Bi-Level Online Provisioning and Scheduling with Switching Costs and Cross-Level Constraints

Jialei Liu
Dept. of Electrical Engineering
University at Buffalo, Buffalo, NY
jialeili@buffalo.edu

C. Emre Koksal
Dept. of Electrical and Computer Engineering
The Ohio State University, Columbus, OH
koksal.2@osu.edu

Ming Shi
Dept. of Electrical Engineering
University at Buffalo, Buffalo, NY
mshi24@buffalo.edu

***Abstract*—We study a bi-level online provisioning and scheduling problem motivated by network resource allocation, where provisioning decisions are made at a slow time scale while queue-/state-dependent scheduling is performed at a fast time scale. We model this two-time-scale interaction using an upper-level online convex optimization (OCO) problem and a lower-level constrained Markov decision process (CMDP). Existing OCO typically assumes stateless decisions and thus cannot capture MDP network dynamics such as queue evolution. Meanwhile, CMDP algorithms typically assume a fixed constraint threshold, whereas in provisioning-and-scheduling systems, the threshold varies with online budget decisions. To address these gaps, we study bi-level OCO-CMDP learning under switching costs (budget reprovisioning/system reconfiguration) and cross-level constraints that couple budgets to scheduling decisions. We build an algorithm to solve this bi-level problem by developing several new techniques, including a carefully designed dual feedback that returns the budget multiplier as sensitivity information for the upper-level update and a lower level that solves a budget-adaptive safe exploration problem via an extended occupancy-measure linear program. We establish near-optimal regret and high-probability satisfaction of the cross-level constraints.**

***Index Terms*—bi-level provisioning and scheduling, two-time-scale decision making, online convex optimization, constrained Markov decision process, switching costs, cross-level budget constraints, reinforcement learning**

I. Introduction

Modern networking systems, such as 5G/6G network and edge computing, must support heterogeneous services (e.g., video streaming, augmented reality, virtual reality, and autonomous driving) over shared wireless infrastructure [1]. A central operational challenge in these systems is online resource provisioning under uncertainty. In particular, a network operator must decide how much bandwidth/compute/energy to allocate to a slice over time, balancing quality of service (QoS) against operational expenditure (OpEx). Practical systems exhibit *multiple coupled time scales*. As illustrated in our motivating example online bandwidth provisioning with queue-aware scheduling (see Fig. 1), the operator provisions an episode-level resource budget at a slow time scale (e.g., bandwidth reserved for a slice over seconds/minutes). This budget directly incurs OpEx and influences the QoS that can be achieved through scheduling. Within each episode, the scheduler performs state-dependent, packet-level actions at a fast time scale (e.g., allocating resource blocks every millisecond). These actions are *constrained* by the provisioned

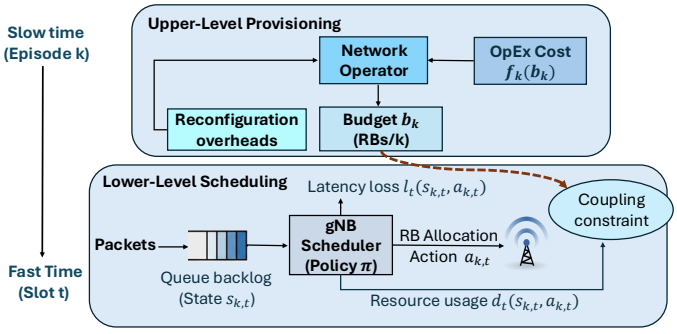

Fig. 1. Online bandwidth provisioning with queue-aware scheduling.

resource budget and aim to reduce latency and congestion, improving QoS [2], [3]. Moreover, changing the provisioning level across episodes is not free. It induces reconfiguration overheads (e.g., hardware retuning, virtualization/migration cost, and instability), which are naturally captured by *switching costs* [4], [5]. See Sec. III for details.

Online convex optimization (OCO) is a classical abstraction for sequential resource allocation and has been extensively studied under adversarial and stochastic settings [4]–[6]. Recent work incorporates switching costs and related operational constraints to penalize aggressive changes in decisions, which is attractive for network slicing. However, OCO with switching costs typically does not model endogenous Markovian state evolution nor provide safety under unknown state dynamics. For example, the per-round loss depends on the current decision and does not capture the evolution of system states, while in networked systems, state variables such as queue lengths evolve endogenously. Consequently, provisioning decisions that appear cost-effective in the short term may cause persistent congestion and latency violations.

Stateful scheduling problems are naturally modeled as Markov decision processes (MDPs), and a rich literature develops provably efficient reinforcement learning (RL) algorithms [7]–[9]. Yet, existing constrained MDP (CMDP) works typically assume that the budget, i.e., the constraint threshold, is *fixed and exogenously specified* [8], [9]. However, in practice, the operator must learn what and how much budget to provision, while accounting for switching overhead. Thus, the constraint threshold itself is a primary decision variable.

Moreover, hierarchical RL (HRL) is a widely studied learning framework for multi-time-scale decision making [10]. However, its objective structure differs fundamentally from our resource provisioning problem. HRL typically optimizes a single, unified reward through task decomposition. In contrast, our problem involves a bi-level trade-off between provisioning cost and service performance. Moreover, HRL generally allows frequent switching between sub-policies.

