, we observe that

$$\frac{|\mathcal{E}_\ell|}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \le \frac{(2^{\ell-1})}{2^{\ell-1}-1} \le 2, \quad \forall\ell > 1.$$

Therefore the informal bound in Eq. (29) becomes $\widetilde{\mathcal{O}}_T(\eta_L^{-1}+\sum_{\ell=1}^L \eta_\ell|\mathcal{E}_\ell|)$. This explains our choice that $\eta_\ell = (\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|)^{-1/2}$. To make it formal, substituting $\frac{|\mathcal{E}_\ell|}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \le 2$ into Eq. (28):

$$\begin{aligned}\mathfrak{R}_L^{\boldsymbol{\lambda}} &\le \sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}} \frac{\Psi(\boldsymbol{\lambda}^*)}{\eta_L} + \sum_{\ell=1}^L \eta_\ell|\mathcal{E}_\ell| \left(8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^d \log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right)\\ &\quad + \sum_{\ell=1}^L \eta_\ell\left(2N_\ell^2 + 8dM_\ell^2\right) + 2\|\boldsymbol{\rho}^{-1}\|_1.\end{aligned}$$

Under the specific choice of $\Psi(\boldsymbol{\lambda}^*) = \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2$ and that $\eta_\ell = (\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|)^{-1/2}$, we get

$$\begin{aligned}\mathfrak{R}_L^{\boldsymbol{\lambda}} &\le \frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2}\left(\sum_{\ell=1}^L|\mathcal{E}_\ell|\right)^{1/2}\\ &\quad + \sum_{\ell=1}^L \frac{|\mathcal{E}_\ell|}{\sqrt{\sum_{\ell'=1}^{\ell}|\mathcal{E}_{\ell'}|}}\left(8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^d \log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right)\\ &\quad + \sum_{\ell=1}^L \eta_\ell\left(2N_\ell^2 + 8dM_\ell^2\right) + 2\|\boldsymbol{\rho}^{-1}\|_1\\ &\le 2\sqrt{\sum_{\ell=1}^L|\mathcal{E}_\ell|}\left(\frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2} + 8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^d \log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right)\\ &\quad + \sum_{\ell=1}^L\left(2N_\ell^2 + 8dM_\ell^2\right) + 2\|\boldsymbol{\rho}^{-1}\|_1.\end{aligned}$$

where the second inequality again uses the folklore summation lemma that $\sum_{t=1}^T \frac{x_t}{\sqrt{\sum_{s=1}^t x_s}} \le 2\sqrt{\sum_{t=1}^T x_t}$ for all $x_1, x_2, \ldots, x_T \in \mathbb{R}_{\ge 0}$ (Duchi et al., 2011, Lemma 4); we also used the trivial bound that $\eta_\ell \le 1$ for the non-dominant terms. Once again, we remark that the final bound is only optimized *w.r.t.* $\mathrm{poly}(T)$ dependencies.

To translate the online learning regret $\mathfrak{R}_L^{\boldsymbol{\lambda}}$ to the $\mathbb{E}[\textsc{DualVar}]$ guarantee, we use Lemma D.1 (we also usd the fact that $\sum_{\ell=1}^L|\mathcal{E}_\ell| = T$):

$$\begin{aligned}\mathbb{E}[\textsc{DualVar}] &\le \mathbb{E}\left[\sup_{\boldsymbol{\lambda}^*\in\boldsymbol{\Lambda}}\sum_{t=1}^{\mathcal{T}_v}(\boldsymbol{\rho}-\boldsymbol{c}_{t,i_t})^\intercal(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}^*)\right] + \sum_{\ell=1}^L(N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + \|\boldsymbol{\rho}^{-1}\|_1\\ &\le 2\sqrt{T}\left(\frac{\|\boldsymbol{\rho}^{-1}\|_2^2}{2} + 8d^2 + 48d\log(dTL) + 48d\sum_{j=1}^d \log\frac{dK^2\epsilon_c T}{T\rho_j} + 16d + 10\right)\\ &\quad + \sum_{\ell=1}^L\left(2N_\ell^2 + (N_\ell+3)\|\boldsymbol{\rho}^{-1}\|_1 + 8dM_\ell^2\right) + 3\|\boldsymbol{\rho}^{-1}\|_1.\end{aligned}$$

This finishes the proof. □

### D.3.1 O-FTRL-FP Framework for Non-Continuous Predictions

**Lemma D.4** (O-FTRL-FP Guarantee)**.** *For a convex and compact region $\mathcal{X}$ within an Euclidean space $\mathbb{R}^d$, an 1-strongly-convex regularizer $\Psi\colon \mathcal{X}\to\mathbb{R}$ w.r.t. some norm $\|\cdot\|$ with $\min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x}) = 0$, a sequence of continuous, differentiable, convex, L-Lipschitz, and $L_f$-smooth losses $f_1, f_2, \ldots, f_R$, a sequence of learning rates $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_R \ge 0$, a (stochastic) **action-dependent** prediction sequence $\{\widetilde{f}_r\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}\}_{r\in[R]}$ such that the following conditions hold:*

1. *$\widetilde{f}_r$ is $\mathcal{F}_{r-1}$-measurable where $(\mathcal{F}_r)_t$ is the natural filtration that $\mathcal{F}_r = \sigma(f_1, f_2, \ldots, f_r)$,*
2. *For any fixed $\boldsymbol{y} \in \mathcal{X}$, $\widetilde{f}_r(\boldsymbol{y}, \cdot)$ is continuous, differentiable, convex, and $L_f$-smooth, and*
3. *$\widetilde{f}_r(\boldsymbol{y}, \boldsymbol{x})$ is $(\epsilon_r, L_r)$-approximately-continuous w.r.t. its first parameter $\boldsymbol{y}$ in the sense that*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left\| \nabla_2 \widetilde{f}_r(\boldsymbol{y}_1, \boldsymbol{x}) - \nabla_2 \widetilde{f}_r(\boldsymbol{y}_2, \boldsymbol{x}) \right\|_* \le L_r, \quad \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{X} \text{ s.t. } \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \epsilon_r,$$

*where $\nabla_2 \widetilde{f}_r$ is the gradient of $\widetilde{f}_r(\boldsymbol{y}, \boldsymbol{x})$ taken only w.r.t. the second parameter $\boldsymbol{x}$, and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.*

*Then, for all $r = 1, 2, \ldots, R$, consider the following fixed-point system:*

$$\boldsymbol{x}_r = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( \sum_{\mathfrak{r}=1}^{r-1} f_{\mathfrak{r}}(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{x}_r, \boldsymbol{x}) + \frac{1}{\eta_r}\Psi(\boldsymbol{x}) \right). \tag{30}$$

*First of all, the argmin in the RHS of Eq. (30) exists and is unique. Furthermore, Eq. (30) allows an $(\eta_r L_r)$-approximate fixed point $\boldsymbol{x}_r$ such that*

$$\left\| \boldsymbol{x}_r - \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( \sum_{\mathfrak{r}=1}^{r-1} f_{\mathfrak{r}}(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{x}_r, \boldsymbol{x}) + \frac{1}{\eta_r}\Psi(\boldsymbol{x}) \right) \right\| \le \eta_r L_r, \quad \forall r \in [R].$$

*Using this $\{\boldsymbol{x}_r\}_{r\in[R]}$, we have the following guarantee for all $\boldsymbol{x}^* \in \mathcal{X}$:*

$$\sum_{r=1}^{R} (f_r(\boldsymbol{x}_r) - f_r(\boldsymbol{x}^*)) \le \frac{\Psi(\boldsymbol{x}^*)}{\eta_R} + \frac{1}{2}\sum_{r=1}^{R} \eta_r \|\nabla f_r(\boldsymbol{x}_r) - \nabla \widehat{f}_r(\boldsymbol{x}_r)\|_*^2 + \sum_{r=1}^{R} (L\eta_r L_r + 2L_f^2 \eta_r^2 L_r^2).$$

*Proof.* We first study the system Eq. (30) for any fixed $r \in [R]$. For simplicity, we drop the subscripts $r$ from $F(\boldsymbol{x})$ and $\boldsymbol{G}(\boldsymbol{y})$ defined soon. Since $\Psi(x)$ is 1-strongly-convex and $f_{\mathfrak{r}}(\boldsymbol{x})$ is convex, $\forall \mathfrak{r} < r$,

$$F(\boldsymbol{x}) := \sum_{\mathfrak{r}=1}^{r-1} f_{\mathfrak{r}}(\boldsymbol{x}) + \frac{1}{\eta_r}\Psi(\boldsymbol{x}) \text{ is } \eta_r^{-1}\text{-strongly-convex}, \quad \forall r \in [R].$$