Motivated by the above gaps, we propose a bi-level online optimization and learning framework that integrates *online budget provisioning* with *state-aware safe scheduling* (without violating cross-level budget constraints). At the upper level, an OCO operator optimizes and provisions budgets $\{b_k\}_{k=1}^K$ across episodes, incurring a convex service cost and a quadratic switching cost. At the lower level, the system evolves as an episodic CMDP under *varying* thresholds. Given the optimized provisioning budget b_k, the agent must learn a safe policy via RL under unknown dynamics. The two levels are coupled through an episode-wise budget constraint and an episode-wise objective that aggregates cross-level costs.

Designing an online algorithm for this coupled problem is technically challenging. *(i) The upper-level effective optimization loss is implicit and non-stationary.* Beyond the unknown convex service cost and switching penalty, it depends on the unknown optimal performance of the lower-level controller. To address this, we leverage a carefully designed dual variable of the budget constraint counterpart to construct an approximated subgradient for evaluating the upper-level sensitivity. *(ii) The lower-level CMDP faces a varying constraint threshold, since budget b_k changes.* This renders standard safe/constrained RL methods with fixed constraints inapplicable. We address this by designing a budget-adaptive exploration algorithm. In summary, our main contributions are as follows.

- We propose a new bi-level online optimization and learning algorithm for provisioning and scheduling with budget switching costs and cross-level budget constraints. At the upper level, we leverage dual multipliers returned by the lower-level (interpretable as the marginal value of extra budget), yielding a subgradient signal despite unknown stateful dynamics. At the lower level, we solve the episode-varying (decision-dependent) budget via budget-adaptive safe exploration and an extended occupancy-measure linear programming (LP), which both enforces feasibility and produces the dual feedback needed to couple the two levels.
- We establish near-optimal regret bounds with high-probability safety guarantees. Our analysis introduces a bi-level regret decomposition that explicitly quantifies how lower-level model-estimation and dual-signal errors propagate into the upper-level OCO updates, and it leverages the extended-LP strong duality to couple constraint satisfaction with sensitivity-based budget learning under switching costs.
- We instantiate the framework in a wireless network scenario, modeling the coupling between episode-level bandwidth provisioning and queue scheduling. Compared to decoupled baselines, our approach achieves a balance between cost and latency while maintaining strict physical feasibility.

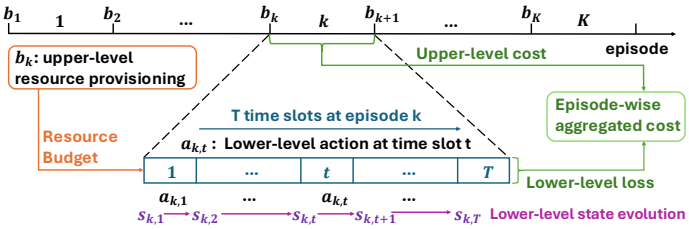

Fig. 2. Coupled two-time-scale decision making: at the beginning of episode k, the agent provisions a resource budget b_k; within the episode, it follows a state-dependent policy selecting actions $a_{k,t}$ over T time slots, incurring losses that aggregate to the episode cost and resource usage that must satisfy the cross-level budget constraint.

Notations: $\Delta(\mathcal{X})$ denotes the probability simplex over a given set $\mathcal{X}$. The operator $a \vee b$ denotes $\max\{a, b\}$. We use $[T] = \{1, \ldots, T\}$ and $[x]^+ = \max\{0, x\}$.

II. Problem Formulation

We study a sequential decision-making problem with *two coupled time scales*. An agent interacts with an environment over K episodes (slow time scale). Within each episode, the environment evolves over T time slots (fast time scale). In each episode $k \in [K]$, the agent chooses an *upper-level* budget decision b_k once per episode and a *lower-level* sequence of state-dependent actions $\{a_{k,t}\}_{t=1}^T$ within the episode. The two levels are coupled through an episode-wise budget constraint (relating b_k to $\{a_{k,t}\}_{t=1}^T$) and an episode-wise objective aggregating upper-level and lower-level costs, as depicted in Fig. 2.

A. Upper-Level Budget Provisioning

At the slow time scale, at the beginning of each episode $k \in [K]$, the agent selects an upper-level budget $b_k \in \mathcal{B} \triangleq [B_0, T]$, which can represent the episode-level resource budget provisioned for a network slice (e.g., bandwidth) over seconds/minutes. Choosing b_k incurs an episode-dependent convex service cost $f_k(b_k)$, which is *not* revealed before the decision is made and is revealed at the end of episode k. As is typical in online convex optimization with switching costs [4], [5], [11], the service cost $f_k(\cdot)$ could change arbitrarily, and it is θ_f-strongly convex and differentiable on $\mathcal{B}$ with bounded gradients $\|\nabla f_k(\cdot)\| \le F$.

In addition, we consider the switching cost, i.e., changing the budget across episodes incurs an additional cost $\alpha\|b_k - b_{k-1}\|^2$, where $b_0 = 0$, and $\alpha > 0$ quantifies reconfiguration overhead (e.g., signaling and control-plane costs, hardware retuning delays, and virtualization/migration overhead).

B. Lower-Level Stateful Scheduling

At the fast time scale, within each episode, the environment evolves over $t = 1, \ldots, T$ time slots according to a finite-horizon (possibly time-inhomogeneous) MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, \{l_t\}_{t=1}^T, \{d_t\}_{t=1}^T, \{P_t\}_{t=1}^T)$. Here, $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $l_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the stage loss, $d_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the per-step resource consumption, and $P_t(\cdot \mid s, a) \in \Delta(\mathcal{S})$ is the *unknown* transition kernel.