Fix any $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{X}$ such that $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \epsilon_r$. From Condition 2 of $\widetilde{f}_r$, $H_1(\boldsymbol{x}) := F(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{y}_1, \boldsymbol{x})$ and $H_2(\boldsymbol{x}) := F(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{y}_2, \boldsymbol{x})$ are continuous, differentiable, and $\eta_r^{-1}$-strongly-convex. Hence $\boldsymbol{x}_1 := \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} H_1(\boldsymbol{x})$ and $\boldsymbol{x}_2 := \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} H_2(\boldsymbol{x})$ exist and are unique. Further utilizing the $\eta_r^{-1}$-strong-convexity of $H_1$, we have (Bertsekas et al., 2003, Exercise 1.10)

$$\begin{aligned}
\eta_r^{-1}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 &\le \langle \nabla H_1(\boldsymbol{x}_1) - \nabla H_1(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \\
&= \langle \nabla H_1(\boldsymbol{x}_1), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle + \langle \nabla H_1(\boldsymbol{x}_2), \boldsymbol{x}_2 - \boldsymbol{x}_1 \rangle \\
&\overset{(a)}{\le} \langle \nabla H_1(\boldsymbol{x}_2), \boldsymbol{x}_2 - \boldsymbol{x}_1 \rangle \\
&\overset{(b)}{\le} \langle \nabla H_1(\boldsymbol{x}_2), \boldsymbol{x}_2 - \boldsymbol{x}_1 \rangle + \langle \nabla H_2(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \\
&\overset{(c)}{\le} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \cdot \|\nabla H_1(\boldsymbol{x}_2) - \nabla H_2(\boldsymbol{x}_2)\|_*,
\end{aligned}$$

where (a) uses the first-order condition of $H_1(\boldsymbol{x}_1)$ that $\langle \nabla H_1(\boldsymbol{x}_1), \boldsymbol{x}_2 - \boldsymbol{x}_1 \rangle \ge 0$, (b) uses the first-order condition of $H_2(\boldsymbol{x}_2)$ that $\langle \nabla H_2(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \ge 0$, and (c) applies Cauchy-Schwartz inequality. Using Condition 3 of $\widetilde{f}_r$ and rearranging terms, we further have

$$\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \le \eta_r \cdot \|\nabla H_1(\boldsymbol{x}_2) - \nabla H_2(\boldsymbol{x}_2)\|_* = \eta_r \cdot \|\nabla_2 \widetilde{f}_r(\boldsymbol{y}_1, \boldsymbol{x}_2) - \nabla_2 \widetilde{f}_r(\boldsymbol{y}_2, \boldsymbol{x}_2)\|_* \le \eta_r L_r.$$

For any $\boldsymbol{y} \in \mathcal{X}$, due to strong convexity of $F(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{y}, \boldsymbol{x})$, the following $\boldsymbol{G}(\boldsymbol{y})$ is well-defined:

$$\boldsymbol{G}(\boldsymbol{y}) := \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( \sum_{\mathfrak{r}=1}^{r-1} f_{\mathfrak{r}}(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{y}, \boldsymbol{x}) + \frac{1}{\eta_r}\Psi(\boldsymbol{x}) \right) = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( F(\boldsymbol{x}) + \widetilde{f}_r(\boldsymbol{y}, \boldsymbol{x}) \right).$$

Eq. (30) translates to $\boldsymbol{x}_r = \boldsymbol{G}(\boldsymbol{x}_r)$, and the aforementioned $\boldsymbol{x}_1 = \boldsymbol{G}(\boldsymbol{y}_1)$, $\boldsymbol{x}_2 = \boldsymbol{G}(\boldsymbol{y}_2)$. Thus

$$\|\boldsymbol{G}(\boldsymbol{y}_1) - \boldsymbol{G}(\boldsymbol{y}_2)\| = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \le \eta_r L_r, \quad \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{X} \text{ s.t. } \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| < \epsilon_r. \tag{31}$$

We now utilize the partitions of unity tool in topology. Consider an $\frac{\epsilon_r}{2}$-net of $\mathcal{X}$, whose size is finite since $\mathcal{X}$ is a compact subset of the Euclidean space. Denote it by $\mathcal{X} \subseteq \bigcup_{i=1}^M \mathcal{B}_i$ where $\mathcal{B}_i$ is a ball with radius $\frac{\epsilon_r}{2}$ centered at some $\widetilde{\boldsymbol{y}}_i \in \mathcal{X}$. It induces a continuous partition of unity $\{\phi_i \colon \mathcal{B}_i \to [0,1]\}_{i\in[M]}$ such that $\sum_{i=1}^M \phi_i(\boldsymbol{y}) = 1$ for all $\boldsymbol{y} \in \mathcal{X}$ (Munkres, 2000, Theorem 36.1). Consider

$$\widetilde{\boldsymbol{G}}(\boldsymbol{y}) := \sum_{i=1}^M \phi_i(\boldsymbol{y}) \boldsymbol{G}(\widetilde{\boldsymbol{y}}_i), \quad \forall \boldsymbol{y} \in \mathcal{X},$$

which is continuous since every $\phi_i$ is. Furthermore, as $\boldsymbol{G}(\widetilde{\boldsymbol{y}}_i) \in \mathcal{X}$ for all $i \in [M]$, we know $\widetilde{\boldsymbol{G}}$ is a continuous map from $\mathcal{X}$ to $\mathcal{X}$. As $\mathcal{X}$ is a non-empty, convex, and compact set, the Brouwer's Fixed Point theorem (Munkres, 2000, Theorem 55.6) suggests the existence of $\boldsymbol{y}^* \in \mathcal{X}$ such that $\widetilde{\boldsymbol{G}}(\boldsymbol{y}^*) = \boldsymbol{y}^*$. This $\boldsymbol{y}^* \in \mathcal{X}$ then ensures

$$\begin{aligned}
\|\boldsymbol{y}^* - \boldsymbol{G}(\boldsymbol{y}^*)\| = \|\widetilde{\boldsymbol{G}}(\boldsymbol{y}^*) - \boldsymbol{G}(\boldsymbol{y}^*)\| = \left\| \sum_{i=1}^M \phi_i(\boldsymbol{y}) (\boldsymbol{G}(\widetilde{\boldsymbol{y}}_i) - \boldsymbol{G}(\boldsymbol{y}^*)) \right\| \\
\le \sum_{i=1}^M \phi_i(\boldsymbol{y}^*) \|\boldsymbol{G}(\widetilde{\boldsymbol{y}}_i) - \boldsymbol{G}(\boldsymbol{y}^*)\| \overset{(a)}{\le} \sum_{i=1}^M \phi_i(\boldsymbol{y}^*) \cdot \eta_r L_r \le \eta_r L_r,
\end{aligned}$$

where (a) uses the fact that if $\phi_i(\boldsymbol{y}^*) \neq 0$, then $\boldsymbol{y}^* \in \mathcal{B}_i$ and hence $\|\boldsymbol{y}^* - \widetilde{\boldsymbol{y}}_i\| < \epsilon_r$, which gives $\|\boldsymbol{G}(\boldsymbol{y}^*) - \boldsymbol{G}(\widetilde{\boldsymbol{y}}_i)\| \le \eta_r L_r$ from Eq. (31). Therefore, for any round $r \in [R]$, the system Eq. (30) indeed allows an $(\eta_r L_r)$-approximate fixed point $\boldsymbol{x}_r \in \mathcal{X}$.