At each time t of episode k, the agent observes the current state $s_{k,t}$ (e.g., queue backlogs), and then selects a (possibly randomized) scheduling/allocation action $a_{k,t} \sim \pi_{k,t}(\cdot \mid s_{k,t})$, where $\pi_k^{\mathrm{A}} = (\pi_{k,t})_{t=1}^T$ is a policy with $\pi_{k,t} : \mathcal{S} \to \Delta(\mathcal{A})$. The agent then incurs a stage loss $l_t(s_{k,t}, a_{k,t})$ (e.g., delay/backlog penalty), consumes resources $d_t(s_{k,t}, a_{k,t})$ (e.g., used resource blocks), and the system evolves to next state $s_{k,t+1} \sim P_t(\cdot \mid s_{k,t}, a_{k,t})$ due to unknown arrivals and service dynamics.

For compactness, we define the expected cumulative loss and consumption under policy π and model P as follows,

$$V_r(\pi; P) \triangleq \mathbb{E}_{\pi,P}\left[\sum\nolimits_{t=1}^{T} r_t(s_t, a_t)\right], \tag{1}$$

where $r \in \{l, d\}$, and the expectation is taken with respect to the randomness of actions under π and states under P.

C. Cross-Level Coupling

The decisions in the two levels are coupled as follows.

1) ***Coupling 1 (Cross-Level Budget Constraint)****:* Given upper-level budget b_k, the lower-level policy executed in episode k must satisfy the episode-wise budget constraint, i.e.,

$$V_d(\pi_k^{\mathrm{A}}; P) \le b_k. \tag{2}$$

We define the feasible policy set under budget b as

$$\Pi(b) \triangleq \{\pi^{\mathrm{A}} : V_d(\pi^{\mathrm{A}}; P) \le b\}. \tag{3}$$

Then, (2) is equivalently $\pi_k^{\mathrm{A}} \in \Pi(b_k)$. This is a *cross-level* constraint because its left-hand side is determined by lower-level actions while its right-hand side is the upper-level budget.

2) ***Coupling 2 (Cross-Level Joint Cost)****:* Given $(b_k, \pi_k^{\mathrm{A}})$, we define the episode cost as follows,

$$C_k(b_k, \pi_k^{\mathrm{A}}) \triangleq f_k(b_k) + \alpha\|b_k - b_{k-1}\|^2 + \beta V_l(\pi_k^{\mathrm{A}}; P), \tag{4}$$

where $\beta > 0$ weighs within-episode performance relative to provisioning and reconfiguration costs. The first two terms capture upper-level operating and switching overhead, and the last term captures the cumulative loss induced by stateful allocation/control within the episode, subject to $\pi_k^{\mathrm{A}} \in \Pi(b_k)$.

D. Online Optimization Objective and Challenges

The agent aims to minimize the cumulative cross-level joint cost over K episodes under the cross-level constraints, i.e.,

$$\min_{\{b_k, \pi_k^{\mathrm{A}}\}_{k=1}^K} \quad \sum\nolimits_{k=1}^{K} C_k(b_k, \pi_k^{\mathrm{A}}) \tag{5a}$$

$$\text{s.t.} \quad b_k \in \mathcal{B}, \pi_k^{\mathrm{A}} \in \Pi(b_k), \text{ for all } k \in [K]. \tag{5b}$$

Solving (5) online is non-trivial due to the two couplings. First, the upper-level decision must be made without knowing f_k in advance and its effective loss depends on the (learned) lower-level performance, since the value induced by π_k^{A} is only realized through stateful dynamics with unknown transitions. Second, the lower level faces a constraint whose threshold is episode-dependent (through b_k), and thus requires safety guarantees that remain valid uniformly over the varying budget. Hence, traditional CMDP cannot be applied anymore.

E. Performance Metric

We evaluate an online algorithm using the static regret, which compares against the best fixed budget and fixed lower-level policy in hindsight. Define the best static optimal solution

$$(b^*, \pi^{\mathrm{A},*}) \in \arg\min\nolimits_{b \in \mathcal{B}, \pi^{\mathrm{A}} \in \Pi(b)} \sum\nolimits_{k=1}^{K} C_k(b, \pi^{\mathrm{A}}), \tag{6}$$

where $\pi^{\mathrm{A},*}$ could be non-stationary in an episode, i.e., $\pi^{\mathrm{A},*} = (\pi_t^*)_{t=1}^T$. The static regret after K episodes is then defined as

$$\mathrm{Reg}(K, b, \pi) \triangleq \sum\nolimits_{k=1}^{K} [C_k(b_k, \pi_k^{\mathrm{A}}) - C_k(b^*, \pi^{\mathrm{A},*})]. \tag{7}$$

A sublinear bound $\mathrm{Reg}(K, b, \pi) = o(K)$ implies that the average performance of the online algorithm converges to that of the best fixed budget-policy pair in hindsight.

III. Motivating Example

The formulation in Sec. II captures network settings where a *slow* provisioning decision must be coordinated with *fast*, MDP-dependent scheduling, e.g., queue-aware stochastic network control and scheduling [2], [3], distributed low-complexity scheduling under interference constraints [12], and delay-sensitive wireless scheduling with queue information [13]. We instantiate the framework in **online bandwidth provisioning with queue-aware scheduling**, matching the abstraction in Fig. 1 and the two-time-scale timeline in Fig. 2.

a) Scenario (two time scales): Consider a next-generation NodeB (gNB) serving an ultra-reliable low-latency communications (URLLC) slice with stochastic packet arrivals in a radio access network (RAN). Slice provisioning/configuration is updated over coarse control intervals (seconds/minutes), while the medium access control (MAC) scheduler allocates resource blocks (RBs) at fine granularity (milliseconds) in response to queue and channel fluctuations.

b) Upper level (provisioning with switching overhead): At the beginning of episode k, the operator provisions a bandwidth budget b_k, e.g., total RB·slot units over T slots. Provisioning incurs an operating cost $f_k(b_k)$ (spectrum leasing, activation energy, or opportunity cost relative to other slices), which may vary with network conditions and is revealed only after the episode. Reconfiguring the provisioned bandwidth across episodes induces overhead, modeled by $\alpha\|b_k - b_{k-1}\|^2$.

c) Lower level (queue-aware scheduling under a budget): Within episode k, over slots $t = 1, \ldots, T$, the scheduler observes the state $s_{k,t}$ (e.g., slice queue backlog) and selects an action $a_{k,t}$ (RB allocation/transmission decision). The action incurs a stage loss $l_t(s_{k,t}, a_{k,t})$ (delay/backlog penalty reflecting QoS) and consumes resources $d_t(s_{k,t}, a_{k,t})$ (RB usage), while the queue evolves via arrivals and service dynamics. This statefulness is essential, since insufficient service can accumulate backlog and trigger persistent latency violations.

d) Online optimization and learning with cross-level coupling: The provisioned budget constrains scheduling through the episode-wise budget constraint (2), requiring the expected cumulative RB usage within episode to not exceed budget. Hence, the upper level must adapt b_k online under switching costs, while the performance of a given b_k is realized through a lower-level CMDP scheduler under unknown dynamics.

Algorithm 1 Bi-Level Optimization and Learning (BLOL)

Initialization: $b_1 \leftarrow B_0$; history $\mathcal{H} \leftarrow \emptyset$.
for $k = 1$ to K **do**
  Execute BALDE (Algorithm 2) in episode k under budget b_k and obtain $(\pi_k, \lambda_k, \mathcal{H})$.
  if $k \leq K_0$ **then**
    $b_{k+1} \leftarrow B_0$
  else
    Compute approximate subgradient according to (11) and update resource budget according to (12).
  end if
end for

IV. Bi-Level Algorithm Design

We design an online algorithm for the problem in Sec. II.

A. Upper Level OCO via Lower Level Dual Feedback

The upper-level selects a budget sequence $\{b_k\}_{k=1}^K$ to provision episode-wise resources while accounting for the unknown f_k and switching overhead. Moreover, the effective loss of b_k depends on the *best achievable lower-level performance* under the provisioned budget. This relationship is implicit because the operator lacks an analytical model of the packet arrival and scheduling dynamics, i.e., unknown MDP dynamics. We address these challenges as follows and see Algorithm 1.

1) Reduce bi-level coupling to an "oracle" effective loss: To isolate the upper-level design, we first analyze an *idealized* setting in which the lower-level scheduling policy can be optimized for any budget b_k. Define the corresponding cost

$$g_k(b_k) \triangleq f_k(b_k) + \beta L_k^*(b_k), \tag{8}$$

where $f_k(b_k)$ is the upper-level service cost, and $L_k^*(b_k)$ is the optimal lower-level scheduling loss under budget b_k, i.e.,

$$L_k^*(b_k) \triangleq \min_{\pi^{\mathrm{A}} \in \Pi(b_k)} \mathbb{E}_{\pi^{\mathrm{A}}, P}\left[\sum\nolimits_{t=1}^{T} l_t(s_t, a_t)\right]. \tag{9}$$

Under this *fictitious* oracle view, the upper level would be reduced to an ideal OCO problem over $\mathcal{B}$, i.e.,

$$\min_{b_k \in \mathcal{B}} \Big\{ \sum\nolimits_{k=1}^{K} g_k(b_k) + \sum\nolimits_{k=1}^{K} \alpha \|b_k - b_{k-1}\|^2 \Big\}. \tag{10}$$

2) Dual-based subgradient approximation: Directly optimizing (10) is still challenging because $L_k^*(b_k)$ depends on the unknown lower-level MDP dynamics. Our key observation is that, by strong duality of the linear programing (LP) formulation of finite-horizon CMDPs, the optimal dual multiplier $\lambda_k^*(b_k)$ of the budget constraint b_k provides the sensitivity of the optimal value with respect to the budget. Physically, $\lambda_k^*(b_k)$ signal the upper level how much QoS could improve if more RBs were allocated. Thus, $-\lambda_k^*(b_k)$ constitutes a valid subgradient of loss $L_k^*(b_k)$ (Please see Appendix B for the complete proof of Lemma 12). Specifically, for any $b_k \in \mathcal{B}$, the subgradient of the true loss is given by

$$\partial g_k(b_k) = \nabla f_k(b_k) + \beta \partial L_k^*(b_k) = \nabla f_k(b_k) - \beta \lambda_k^*(b_k),$$