After proving the existence of approximate fixed points in Eq. (30), we utilize the vanilla O-FTRL result stated as Lemma E.2. Since every term in Eq. (30) is $\mathcal{F}_{r-1}$-measurable, the approximate fixed point $\boldsymbol{x}_r$ and its induced prediction $\widehat{\ell}_r(\cdot) := \widetilde{\ell}_r(\boldsymbol{x}_r, \cdot)$ are $\mathcal{F}_{r-1}$-measurable. Therefore, the $\{\widehat{\ell}_r\}_{r\in[R]}$ serves as a valid prediction required by Lemma E.2. Applying Lemma E.2, for the sequence

$$\boldsymbol{x}_r^* = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( \sum_{\mathfrak{r}=1}^{r-1} f_{\mathfrak{r}}(\boldsymbol{x}) + \widehat{f}_r(\boldsymbol{x}) + \frac{1}{\eta_r} \Psi(\boldsymbol{x}) \right) = \boldsymbol{G}(\boldsymbol{x}_r), \quad \forall r \in [R],$$

we have

$$\sum_{r=1}^R (f_r(\boldsymbol{x}_r^*) - f_r(\boldsymbol{x}^*)) \le \frac{\Psi(\boldsymbol{x}^*)}{\eta_R} + \frac{1}{2} \sum_{r=1}^R \eta_r \|\nabla f_r(\boldsymbol{x}_r^*) - \nabla \widehat{f}_r(\boldsymbol{x}_r^*)\|_*^2, \quad \forall \boldsymbol{x}^* \in \mathcal{X}.$$

Since $\boldsymbol{x}_r$ is an $(\eta_r L_r)$-approximate fixed point of $\boldsymbol{G}$, we know $\|\boldsymbol{x}_r - \boldsymbol{x}_r^*\| = \|\boldsymbol{x}_r - \boldsymbol{G}(\boldsymbol{x}_r)\| \le \eta_r L_r \epsilon_r$. Further realizing that $\nabla \widehat{f}_r(\boldsymbol{x}) = \nabla_2 \widetilde{f}_r(\boldsymbol{x}_r, \boldsymbol{x})$ by definition of $\widehat{f}_r(\cdot) = \widetilde{f}_r(\boldsymbol{x}_r, \cdot)$, we get

$$\begin{aligned}
|f_r(\boldsymbol{x}_r) - f_r(\boldsymbol{x}_r^*)| &\le L \cdot \eta_r L_r, \\
\|\nabla f_r(\boldsymbol{x}_r) - \nabla f_r(\boldsymbol{x}_r^*)\|_* &\le L_f \cdot \eta_r L_r, \\
\|\nabla_2 \widetilde{f}_r(\boldsymbol{x}_r, \boldsymbol{x}_r) - \nabla_2 \widetilde{f}_r(\boldsymbol{x}_r, \boldsymbol{x}_r^*)\|_* &\le L_f \cdot \eta_r L_r, \quad \forall r \in [R],
\end{aligned}$$

where the first inequality uses the $L$-Lipschitzness of $f_r$, and the second and third inequalities use the $L_f$-smoothness of $f_r$ and $\widetilde{f}_r(\boldsymbol{x}_r, \cdot)$. Plugging back gives our conclusion that for any $\boldsymbol{x}^* \in \mathcal{X}$,

$$\sum_{r=1}^R (f_r(\boldsymbol{x}_r) - f_r(\boldsymbol{x}^*)) \le \frac{\Psi(\boldsymbol{x}^*)}{\eta_R} + \frac{1}{2} \sum_{r=1}^R \eta_r \|\nabla f_r(\boldsymbol{x}_r) - \nabla \widehat{f}_r(\boldsymbol{x}_r)\|_*^2 + \sum_{r=1}^R (L \eta_r L_r + 2 L_f^2 \eta_r^2 L_r^2).$$

This finishes the proof. □

### D.3.2 Approximate Continuity of Predictions in O-FTRL-FP

**Lemma D.5** (Approximate Continuity of Predictions)**.** *Recall the definition of $\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda})$ from Eq. (6):*

$$\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}) = |\mathcal{E}_\ell| \cdot \frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sum_{\ell'<\ell, \tau\in\mathcal{E}_{\ell'}} \left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau(\boldsymbol{\lambda})}\right),$$
$$\widetilde{i}_\tau(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]} \left(u_{\tau,i} - \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right), \quad \forall \ell' < \ell, \tau \in \mathcal{E}_{\ell'}.$$

*For any fixed $\epsilon > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_1) - \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_2)\|_2 \le 4|\mathcal{E}_\ell|K^2\epsilon\epsilon_c d + \frac{4|\mathcal{E}_\ell|(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{\sqrt{d}}{\rho_j\epsilon})}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sqrt{d},$$
$$\forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda} \text{ s.t. } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \le \epsilon.$$

*Proof.* Take an $\frac{\epsilon}{2}$-net of $\boldsymbol{\Lambda}$ defined in Lemma E.3, and denote it by $\boldsymbol{\Lambda}_\epsilon$ (which we slightly abused the notation; note that $\frac{\epsilon}{2}$-nets are also $\epsilon$-nets since $\frac{\epsilon}{2} < \epsilon$). Then we have $|\boldsymbol{\Lambda}_\epsilon| \le \prod_{j=1}^d (2d/\rho_j\epsilon)$.

Fix a $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$ and consider the stochastic process $(X_\tau)_{\tau\ge1}$ adapted to $(\mathcal{F}_\tau)_{\tau\ge0}$: [6]

$$X_\tau := \mathbb{1}[\exists \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) \text{ s.t. } \widetilde{i}_\tau(\boldsymbol{\lambda}) \ne \widetilde{i}_\tau(\boldsymbol{\lambda}_\epsilon)], \quad \forall \ell' < \ell, \tau \in \mathcal{E}_{\ell'},$$

where $(\mathcal{F}_\tau)_{\tau\ge0}$ is defined as $\mathcal{F}_\tau = \sigma(\bigcup_{i\in\{0\}\cup[K]} \mathcal{H}_{\tau+1,i})$, *i.e.*, the smallest $\sigma$-algebra containing all revealed history up to the end of round $\tau$. Then $X_\tau$ is $\mathcal{F}_\tau$-measurable. Using the Multiplicative Azuma-Hoeffding inequality given as Lemma E.4 with $Y_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ and $\epsilon = \frac{1}{2}$,

$$\Pr\left\{\frac{1}{2}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} X_\tau \ge \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{E}[X_\tau \mid \mathcal{F}_{\tau-1}] + A\right\} \le \exp\left(-\frac{A}{2}\right), \quad \forall A > 0.$$

For any $\ell' < \ell$ and $\tau \in \mathcal{E}_{\ell'}$, the distribution of $\boldsymbol{u}_\tau$ conditional on all previous history, namely $\bigcup_{i\in\{0\}\cup[K]} \mathcal{H}_{\tau,i}$, is $\mathcal{F}_{\tau-1}$-measurable (*i.e.*, it follows a $\mathcal{F}_{\tau-1}$-measurable joint distribution $\mathcal{U} \in \triangle([0,1]^d)$). $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$ is fixed before the game and thus also $\mathcal{F}_{\tau-1}$-measurable. Lemma E.3 gives

$$\mathbb{E}[X_\tau \mid \mathcal{F}_{\tau-1}] = \Pr\left\{\exists \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) \text{ s.t. } \operatorname*{argmax}_{i\in[K]} \left(u_{\tau,i} - \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right) \ne \operatorname*{argmax}_{i\in[K]} \left(u_{\tau,i} - \boldsymbol{\lambda}_\epsilon^\mathsf{T}\boldsymbol{c}_{\tau,i}\right)\right\}$$
$$\le K^2\epsilon\epsilon_c, \quad \forall \ell' < \ell, \tau \in \mathcal{E}_{\ell'},$$

where the Pr is taken *w.r.t.* the randomness of generating $\boldsymbol{u}_\tau$ according to the conditional joint distribution $\boldsymbol{u}_\tau \mid \bigcup_{i\in\{0\}\cup[K]} \mathcal{H}_{\tau,i}$ and the independent sampling of $\boldsymbol{c}_\tau \sim \mathcal{C}$.

Hence, for any failure probability $\delta > 0$ that we determine later, with probability $1-\delta$,

$$\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}[\exists \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) \text{ s.t. } \widetilde{i}_\tau(\boldsymbol{\lambda}) \ne \widetilde{i}_\tau(\boldsymbol{\lambda}_\epsilon)] \le 2\sum_{\ell'<\ell} |\mathcal{E}_{\ell'}| \cdot K^2\epsilon\epsilon_c + 4\log\frac{1}{\delta}.$$

Taking Union Bound over $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$, with probability $1 - \delta\prod_{j=1}^d (2d/\rho_j\epsilon)$, the above good event holds for all $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}$ at the same time. Consider any $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda}$ such that $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \le \frac{\epsilon}{2\sqrt{d}}$, which immediately gives $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_1 \le \frac{\epsilon}{2}$. Take $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$ such that $\boldsymbol{\lambda}_1 \in \mathcal{B}_{\epsilon/2}(\boldsymbol{\lambda}_\epsilon)$ (recall that $\boldsymbol{\Lambda}_\epsilon$ is in fact a $\frac{\epsilon}{2}$-net), we therefore have $\|\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_\epsilon\|_1 \le \epsilon$, which means $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$. Thus

$$\left\|\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau(\boldsymbol{\lambda}_1)}) - \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau(\boldsymbol{\lambda}_2)})\right\|_2 \le \left(2\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}| \cdot K^2\epsilon\epsilon_c + 4\log\frac{1}{\delta}\right)\sqrt{d},$$