Algorithm 2 Budget-Adaptive Learning with Dual Estimation

Input: episode k, budget b_k, history $\mathcal{H}$, confidence level δ, baseline policy π_{base}.
Compute $\hat{P}_{k,t}$ and β^p using $\mathcal{H}$ according to (13-15).
Calculate estimates $\bar{l}_{k,t}$ and $\bar{d}_{k,t}$ according to (16-17).
if $k \leq K_0$ **then**
  $\pi_k^{\mathrm{A}} \leftarrow \pi_{\mathrm{base}}$ ▷ Use safe baseline during warm-up
  $\lambda_k \leftarrow 0$ ▷ Dual variable is not used in warm-up
else
  Solve the extended LP (23) with budget b_k, and obtain optimal policy π_k^{A} and dual multiplier λ_k of the constraint.
end if
for $t = 1$ to T **do**
  Observe state $s_{k,t}$, next execute action $a_{k,t} \sim \pi_{k,t}(\cdot|s_{k,t})$, and then observe loss $l_t(s_{k,t}, a_{k,t})$, consumption $d_t(s_{k,t}, a_{k,t})$, and next state $s_{k,t+1}$.
  Update history: $\mathcal{H} \leftarrow \mathcal{H} \cup \{(s_{k,t}, a_{k,t}, s_{k,t+1})\}$.
end for
Output: $(\pi_k, \lambda_k, \mathcal{H})$.

However, since the true MDP dynamics are unknown, the true multiplier $\lambda_k^*(b_k)$ is inaccessible. In our algorithm, we approximates it with the empirical dual estimate λ_k returned by the lower-level scheduler. Then, the update direction is

$$\hat{h}_k \triangleq \nabla f_k(b_k) - \beta \lambda_k. \tag{11}$$

While $\hat{h}_k$ is an approximation of the $\partial L_k^*(b_k)$, it acts as an exact subgradient for the *surrogate objective* $\hat{g}_k(b_k) = f_k(b_k) + \beta \hat{L}_k(b_k)$, where $\hat{L}_k(b_k)$ is the surrogate lower-level optimal loss. This connection allows us to analyze the algorithm as performing projected subgradient descent on the surrogate functions $\{\hat{g}_k\}$, with the gap between true and surrogate objectives controlled by the lower-level learning guarantees, and it will appear in the upper-level regret bound.

3) Switching cost control via stable OCO updates: While the total episode cost includes the switching term $\alpha\|b_k - b_{k-1}\|^2$, our update rule operates on (approximate) subgradients of g_k. As we show in Lemma 2, a decaying step size $\eta_k \propto 1/k$ yields stable budget trajectories with controlled path length, which in turn controls the cumulative switching costs without requiring a switching regularizer in the update.

4) Projected subgradient descent with a warm-up phase: We adopt a warm-up period of $K_0 = \tilde{\mathcal{O}}(S^2AT^4/\gamma^2)$ episodes to allow the lower-level learner to collect sufficient samples for dual estimates, and see the proof sketch of Lemma 1.

- *Warm-up ($k \leq K_0$):* set $b_k = B_0$ and run the lower-level learner to explore the environment.
- *Online update ($k > K_0$):* after observing $f_k(\cdot)$ and receiving λ_k from the lower level, update the budget

$$b_{k+1} = \Pi_{\mathcal{B}}\left(b_k - \eta_k \hat{h}_k\right), \tag{12}$$

where the step-size is $\eta_k = 1/[\theta_g(k - K_0)]$.

B. Lower Level RL Under Upper Level Dynamic Budget

In episode k, given the budget b_k, the lower-level scheduler seeks a policy π_k (for brevity, we write π_k for π_k^{A} in this subsection) that minimizes the expected cumulative loss $V_l(\pi_k; P)$ while satisfying the episode budget constraint $V_d(\pi_k; P) \le b_k$.

To guarantee safety *uniformly* for dynamic constraint threshold and return a dual variable λ_k, we propose *Budget-Adaptive Learning with Dual Estimation* (BALDE), see Algorithm 2.

1) Empirical estimates and confidence sets: We focus on uncertainty in the packet arrival and channel dynamics (transition probabilities), and the results can be extended to unknown l_t and d_t via standard concentration arguments. For each state-action pair (s, a) at time sot t, define the visitation counts prior to episode k as $n_{k,t}(s,a) = \sum_{k'=1}^{k-1} \mathbf{1}\{s_{k',t} = s,\ a_{k',t} = a\}$, $n_{k,t}(s,a,s') = \sum_{k'=1}^{k-1} \mathbf{1}\{s_{k',t} = s,\ a_{k',t} = a,\ s_{k',t+1} = s'\}$. At the start of episode k, the empirical transition estimate is

$$\hat{P}_{k,t}(s'|s,a) = n_{k,t}(s,a,s')/[n_{k,t}(s,a) \vee 1], \tag{13}$$

Then, we construct a confidence set $\mathcal{P}_k$ centered at $\hat{P}_{k,t}$:

$$\mathcal{P}_k \triangleq \bigcap\nolimits_{t\in[T],\, s\in\mathbb{S},\, a\in\mathbb{A}} \mathcal{P}_{k,t}(s,a), \tag{14}$$

where, based on the empirical Bernstein inequality [14], $\mathcal{P}_{k,t}(s,a) = \{P' : \left|P'_t(s'|s,a) - \hat{P}_{k,t}(s'|s,a)\right| \le \beta^p_{k,t}(s,a,s')$ for all $s'\}$, with radius

$$\beta^p_{k,t}(s,a,s') = \sqrt{\frac{4\,\mathrm{Var}(\hat{P}_{k,t}(s'|s,a))\,L'}{n_{k,t}(s,a)\vee 1}} + \frac{14L'/3}{n_{k,t}(s,a)\vee 1}, \tag{15}$$

and $L' = \log\left(\frac{2SATK}{\delta}\right)$. With probability at least $1-2\delta$, the true model P lies in $\mathcal{P}_k$ for all k (Please see Appendix A for the complete proof of Lemma 1).

2) Pessimism for safety: To ensure the scheduled resources do not exceed the budget b_k under model uncertainty, we inflate the constraint cost using a pessimistic penalty, i.e.,

$$\bar{d}_{k,t}(s,a) = d_t(s,a) + T\beta^p_{k,t}(s,a), \tag{16}$$

where $\beta^p_{k,t}(s,a) \triangleq \sum_{s'} \beta^p_{k,t}(s,a,s')$. $\bar{d}_{k,t}$ enforces a conservative constraint that holds uniformly for all $P' \in \mathcal{P}_k$.

3) Budget-aware optimism for efficient exploration: Pure pessimism can severely restrict exploration, especially for rarely visited state-action pairs. We therefore introduce an optimistic adjustment to the scheduling loss, i.e.,

$$\bar{l}_{k,t}(s,a) = l_t(s,a) - [T^2\beta^p_{k,t}(s,a)]/(b_k - b_{\mathrm{base}}). \tag{17}$$

The scaling by $(b_k - b_{\mathrm{base}})^{-1}$ is crucial, since when the current budget b_k is tight (close to b_{base}), the algorithm reduces aggressive exploration, and when b_k is larger, the learner can explore more aggressively while still maintaining feasibility.

4) Optimization under extended LP: Based on the construction above, we select the scheduling policy as follows,

$$\pi_k = \arg\min\nolimits_{\{\pi': V_{\bar{d}_k}(\pi';P')\le b_k;\, P'\in\mathcal{P}_k\}} V_{\bar{l}_{k,t}}(\pi'; P'). \tag{18}$$

This formulation is optimistic in the objective (to learn low-latency scheduling) and pessimistic in the constraint (to ensure RB limits are respected). However, solving (18) efficiently and *extracting the dual multiplier* associated with the budget constraint are essential and non-trivial in our bi-level design.

To this end, we cast (18) as an *extended LP* in occupancy measures, which serves two roles: (i) it provides a tractable solver for the lower-level under time-inhomogeneous dynamics, and (ii) it yields the *dual multiplier* estimate λ_k for the budget constraint, which acts as a sensitivity signal (approximately $\partial_b L^*_k(b_k)$) required by the upper-level update. This primal-dual interface is critical not only for algorithm implementation but also for regret analysis, where λ_k connects lower-level learning error to upper-level optimization error. Specifically, for a policy π and transition kernels $P = \{P_t\}_{t=1}^T$, define the state-action occupancy measure $w_t^\pi(s,a;P) \triangleq \Pr(s_t = s, a_t = a \mid \pi, P)$ [15]. The expected cumulative loss and resource consumption are linear, i.e.,

$$V_r^\pi(P) = \sum\nolimits_{t=1}^T \sum\nolimits_{s,a} w_t^\pi(s,a;P) r_t(s,a) = \langle r, w^\pi\rangle, \tag{19}$$

where $r \in \{l, d\}$ and $w^\pi = \{w_t^\pi\}_{t=1}^T$. Any feasible occupancy measure must satisfy standard flow constraints. Let $\mathcal{W}(P)$ denote the set of nonnegative w^π satisfying, for all $s \in \mathcal{S}$,

$$\sum\nolimits_a w_t^\pi(s,a) = \sum\nolimits_{s',a'} P_{t-1}(s \mid s',a') w_{t-1}^\pi(s',a'). \tag{20}$$

Then, the CMDP under dynamic budget b_k is equivalent to

$$\min\nolimits_{w^\pi\in\mathcal{W}(P)} \langle l, w^\pi\rangle,\ \text{s.t. } \langle d, w^\pi\rangle \le b_k. \tag{21}$$

An optimal policy can be obtained by normalization $\pi_t^*(a \mid s) = w_t^*(s,a)/\sum_{a'} w_t^*(s,a')$ for nonzero denominators. In BALDE, the transition model is unknown and is optimized within a confidence set $\mathcal{P}_k$. To avoid bilinearity between $P' \in \mathcal{P}_k$ and w^π, we introduce an augmented occupancy measure

$$q_t^\pi(s,a,s') \triangleq w_t^\pi(s,a)\, P'_t(s' \mid s,a), \tag{22}$$

which represents the joint occupancy of (s_t, a_t, s_{t+1}). This yields a linear program in q^π to represent the problem (18) :

$$\min\nolimits_{q^\pi} \sum\nolimits_{t,s,a,s'} q_t^\pi(s,a,s')\bar{l}_t(s,a) \tag{23a}$$

$$\text{s.t.} \sum\nolimits_{t,s,a,s'} q_t^\pi(s,a,s')\bar{d}_t(s,a) \le b_k, \tag{23b}$$

$$\sum\nolimits_{a,s'} q_t^\pi(s,a,s') = \sum\nolimits_{s'',a''} q_{t-1}^\pi(s'',a'',s), \tag{23c}$$

$$\hat{P}_{k,t}(s' \mid s,a) - \beta^p_{k,t}(s,a,s') \le \frac{q_t(s,a,s')}{\sum_y q_t(s,a,y)}$$

$$\le \hat{P}_{k,t}(s' \mid s,a) + \beta^p_{k,t}(s,a,s'), \forall s,a,s',t, \tag{23d}$$

$$q_t(s,a,s') \ge 0, \forall s,a,s',t, \tag{23e}$$

where (23c) holds for all state s and (23d) encodes the transition-dynamics confidence set constraint $P'_t(\cdot \mid s,a) \in \mathcal{P}_k$ in a linear form. Crucially, the dual variable associated with the budget constraint (23b) is exactly the multiplier λ_k, which the lower level returns to the upper level as a subgradient and sensitivity feedback for updating the budget b_k.