[6] This $\exists \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$ is pivotal because we cannot afford to take a Union Bound over all $\boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$. We call this step "uniform smoothness", since it ensures the similarity holds uniformly in the neighborhood of $\boldsymbol{\lambda}_\epsilon$.

$$\forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda} \text{ s.t. } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \le \frac{\epsilon}{2\sqrt{d}}, \quad \textit{w.p. } 1 - \delta \prod_{j=1}^{d} (2d/\rho_j \epsilon),$$

where we used the fact that $\boldsymbol{c}_{\tau,i} \in [0,1]^d$ for any $i \in [K]$. This ensures that

$$\|\widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_1) - \widetilde{\boldsymbol{g}}_\ell(\boldsymbol{\lambda}_2)\|_2 \le \frac{|\mathcal{E}_\ell|}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \left( 2 \sum_{\ell'<\ell} |\mathcal{E}_{\ell'}| \cdot K^2 \epsilon \epsilon_c + 4 \log \frac{1}{\delta} \right) \sqrt{d}$$

$$= 2|\mathcal{E}_\ell| K^2 \epsilon \epsilon_c \sqrt{d} + \frac{4|\mathcal{E}_\ell| \log \frac{1}{\delta}}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sqrt{d}, \quad \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \boldsymbol{\Lambda} \text{ s.t. } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \le \frac{\epsilon}{2\sqrt{d}}$$

with probability at least $1 - \delta \prod_{j=1}^{d} (2d/\rho_j \epsilon)$. Substituting $\epsilon' = \frac{\epsilon}{2\sqrt{d}}$ and $\delta' = \delta \prod_{j=1}^{d} (2d/\rho_j \epsilon)$ gives the conclusion. □

#### D.3.3 `Stability` Term Bounds in O-FTRL-FP

**Lemma D.6** (Difference between $\{i_t\}_{t\in\mathcal{E}_\ell}$ and $\{\widetilde{i}_t^*\}_{t\in\mathcal{E}_\ell}$)**.** *For any epoch $\ell \in [L]$,*

$$\mathbb{E}\left[ \left\| \sum_{t\in\mathcal{E}_\ell} (\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t}) - \widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell) \right\|_2^2 \right] \le (d+3)|\mathcal{E}_\ell| + N_\ell^2$$

*where $N_\ell$ is defined as in Lemma D.1:*

$$N_\ell = 1 + 4K^2\epsilon_c \|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4),$$

*and we recall that*

$$\widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}) = |\mathcal{E}_\ell| \cdot \mathop{\mathbb{E}}_{\boldsymbol{v}_*\sim\mathcal{V},\boldsymbol{c}_*\sim\mathcal{C}} \left[ \boldsymbol{\rho} - \boldsymbol{c}_{*,\widetilde{i}^*(\boldsymbol{\lambda})} \right], \quad \widetilde{i}^*(\boldsymbol{\lambda}) = \mathop{\mathrm{argmax}}_{i\in[K]} \left( v_{*,i} - \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{c}_{*,i} \right).$$

*Proof.* In this proof, we first control $\mathbb{E}[\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*}) - \widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell)\|_2^2]$, *i.e.*, the squared $\ell_2$-error of $|\mathcal{E}_\ell|$ random vectors from their mean – which is of order $|\mathcal{E}_\ell|$ because they are *i.i.d.* We then relate it to $\mathbb{E}[\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t}) - \widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell)\|_2^2]$ by utilizing the similarity between $\{\widetilde{i}_t^*\}_{t\in\mathcal{E}_\ell}$ and $\{i_t\}_{t\in\mathcal{E}_\ell}$ that we derived in Theorem C.5.

**Step 1: Control $\mathbb{E}[\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*}) - \widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell)\|_2^2]$.** Recall the definition of $\{\widetilde{i}_t^*\}_{t\in[T]}$ from Eq. (9):

$$\widetilde{i}_t^* = \mathop{\mathrm{argmax}}_{i\in[K]} \left( v_{t,i} - \boldsymbol{\lambda}_\ell^\mathsf{T} \boldsymbol{c}_{t,i} \right), \quad \forall t \in \mathcal{E}_\ell,$$

and noticing that $\boldsymbol{v}_t$ and $\boldsymbol{c}_t$ are *i.i.d.* samples from $\mathcal{V}$ and $\mathcal{C}$, we have $\mathrm{PDF}(\widetilde{i}_t^*) = \mathrm{PDF}(\widetilde{i}^*(\boldsymbol{\lambda}_\ell))$ for all $t \in \mathcal{E}_\ell$. Therefore,

$$\mathbb{E}\left[ \boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*} \right] = \mathop{\mathbb{E}}_{\boldsymbol{v}_*\sim\mathcal{V},\boldsymbol{c}_*\sim\mathcal{C}} \left[ \boldsymbol{\rho} - \boldsymbol{c}_{*,\widetilde{i}^*(\boldsymbol{\lambda}_\ell)} \right] = \frac{1}{|\mathcal{E}_\ell|} \widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell), \quad \forall t \in \mathcal{E}_\ell,$$

where the last equation is precisely the definition of $\widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell)$. Since for a $d$-dimensional random vector $X$, $\mathbb{E}[\|X - \mathbb{E}[X]\|_2^2] = \mathbb{E}[\sum_{i=1}^d (X_i - \mathbb{E}[X_i])^2] = \sum_{i=1}^d \mathrm{Var}(X_i) = \mathrm{Tr}(\mathrm{Cov}(X))$ where Tr is the trace and Cov is the covariance matrix, we have

$$\mathbb{E}\left[ \left\| \sum_{t\in\mathcal{E}_\ell} (\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*}) - \widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}_\ell) \right\|_2^2 \right] = \mathrm{Tr}\left( |\mathcal{E}_\ell| \cdot \mathop{\mathrm{Cov}}_{\boldsymbol{v}_*\sim\mathcal{V},\boldsymbol{c}_*\sim\mathcal{C}} \left( \boldsymbol{\rho} - \boldsymbol{c}_{*,\widetilde{i}^*(\boldsymbol{\lambda}_\ell)} \right) \right) \le |\mathcal{E}_\ell| d,$$

using the fact that $\widetilde{i}_t^*$'s are independent from each other and that $\boldsymbol{\rho}$ and $\boldsymbol{c}_{*,i}$ are all within $[0,1]^d$.

**Step 2: Relate $\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}_t^*})$ to $\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t})$.** Recall Eq. (16) from Theorem C.5:

$$\Pr\left\{ \sum_{t\in\mathcal{E}_\ell} \mathbb{1}[i_t \ne \widetilde{i}_t^*] > N_\ell \right\} \le \frac{3}{|\mathcal{E}_\ell|}, \tag{32}$$

where $N_\ell := 1 + 4K^2\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 5\log|\mathcal{E}_\ell| + \log_{\gamma^{-1}}(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^4)$.

As $(\boldsymbol{\rho} - \boldsymbol{c}_{t,i}) \in [-1, 1]^d$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,i_t}) - \widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell)\right\|_2^2\right] \\
&\le \mathbb{E}\left[\left\|\sum_{t\in\mathcal{E}_\ell}(\boldsymbol{\rho} - \boldsymbol{c}_{t,\widetilde{i}^*_t}) - \widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell)\right\|_2^2\right] + \mathbb{E}\left[d\left(\sum_{t\in\mathcal{E}_\ell}\mathbb{1}[i_t \neq \widetilde{i}^*_t]\right)^2\right] \\
&\le |\mathcal{E}_\ell|d + N_\ell^2 + 3\frac{|\mathcal{E}_\ell|^2}{|\mathcal{E}_\ell|},
\end{aligned}$$

where the last term considers the failure probability of Eq. (32), in which case we use the trivial bound $(\sum_{t\in\mathcal{E}_\ell}\mathbb{1}[i_t \neq \widetilde{i}^*_t])^2 \le |\mathcal{E}_\ell|^2$. Rearranging gives the desired conclusion. □

**Lemma D.7** (Empirical Estimation)**.** *For any $\ell \in [L]$, $\epsilon > 0$, and $\delta \in (0, 1)$, with probability $1 - 2\delta$,*

$$\begin{aligned}
\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})\|_2^2 &\le 3d|\mathcal{E}_\ell|^2 \cdot (K^2\epsilon\epsilon_c)^2 + 6d|\mathcal{E}_\ell|^2 \cdot \frac{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \\
&= \widetilde{\mathcal{O}}\left(|\mathcal{E}_\ell|^2\epsilon^2 + \frac{|\mathcal{E}_\ell|^2}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right)\right), \quad \forall\boldsymbol{\lambda} \in \boldsymbol{\Lambda},
\end{aligned}$$