5) Initial safety under dynamic budgets: As is common in safe RL [8], [9], we assume access to a baseline safe scheduler π_{base} (e.g., a conservative rate-limiting policy) satisfying $V_d(\pi_{\text{base}}; P) = b_{\text{base}}$ with $b_{\text{base}} < B_0$, which is used to guarantee feasibility in early learning episodes without any pre-samples. We only execute π_{base} during the warm-up phase.

V. Regret Analysis and Budget Guarantee

In this section, we establish the regret upper bound and cross-level budget constraint satisfaction guarantee for our algorithm. The main technical challenges lie in quantifying the error quantification from the lower-level reinforcement learning under varying threshold to the upper-level optimization updates under coupling and estimated sensitivity feedback. We provide proof sketches for the lemmas and theorems (Please see complete proofs in our Appendix).

A. Regret Decomposition

Recall the total regret after K episodes $\text{Reg}(K, b, \pi) \triangleq \sum_{k=1}^{K} C_k(b_k, \pi_k^{\text{A}}) - \sum_{k=1}^{K} C_k(b^*, \pi^{\text{A},*})$. For each episode k and a fixed upper-level decision b_k, define the *episode-wise optimal safe lower-level policy* under the true model P:

$$\pi_k^*(b_k) \in \arg\min_{\pi \in \Pi(b_k)} V_l(\pi; P), \tag{24}$$

where $\Pi(b_k)$ is the feasible policy set induced by budget b_k. Using $\pi_k^*(b_k)$ as an intermediate comparator, we can decompose the final regret as follows,

$$\begin{aligned}\text{Reg}(K, b, \pi) = &\underbrace{\sum_{k=1}^{K} \Big(C_k(b_k, \pi_k^{\text{A}}) - C_k(b_k, \pi_k^*(b_k)) \Big)}_{\text{(I) Lower-level BALDE regret under budget } b_k\text{: } \text{Reg}_{\text{LL}}(K,\pi)} \\ + &\underbrace{\sum_{k=1}^{K} \Big(C_k(b_k, \pi_k^*(b_k)) - C_k(b^*, \pi^{\text{A},*}) \Big)}_{\text{(II) Upper-level BLOL regret with switching costs: } \text{Reg}_{\text{UL}}(K,b)} . \end{aligned} \tag{25}$$

Term (I) in (25) represents the lower-level BALDE regret under varying budget b_k, i.e., the scheduling regret. By the definition of C_k in (4), for a fixed budget b_k, the upper-level service cost and switching cost diminishes to zero and we have

$$\text{Reg}_{\text{LL}}(K, \pi) = \beta \sum_{k=1}^{K} (V_l(\pi_k; P) - V_l(\pi_k^*(b_k); P)). \tag{26}$$

Thus, $\text{Reg}_{\text{LL}}(K, \pi)$ is exactly the scaled learning regret of the lower-level algorithm BALDE under the budget sequence generated by the upper level. Term (II) in (25) represents the upper-level BLOL regret with switching costs, i.e., the provisioning regret. Using the oracle effective loss $g_k(b) \triangleq f_k(b) + \beta L_k^*(b)$, where $L_k^*(b) = \min_{\pi \in \Pi(b)} V_l(\pi; P)$ (see (9)), the upper-level regret term $\text{Reg}_{\text{UL}}(K, b)$ becomes

$$\sum_{k=1}^{K} (g_k(b_k) - g_k(b^*)) + \alpha \sum_{k=1}^{K} \|b_k - b_{k-1}\|^2. \tag{27}$$

B. Lower-Level Regret Bound

We first bound the lower-level CMDP scheduling regret.

Lemma 1 (Lower-level BALDE regret)**.** *Fix $\delta \in (0, 1)$. With probability at least $1 - 3\delta$, executing BALDE (Algorithm 2) under budget sequence $\{b_k\}_{k=1}^K$, we have ($\gamma = B_0 - b_{\text{base}}$)*

$$\sum_{k=1}^{K} \Big(V_l(\pi_k; P) - V_l(\pi_k^*(b_k); P) \Big) \le \tilde{\mathcal{O}} \left(S T^3 \sqrt{AK}/\gamma \right).$$

Lemma 1 provides a *high-probability* regret guarantee for the lower-level scheduler under an *episode-varying* budget sequence $\{b_k\}_{k=1}^K$, which is crucial for coupling with the upper-level provisioning decisions. The bound scales as $\tilde{\mathcal{O}}(\sqrt{K})$, i.e., the average per-episode suboptimality decays as $\tilde{\mathcal{O}}(1/\sqrt{K})$. This $\sqrt{K}$ dependence is optimal and matches the canonical statistical rate for learning under unknown Markov dynamics. The dependence on γ makes explicit a fundamental *safety-learning tradeoff*. γ is the minimum safety slack that allows the algorithm to enforce a pessimistically inflated constraint (to guarantee feasibility under uncertainty) while still retaining room for exploration. When γ is small, safe learners must behave conservatively, so a blow-up with $1/\gamma$ is unavoidable [8], [9]. Finally, the T^3 factor is characteristic of finite-horizon optimistic exploration with extended-LP-based planning. It reflects that estimation errors compound along length-T trajectories and that additional factors appear when enforcing constraints pessimistically while maintaining optimism. Please see Appendix A for the complete proof of Lemma 1.