*where we recall that*

$$\begin{aligned}
\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}) &= |\mathcal{E}_\ell| \cdot \frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}^v_\tau(\boldsymbol{\lambda})}\right), \quad \widetilde{i}^v_\tau(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]}\left(v_{\tau,i} - \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right); \\
\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}) &= |\mathcal{E}_\ell| \cdot \mathop{\mathbb{E}}_{\boldsymbol{v}_*\sim\mathcal{V},\boldsymbol{c}_*\sim\mathcal{C}}\left[\boldsymbol{\rho} - \boldsymbol{c}_{*,\widetilde{i}^*(\boldsymbol{\lambda})}\right], \quad \widetilde{i}^*(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]}\left(v_{*,i} - \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{*,i}\right).
\end{aligned}$$

*Proof.* In contrast to Lemma D.6 where we directly applied concentration bounds at the realized dual iterate $\boldsymbol{\lambda}_\ell$, here we cannot proceed in the same way. The reason is that the value samples $\boldsymbol{v}_\tau$ used to compute $\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\ell)$ were drawn in the past, and $\boldsymbol{\lambda}_\ell$ itself is computed based on reports dependent on these values. As a result, conditioning on the event that $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\ell) \approx \widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\ell)$ introduces a dependence on future information from the perspective of those past realizations, violating valid conditioning.

To overcome this, we establish uniform concentration over all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ by discretizing the domain. Specifically, we construct an $\epsilon$-net $\boldsymbol{\Lambda}_\epsilon \subseteq \boldsymbol{\Lambda}$ and first show that for every $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$, the approximation $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon) \approx \widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)$ holds with high probability. We then extend this guarantee to all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ by considering the stochastic process $\{\mathbb{1}[\exists\boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) \text{ s.t. } \widetilde{i}^v_\tau(\boldsymbol{\lambda}) \neq \widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)]\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$ where $\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$ is the $\epsilon$-radius ball centered at $\boldsymbol{\lambda}_\epsilon$. The proof goes in three steps.

**Step 1: Cover $\boldsymbol{\Lambda}$ with an $\epsilon$-net.** From Lemma E.3, for any $\epsilon > 0$, there exists an $\epsilon$-net $\boldsymbol{\Lambda}_\epsilon \subseteq \boldsymbol{\Lambda}$ of size $\mathcal{O}((d/\epsilon)^d)$, such that every $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ has some $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$ with $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_\epsilon\|_1 \le \epsilon$. We remark that our final guarantee does not have a $d$-exponent, because the dependency on $|\boldsymbol{\Lambda}_\epsilon|$ is logarithmic.

**Step 2: Yield concentration for any fixed $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$.** Fix $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$. Let

$$\boldsymbol{x}_\tau := (\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)}) - \mathop{\mathbb{E}}_{\boldsymbol{v}_*\sim\mathcal{V},\boldsymbol{c}_*\sim\mathcal{C}}\left[\boldsymbol{\rho} - \boldsymbol{c}_{*,\widetilde{i}^*(\boldsymbol{\lambda}_\epsilon)}\right], \quad \forall\ell' < \ell, \tau \in \mathcal{E}_{\ell'}.$$

Since $\boldsymbol{x}_\tau$ only depends on $\boldsymbol{v}_\tau$ and $\boldsymbol{c}_\tau$ which are *i.i.d.* samples from $\mathcal{V}$ and $\mathcal{C}$, these vectors are *i.i.d.*, zero-mean, and ensures $\|\boldsymbol{x}_\tau\|_2 \le \sqrt{d}$ *a.s.* since $\boldsymbol{\rho} - \boldsymbol{c}_\tau \in [-1, 1]^d$. Applying the vector Bernstein inequality (Kohler and Lucchi, 2017, Lemma 18) restated as Lemma E.5 gives:

$$\Pr\left\{\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon) - \widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2 \ge |\mathcal{E}_\ell|c\right\} = \Pr\left\{\left\|\frac{\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\boldsymbol{x}_\tau}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\right\|_2 \ge c\right\} \le \exp\left(-\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}| \cdot \frac{c^2}{8d} + \frac{1}{4}\right).$$

Taking a union bound over all $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$, we obtain that

$$\Pr\left\{\max_{\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon}\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon)-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2\geq|\mathcal{E}_\ell|^2c^2\right\}\leq\prod_{i=1}^d\frac{d}{\rho_j\epsilon}\cdot\exp\left(-2\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|\cdot\frac{c^2}{8d}+\frac{1}{4}\right),\quad\forall c>0.$$

Therefore, for the given failure probability $\delta$, with probability at least $1-\delta$,

$$\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon)-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2\leq|\mathcal{E}_\ell|^2\cdot 4d\cdot\frac{\log\frac{1}{\delta}+\sum_{j=1}^d\log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|},\quad\forall\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon.\tag{33}$$

**Step 3: Extend the similarity to all $\boldsymbol{\lambda}\in\boldsymbol{\Lambda}$.** We now fix a $\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon$ and try to ensure a *uniform* concentration guarantee for all $\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$. Using boundedness $\|\boldsymbol{\rho}-\boldsymbol{c}_{\tau,i}\|_2\leq\sqrt{d}$, we have:

$$\|\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2\leq d\left(\frac{|\mathcal{E}_\ell|}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\mathbb{1}[\widetilde{i}^v_\tau(\boldsymbol{\lambda})\neq\widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)]\right)^2,$$

$$\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2\leq d\left(|\mathcal{E}_\ell|\cdot\Pr\{\widetilde{i}^*(\boldsymbol{\lambda})\neq\widetilde{i}^*(\boldsymbol{\lambda}_\epsilon)\}\right)^2,\quad\forall\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon).$$

From Lemma E.3, we know

$$\Pr_{\boldsymbol{v}\sim\mathcal{V},\boldsymbol{c}\sim\mathcal{C}}\left\{\exists\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)\text{ s.t. }\widetilde{i}^*(\boldsymbol{\lambda})\neq\widetilde{i}^*(\boldsymbol{\lambda}_\epsilon)\right\}\leq K^2(\epsilon\cdot\epsilon_c),\quad\forall\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon,\tag{34}$$

where we remark that the $\exists\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$ clause is important since it ensures "uniform smoothness" in the neighborhood of $\boldsymbol{\lambda}_\epsilon$. If we instead fix a $\boldsymbol{\lambda}$ and its corresponding $\boldsymbol{\lambda}_\epsilon$ and apply concentration to this specific $\boldsymbol{\lambda}$, we need to do an prohibitively expensive Union Bound afterwards; see also Footnote 6.

Hence the error between $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})$ and $\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon)$ is bounded by

$$\max_{\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)}\|\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^*_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2\leq d|\mathcal{E}_\ell|^2\cdot(K^2\epsilon\epsilon_c)^2,\quad a.s.\tag{35}$$

For the error between $\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})$ and $\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)$, consider a stochastic process $(X_\tau)_{\tau\geq1}$ adapted to $(\mathcal{F}_\tau)_{\tau\geq0}$:

$$X_\tau:=\mathbb{1}[\exists\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)\text{ s.t. }\widetilde{i}^v_\tau(\boldsymbol{\lambda})\neq\widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)],\quad\forall\ell'<\ell,\tau\in\mathcal{E}_{\ell'},$$

where $\mathcal{F}_\tau=\sigma(\bigcup_{i\in\{0\}\cup[K]}\mathcal{H}_{\tau+1,i})=\sigma(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_\tau,\boldsymbol{u}_1,\ldots,\boldsymbol{u}_\tau,\boldsymbol{c}_1,\ldots,\boldsymbol{c}_\tau,i_1,\ldots,i_\tau,\boldsymbol{p}_1,\ldots,\boldsymbol{p}_\tau)$ is the smallest $\sigma$-algebra containing all generated history up to the end of round $\tau$.[7] Then $X_\tau$ is $\mathcal{F}_\tau$-measurable, and we have

$$\mathbb{E}[X_\tau\mid\mathcal{F}_{\tau-1}]=\Pr_{\boldsymbol{v}\sim\mathcal{V},\boldsymbol{c}\sim\mathcal{C}}\left\{\exists\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)\text{ s.t. }\widetilde{i}^*(\boldsymbol{\lambda})\neq\widetilde{i}^*(\boldsymbol{\lambda}_\epsilon)\right\}\leq K^2(\epsilon\cdot\epsilon_c),$$

where the last step uses Eq. (34). Applying Azuma-Hoeffding to martingale difference sequence $\{X_\tau-\mathbb{E}[X_\tau\mid\mathcal{F}_{\tau-1}]\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$, we know for any $c>0$,

$$\Pr\left\{\frac{\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}\mathbb{1}[\exists\boldsymbol{\lambda}\in\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)\text{ s.t. }\widetilde{i}^v_\tau(\boldsymbol{\lambda})\neq\widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)]}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}\geq K^2\epsilon\epsilon_c+c\right\}\leq\exp\left(-2c^2\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|\right).$$

Applying a Union Bound over all $\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon$ and recalling the expression of $\|\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2$,

$$\Pr\left\{\exists\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon,\boldsymbol{\lambda}\in\boldsymbol{\Lambda}\text{ s.t. }\|\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2\geq d|\mathcal{E}_\ell|^2(K^2\epsilon\epsilon_c+c)^2\right\}\leq\prod_{i=1}^d\frac{d}{\rho_j\epsilon}\cdot\exp\left(-2c^2\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|\right).$$

[7] In fact, the $X_\tau$'s here are *i.i.d.* variables since $\boldsymbol{v}_\tau\sim\mathcal{V}$ and $\boldsymbol{c}_\tau\sim\mathcal{C}$ are independent. However, when controlling $\|\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda})-\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2$ in Lemma D.8, since reports $\boldsymbol{u}_\tau$ can be history-dependent, to make Lemma E.3 still applicable we need to make sure the conditional distribution of reports $\boldsymbol{u}_\tau\mid\mathcal{H}_\tau$ is $\mathcal{F}_{\tau-1}$-measurable.

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

$$\|\widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}_\epsilon)\|_2^2 \le d|\mathcal{E}_\ell|^2 \left( K^2\epsilon\epsilon_c + \frac{\sqrt{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}}{\sqrt{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}} \right)^2, \quad \forall \boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon, \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon). \tag{36}$$

**Final Bound.** Via Union Bound, with probability $1-2\delta$, Eqs. (33), (35) and (36) are all true and

$$\|\widetilde{\boldsymbol{g}}_\ell^*(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda})\|_2^2 \le 3d|\mathcal{E}_\ell|^2 \cdot (K^2\epsilon\epsilon_c)^2 + 6d|\mathcal{E}_\ell|^2 \cdot \frac{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}, \quad \forall \boldsymbol{\lambda} \in \boldsymbol{\Lambda}.$$

This finishes the proof. □

**Lemma D.8** (Untruthful Reports)**.** *For any $\ell \in [L]$, $\epsilon > 0$, and $\delta \in (0,1)$, with probability $1-4\delta$,*

$$\begin{aligned}
&\|\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda})\|_2^2 \\
&\le d|\mathcal{E}_\ell|^2 \left( \frac{\ell \log_{\gamma^{-1}}\left(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3 \cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right)}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \right)^2 \\
&\quad + 4d|\mathcal{E}_\ell|^2 \left( (K^2\epsilon\epsilon_c)^2 + \frac{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \right) \\
&= \widetilde{\mathcal{O}}_{T,\delta,\epsilon}\left( |\mathcal{E}_\ell|^2\epsilon^2 + \frac{|\mathcal{E}_\ell|^2}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \left( L + \log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon} \right)^2 \right), \quad \forall \boldsymbol{\lambda} \in \boldsymbol{\Lambda},
\end{aligned}$$

*where we recall that*

$$\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}) = |\mathcal{E}_\ell| \cdot \frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sum_{\ell'<\ell, \tau\in\mathcal{E}_{\ell'}} \left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau^u(\boldsymbol{\lambda})}\right), \quad \widetilde{i}_\tau^u(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]} \left(u_{\tau,i} - \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right);$$

$$\widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}) = |\mathcal{E}_\ell| \cdot \frac{1}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sum_{\ell'<\ell, \tau\in\mathcal{E}_{\ell'}} \left(\boldsymbol{\rho} - \boldsymbol{c}_{\tau,\widetilde{i}_\tau^v(\boldsymbol{\lambda})}\right), \quad \widetilde{i}_\tau^v(\boldsymbol{\lambda}) = \operatorname*{argmax}_{i\in[K]} \left(v_{\tau,i} - \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{c}_{\tau,i}\right).$$

*Proof.* We now bound the impact of untruthful reporting, specifically the difference between past reported values $\boldsymbol{u}_t$ and true values $\boldsymbol{v}_t$. Similar to the reason in Lemma D.7, we cannot directly apply concentration inequalities to $\boldsymbol{\lambda}_\ell$ which is unmeasurable when the reports are generated. Therefore, we still consider an $\epsilon$-net $\boldsymbol{\Lambda}_\epsilon$ of $\boldsymbol{\Lambda}$, ensure that $\widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}_\epsilon) \approx \widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}_\epsilon)$ for all $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$, and then extend it to all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ via Chernoff-Hoeffding inequalities.

**Step 1: Cover $\boldsymbol{\Lambda}$ with an $\epsilon$-net.** From Lemma E.3, for any $\epsilon > 0$, there exists an $\epsilon$-net $\boldsymbol{\Lambda}_\epsilon \subseteq \boldsymbol{\Lambda}$ of size $\mathcal{O}((d/\epsilon)^d)$, such that every $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ has some $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$ with $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_\epsilon\|_1 \le \epsilon$. We remark that our final guarantee does not have a $d$-exponent, because the dependency on $|\boldsymbol{\Lambda}_\epsilon|$ is logarithmic.

**Step 2: Yield concentration for any fixed $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$.** Fix $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$. From Lemma C.3, which underpins the INTEREPOCH analysis in Theorem C.5, we know that for any previous epoch $\ell' < \ell$, the event $|u_{\tau,i} - v_{\tau,i}| \ge \frac{1}{|\mathcal{E}_{\ell'}|}$ occurs in only $\widetilde{\mathcal{O}}(1)$ rounds (with $\tau \in \mathcal{E}_{\ell'}$) with probability at least $1 - \frac{1}{|\mathcal{E}_{\ell'}|}$. Following the approach in Lemma C.4, we now leverage the smoothness of costs in Assumption 3 in combination with Azuma-Hoeffding inequalities to conclude that $\widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}_\epsilon) \approx \widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}_\epsilon)$.

Formally, to compare $\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}_\epsilon)$ and $\widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}_\epsilon)$, we need only to control the number of previous rounds $\tau$ such that $\widetilde{i}_\tau^u(\boldsymbol{\lambda}_\epsilon) \ne \widetilde{i}_\tau^v(\boldsymbol{\lambda}_\epsilon)$. We decompose such events by whether large misreports happen:

$$\begin{aligned}
\sum_{\ell'<\ell,\,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}\left[\widetilde{i}_\tau^u(\boldsymbol{\lambda}_\epsilon) \ne \widetilde{i}_\tau^v(\boldsymbol{\lambda}_\epsilon)\right] &\le \sum_{\ell'<\ell,\,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}\left[\exists i \in [K] \text{ s.t. } |u_{\tau,i} - v_{\tau,i}| \ge \frac{1}{|\mathcal{E}_{\ell'}|}\right] \\
&\quad + \sum_{\ell'<\ell,\,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}\left[\exists i \ne j \in [K] \text{ s.t. } (v_{\tau,i} - v_{\tau,j}) - \boldsymbol{\lambda}_\epsilon^\mathsf{T}(\boldsymbol{c}_{\tau,i} - \boldsymbol{c}_{\tau,j}) \in \left[0, \frac{2}{|\mathcal{E}_{\ell'}|}\right]\right],
\end{aligned} \tag{37}$$

where the second term plugs $|u_{\tau,i} - v_{\tau,i}| \geq \frac{1}{|\mathcal{E}_{\ell'}|}, \forall i \in [K]$ into definitions of $\widetilde{i}^u_\tau(\boldsymbol{\lambda}_\epsilon)$ and $\widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)$.

For the first term, we use the following inequality which appeared as Eq. (13) in Lemma C.3:

$$\Pr\left\{\sum_{\tau\in\mathcal{E}_{\ell'}} \mathbb{1}\left[|u_{\tau,i} - v_{\tau,i}| \geq \frac{1}{|\mathcal{E}_{\ell'}|}\right] \geq c\right\} \leq 2(1 + 2\|\boldsymbol{\rho}^{-1}\|_1) \cdot \frac{2K|\mathcal{E}_{\ell'}|^3}{\gamma^{-c} - 1}, \quad \forall \ell' < \ell, c > 0.$$

For any fixed failure probability $\delta \in (0, 1)$, for every $\ell' < \ell$, picking $c$ so that the RHS is $\frac{\delta}{\ell}$ gives

$$\Pr\left\{\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}\left[|u_{\tau,i} - v_{\tau,i}| \geq \frac{1}{|\mathcal{E}_{\ell'}|}\right] \geq \ell \log_{\gamma^{-1}}\left(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3 \cdot \ell\delta^{-1}\right)\right\} \leq \delta.$$

For the second term, under Assumption 3 that $\mathrm{PDF}(\boldsymbol{\lambda}_\epsilon^\mathsf{T}\boldsymbol{c}_{\tau,i})$ is uniformly bounded by $\epsilon_c, \forall i \in [K]$,

$$\Pr\left\{(v_{\tau,i} - v_{\tau,j}) - \boldsymbol{\lambda}_\epsilon^\mathsf{T}(\boldsymbol{c}_{\tau,i} - \boldsymbol{c}_{\tau,j}) \in \left[0, \frac{2}{|\mathcal{E}_{\ell'}|}\right]\right\} \leq \frac{2}{|\mathcal{E}_{\ell'}|}\epsilon_c\|\boldsymbol{\lambda}_\epsilon\|_1, \ \ \forall i \neq j \in [K], \ell' < \ell, \tau \in \mathcal{E}_{\ell'}.$$

Although we are now standing at epoch $\ell$, the $\epsilon$-nets are fixed before the game (as it only depends on $\boldsymbol{\Lambda}$). Hence, the indicator $X_\tau := \mathbb{1}[(v_{\tau,i} - v_{\tau,j}) - \boldsymbol{\lambda}_\epsilon^\mathsf{T}(\boldsymbol{c}_{\tau,i} - \boldsymbol{c}_{\tau,j}) \in [0, \frac{2}{|\mathcal{E}_{\ell'}|}]]$ is indeed $\mathcal{F}_\tau$-measurable back in the past when $\tau \in \mathcal{E}_{\ell'}$ and $\ell' < \ell$, where $(\mathcal{F}_\tau)_{\tau\geq 0}$ is the natural filtration $\mathcal{F}_\tau = \sigma(X_1, \ldots, X_\tau)$. Thus, applying multiplicative Azuma-Hoeffding inequality in Lemma E.4 to the martingale difference sequence $\{X_\tau - \mathbb{E}[X_\tau \mid \mathcal{F}_{\tau-1}]\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$, we get

$$\Pr\left\{\frac{1}{2}\sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} X_\tau \geq \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{E}[X_\tau \mid \mathcal{F}_{\tau-1}] + 2A\right\} \leq \exp(-A), \quad \forall A \in \mathbb{R}.$$

Since $X_\tau$ only involves $\boldsymbol{v}_\tau \sim \mathcal{V}$ and $\boldsymbol{c}_\tau \sim \mathcal{C}$ which are *i.i.d.*, we know $\mathbb{E}[X_\tau \mid \mathcal{F}_{\tau-1}] = \mathbb{E}[X_\tau] \leq \frac{2}{|\mathcal{E}_{\ell'}|}\epsilon_c\|\boldsymbol{\lambda}_\epsilon\|_1$. Setting the RHS as $\frac{\delta}{|\boldsymbol{\Lambda}_\epsilon|K^2}$ and taking a Union Bound over all $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$, we have

$$\Pr\left\{\max_{\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon} \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}\left[(v_{\tau,i} - v_{\tau,j}) - \boldsymbol{\lambda}_\epsilon^\mathsf{T}(\boldsymbol{c}_{\tau,i} - \boldsymbol{c}_{\tau,j}) \in \left[0, \frac{2}{|\mathcal{E}_{\ell'}|}\right]\right]\right.$$
$$\left. \geq 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right)\right\} \leq \frac{\delta}{K^2}, \quad \forall i \neq j \in [K].$$

Using another Union Bound over all $i \neq j \in [K]$ and plugging it back into Eq. (37), we get

$$\max_{\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon} \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}[\widetilde{i}^u_\tau(\boldsymbol{\lambda}_\epsilon) \neq \widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)]$$
$$\leq \ell \log_{\gamma^{-1}}\left(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3 \cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right),$$

with probability at least $1 - 2\delta$. Therefore, with probability at least $1 - 2\delta$, we have

$$\max_{\boldsymbol{\lambda}_\epsilon\in\boldsymbol{\Lambda}_\epsilon} \|\widetilde{\boldsymbol{g}}^u_\ell(\boldsymbol{\lambda}_\epsilon) - \widetilde{\boldsymbol{g}}^v_\ell(\boldsymbol{\lambda}_\epsilon)\|_2^2 \leq d\left(\frac{|\mathcal{E}_\ell|}{\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|} \sum_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}} \mathbb{1}[\widetilde{i}^u_\tau(\boldsymbol{\lambda}_\epsilon) \neq \widetilde{i}^v_\tau(\boldsymbol{\lambda}_\epsilon)]\right)^2$$
$$\leq \frac{d|\mathcal{E}_\ell|^2}{(\sum_{\ell'<\ell}|\mathcal{E}_{\ell'}|)^2}\left(\ell \log_{\gamma^{-1}}\left(1 + 4(1 + \|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3 \cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1\right.$$
$$\left. + 4\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right)\right)^2,$$

where the first inequality uses the boundedness of $\|\boldsymbol{\rho} - \boldsymbol{c}_{\tau,i}\|_2^2 \le d$.

**Step 3: Extend the similarity to all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$.** After yielding the similarity that $\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}_\epsilon) \approx \widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}_\epsilon)$ for all $\boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon$, we extend it to all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ using the arguments already derived in Lemma D.7. Recall from Eq. (36) that we already proved that with probability $1-\delta$,

$$\|\widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda}_\epsilon)\|_2^2 \le d|\mathcal{E}_\ell|^2 \left( K^2\epsilon\epsilon_c + \frac{\sqrt{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}}{\sqrt{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}} \right)^2, \quad \forall \boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon, \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon).$$

Using exactly the same arguments (see Footnote 7 for the reason why Lemma E.3 is still applicable when reports $\{\boldsymbol{u}_\tau\}_{\ell'<\ell,\tau\in\mathcal{E}_{\ell'}}$ can be history-dependent), with probability $1-\delta$,

$$\|\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}_\epsilon)\|_2^2 \le d|\mathcal{E}_\ell|^2 \left( K^2\epsilon\epsilon_c + \frac{\sqrt{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}}{\sqrt{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|}} \right)^2, \quad \forall \boldsymbol{\lambda}_\epsilon \in \boldsymbol{\Lambda}_\epsilon, \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon).$$

**Final Bound.** Putting the three inequalities together and taking Union Bound,

$$\begin{aligned}
&\sup_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \|\widetilde{\boldsymbol{g}}_\ell^u(\boldsymbol{\lambda}) - \widetilde{\boldsymbol{g}}_\ell^v(\boldsymbol{\lambda})\|_2^2 \\
&\le d|\mathcal{E}_\ell|^2 \left( \frac{\ell \log_{\gamma^{-1}}\left(1 + 4(1+\|\boldsymbol{\rho}^{-1}\|_1)K|\mathcal{E}_\ell|^3 \cdot 4\ell\delta^{-1}\right) + 4\ell\epsilon_c\|\boldsymbol{\rho}^{-1}\|_1 + 4\left(\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}\right)}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \right)^2 \\
&\quad + 4d|\mathcal{E}_\ell|^2 \left( (K^2\epsilon\epsilon_c)^2 + \frac{\log\frac{1}{\delta} + \sum_{j=1}^d \log\frac{d}{\rho_j\epsilon}}{\sum_{\ell'=1}^{\ell-1}|\mathcal{E}_{\ell'}|} \right), \quad \text{w.p. } 1-4\delta.
\end{aligned}$$

This completes the proof. □

# E Auxiliary Lemmas

We first include several classical online learning guarantees.

**Lemma E.1** (FTRL Guarantee (Orabona, 2019, Corollary 7.7)). *For a convex region $\mathcal{X}$, an 1-strongly-convex regularizer $\Psi\colon \mathcal{X} \to \mathbb{R}$ w.r.t. some norm $\|\cdot\|$ with $\min_{\boldsymbol{x}\in\mathcal{X}} \Psi(\boldsymbol{x}) = 0$, a sequence of convex and differentiable losses $f_1, f_2, \dots, f_R$, a sequence of learning rates $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_R \ge 0$,*

$$\boldsymbol{x}_r = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( \sum_{\mathfrak{r}=1}^{r-1} f_\mathfrak{r}(\boldsymbol{x}) + \frac{1}{\eta_r}\Psi(\boldsymbol{x}) \right), \quad \forall r = 1, 2, \dots, R,$$

*we have the following regret guarantee where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$:*

$$\sum_{r=1}^R (f_r(\boldsymbol{x}_r) - f_r(\boldsymbol{x}^*)) \le \frac{\Psi(\boldsymbol{x}^*)}{\eta_R} + \frac{1}{2}\sum_{r=1}^R \eta_r \|\nabla f_r(\boldsymbol{x}_r)\|_*^2, \quad \forall \boldsymbol{x}^* \in \mathcal{X}.$$

**Lemma E.2** (O-FTRL Guarantee (Orabona, 2019, Theorem 7.39)). *For a convex region $\mathcal{X}$, an 1-strongly-convex regularizer $\Psi\colon \mathcal{X} \to \mathbb{R}$ w.r.t. some norm $\|\cdot\|$ with $\min_{\boldsymbol{x}\in\mathcal{X}} \Psi(\boldsymbol{x}) = 0$, a sequence of convex and differentiable losses $f_1, f_2, \dots, f_R$, a sequence of learning rates $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_R \ge 0$, a (stochastic) prediction sequence $\{\widehat{f}_r\colon \mathcal{X} \to \mathbb{R}\}_{r\in[R]}$ that is $\mathcal{F}_{r-1}$-measurable (where $(\mathcal{F}_r)_t$ is the natural filtration that $\mathcal{F}_r = \sigma(f_1, f_2, \dots, f_r)$), and*

$$\boldsymbol{x}_r = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \left( \sum_{\mathfrak{r}=1}^{r-1} f_\mathfrak{r}(\boldsymbol{x}) + \widehat{\ell}_r(\boldsymbol{x}) + \frac{1}{\eta_r}\Psi(\boldsymbol{x}) \right), \quad \forall r = 1, 2, \dots, R,$$

*we have the following regret guarantee where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$:*

$$\sum_{r=1}^R (f_r(\boldsymbol{x}_r) - f_r(\boldsymbol{x}^*)) \le \frac{\Psi(\boldsymbol{x}^*)}{\eta_R} + \frac{1}{2}\sum_{r=1}^R \eta_r \|\nabla(f_r - \widehat{f}_r)(\boldsymbol{x}_r)\|_*^2, \quad \forall \boldsymbol{x}^* \in \mathcal{X}.$$

We now present a covering of the dual variable space $\mathbf{\Lambda}$.

**Lemma E.3** (Covering of Dual Variables). *For any fixed constant $\epsilon > 0$, there exists a subset $\mathbf{\Lambda}_\epsilon$ of $\mathbf{\Lambda} := \bigotimes_{j=1}^d [0, \rho_j^{-1}]$ that has a size no more than $\prod_{j=1}^d (d/\rho_j \epsilon)$ and ensures*

$$\forall \boldsymbol{\lambda} \in \mathbf{\Lambda}, \quad \exists \boldsymbol{\lambda}_\epsilon \in \mathbf{\Lambda}_\epsilon \text{ s.t. } \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_\epsilon\|_1 \le \epsilon.$$

*Furthermore, for any $\epsilon$-net $\mathbf{\Lambda}_\epsilon$ such that $\mathbf{\Lambda} \subseteq \bigcup_{\boldsymbol{\lambda}_\epsilon \in \mathbf{\Lambda}_\epsilon} \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$ where $\mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) = \{\boldsymbol{\lambda} \in \mathbf{\Lambda} \mid \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_\epsilon\|_1 \le \epsilon\}$ is the neighborhood of $\boldsymbol{\lambda}_\epsilon \in \mathbf{\Lambda}_\epsilon$, under Assumption 3, we have for all $\boldsymbol{\lambda}_\epsilon \in \mathbf{\Lambda}_\epsilon$ and any distribution $\mathcal{U} \in \triangle([0,1]^K)$ that*

$$\Pr_{\boldsymbol{u} \sim \mathcal{U}, \boldsymbol{c} \sim \mathcal{C}} \left\{ \exists \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) \text{ s.t. } \operatorname*{argmax}_{i \in [K]} (u_i - \boldsymbol{\lambda}^\intercal \boldsymbol{c}_i) \neq \operatorname*{argmax}_{i \in [K]} (u_i - \boldsymbol{\lambda}_\epsilon^\intercal \boldsymbol{c}_i) \right\} \le K^2 (\epsilon \cdot \epsilon_c).$$

*Proof.* The first claim is standard from covering arguments and the fact that $\mathbf{\Lambda} = \bigotimes_{j=1}^d [0, \rho_j^{-1}]$ is bounded. For the second part, we make use of Assumption 3: For any fixed $i \neq j \in [K]$,

$$\begin{aligned}
&\Pr_{\boldsymbol{u} \sim \mathcal{U}, \boldsymbol{c} \sim \mathcal{C}} \left\{ \exists \boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon) \text{ s.t. } (u_i - \boldsymbol{\lambda}_\epsilon^\intercal \boldsymbol{c}_i > u_j - \boldsymbol{\lambda}_\epsilon^\intercal \boldsymbol{c}_j) \wedge (u_i - \boldsymbol{\lambda}^\intercal \boldsymbol{c}_i < u_j - \boldsymbol{\lambda}^\intercal \boldsymbol{c}_j) \right\} \\
\le &\Pr_{\boldsymbol{u} \sim \mathcal{U}, \boldsymbol{c} \sim \mathcal{C}} \left\{ 0 \le (u_i - \boldsymbol{\lambda}_\epsilon^\intercal \boldsymbol{c}_i) - (u_j - \boldsymbol{\lambda}_\epsilon^\intercal \boldsymbol{c}_j) \le \epsilon \right\} \le \epsilon \cdot \epsilon_c,
\end{aligned}$$

where the first inequality uses $|\langle \boldsymbol{\lambda} - \boldsymbol{\lambda}_\epsilon, \boldsymbol{c}_i - \boldsymbol{c}_j \rangle| \le \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_\epsilon\|_1 \cdot \|\boldsymbol{c}_i - \boldsymbol{c}_j\|_\infty \le \epsilon$ for all $\boldsymbol{\lambda} \in \mathcal{B}_\epsilon(\boldsymbol{\lambda}_\epsilon)$, while the second uses Assumption 3 and the independence of $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$: If two independent real-valued random variables $X \perp Y$ have their PDFs $f_X$ and $f_Y$ uniformly bounded by $\epsilon_c$, then

$$f_{X-Y}(z) = \int_{-\infty}^{\infty} f_X(z+y) f_Y(y) dy \le \epsilon_c \int_{-\infty}^{\infty} f_Y(y) dy \le \epsilon_c, \quad \forall z \in \mathbb{R}.$$

Applying Union Bound to the $K^2$ pairs of $(i, j) \in [K] \times [K]$ gives the second conclusion. □

Finally, we introduce some martingale or random variable concentration inequalities.

**Lemma E.4** (Multiplicative Azuma-Hoeffding Inequality (Koufogiannakis and Young, 2014, Lemma 10)). *Let $\mathcal{T}$ be a stopping time such that $\mathbb{E}[\mathcal{T}] < \infty$. Let $\{X_t\}_{t \ge 1}$ and $\{Y_t\}_{t \ge 1}$ be two sequence of random variables such that*

$$|X_t - Y_t| \le 1 \text{ a.s.}, \mathbb{E}\left[ X_t - Y_t \,\middle|\, \sum_{\tau < t} X_\tau, \sum_{\tau < t} Y_\tau \right] \le 0, \quad \forall 1 \le t \le \mathcal{T},$$

*then for any $\epsilon \in [0,1]$ and $A \in \mathbb{R}$, we have*

$$\Pr\left\{ (1-\epsilon) \sum_{t=1}^{\mathcal{T}} X_t \ge \sum_{t=1}^{\mathcal{T}} Y_t + A \right\} \le \exp(-\epsilon A).$$

**Lemma E.5** (Vector Bernstein Inequality (Kohler and Lucchi, 2017, Lemma 18)). *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be independent random $d$-dimensional vectors such that they are zero-mean $\mathbb{E}[\boldsymbol{x}_i] = \mathbf{0}$, uniformly bounded $\|\boldsymbol{x}_i\|_2 \le C$ a.s. for some $C > 0$, and have bounded variance $\mathbb{E}[\|\boldsymbol{x}_i\|^2] \le \sigma^2$ for some $\sigma > 0$. Let $\boldsymbol{z} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$, then*

$$\Pr\{\|\boldsymbol{z}\|_2 \ge \epsilon\} \le \exp\left( -n \frac{\epsilon^2}{8\sigma^2} + \frac{1}{4} \right), \quad \forall 0 < \epsilon < \frac{\sigma^2}{C}.$$