Proof sketch of Lemma 1. The proof follows an optimism-in-the-face-of-uncertainty argument together with pessimism for the constraint, but adapted to a time-varying budget b_k.

1) Step 1 (High-probability event): We construct confidence sets $\mathcal{P}_k$ around empirical transitions and define a good event $\mathcal{G}$ under which $P \in \mathcal{P}_k$ for all k. Empirical Bernstein bounds then imply $\Pr(\mathcal{G}) \ge 1 - 3\delta$.

2) Step 2 (Safety under a dynamic threshold): We use a safe baseline policy π_{base} during warm-up phase K_0 to ensure feasibility when uncertainty is high in early exploration. For $k > K_0$, the solver ensures safety by incorporating a cumulative transition bonus $\epsilon_k^{\pi_k}(P_k) \triangleq T\mathbb{E}_{\pi_k, P_k}[\sum_{t=1}^{T} \beta_{k,t}^p]$, we have $V_d(\pi_k; P) \le V_d(\pi_k; P_k) + \epsilon_k^{\pi_k}(P_k) = V_{\bar{d}}(\pi_k; P_k)$. Since the solver ensures $V_{\bar{d}}(\pi_k; P_k) \le b_k$, it follows that $V_d(\pi_k; P) \le b_k$ under event $\mathcal{G}$ with probability at least $1 - 3\delta$.

3) Step 3 (Regret bound): We decompose the per-episode regret into two terms, i.e., $V_l(\pi_k; P) - V_l(\pi^*; P) = (V_l(\pi_k; P) - V_{\bar{l}}(\pi_k; P_k)) + (V_{\bar{l}}(\pi_k; P_k) - V_l(\pi^*; P))$, where $V_{\bar{l}}$ is the value under a pessimistic bonus-augmented loss $\bar{l}_k = l_k - mT\beta_k^p$. We show that setting the scaling factor $m = T/(b_k - b_{\text{base}})$ is sufficient to guarantee $V_{\bar{l}}(\pi_k; P_k) \le V_l(\pi^*; P)$. In addition, the first error term $V_l(\pi_k; P) - V_{\bar{l}}(\pi_k; P_k)$ in the decomposition can be bounded by relating $V_{\bar{l}}$ back to V_l via the classic value difference lemma [7]–[9] and the cumulative bonus. Finally, summing over k and bounding the cumulative bonus terms $\sum \epsilon_k^{\pi_k}$ by invoking the pigeon-hole principle yields the final regret bound of $\tilde{\mathcal{O}}\left(T^3 S\sqrt{AK}/\gamma\right)$. □

C. Upper-Level Regret Bound

We now bound the upper-level OCO provisioning regret.

Lemma 2 (Upper-level BLOL Regret with Switching Costs)**.** *Surrogate effective cost $\hat{g}_k(\cdot)$ is θ_g-strongly convex on $\mathcal{B}$ and has bounded subgradients $\|\partial \hat{g}_k(b)\| \le G$ for all $b \in \mathcal{B}$ and all k. Let $\{b_k\}$ be updated by projected subgradient descent using the approximate subgradient $\hat{h}_k$ and step-size $\eta_k = \frac{1}{\theta_g(k-K_0)}$ for $k > K_0$. Then, with probability at least $1-3\delta$, we have*

$$\mathit{Reg}_{\mathrm{UL}}(K,b) \le \underbrace{G^2 \log K/(2\theta_g) + M}_{\textit{Surrogate regret and switching costs}} + \underbrace{\tilde{\mathcal{O}}\left(ST^3\sqrt{AK}/\gamma\right)}_{\textit{Value approximation error}},$$

where $M = K_0 G(T-B_0) + G^2\pi^2/(6\theta_g^2)$ collects constant overheads from the warm-up phase and the finite switching cost series, and π is the mathematical constant.

Lemma 2 isolates two phenomena. First, the term for surrogate regret and switching costs shows that, despite switching costs, the upper-level optimization error scales only as $\tilde{\mathcal{O}}(\log K)$ under strong convexity. The key reason switching costs do not destroy the rate is stability, i.e., with $\eta_k \propto 1/k$, budget updates shrink over time, so the path length $\sum_k \|b_k - b_{k-1}\|^2$ is bounded by a convergent series. Second, the term for value approximation error captures the penalty for optimizing the surrogate $\hat{g}_k$ instead of the true g_k. This term scales as $\tilde{\mathcal{O}}(\sqrt{K})$ and matches the lower-level learning regret, which confirms that the overall performance is dominated by the difficulty of learning dynamics and the resulting sensitivity. Please see Appendix B for the complete proof of Lemma 2.

Proof Sketch of Lemma 2. The core idea is to decompose the regret against the true objective g_k into the regret on the surrogate $\hat{g}_k$ and the approximation error. Specifically, we have

$$\begin{aligned}\mathrm{Reg}_{\mathrm{UL}}(K,b) = &\underbrace{\sum\nolimits_k (\hat{g}_k(b_k) - \hat{g}_k(b^*)) + \text{Switching costs}}_{\text{Surrogate regret and switching costs}} \\ &+ \underbrace{\sum\nolimits_k (g_k(b_k) - \hat{g}_k(b_k)) - (g_k(b^*) - \hat{g}_k(b^*))}_{\text{Value approximation error}}.\end{aligned}$$

1) Surrogate regret and switching costs: Since $\hat{h}_k$ is a valid subgradient of the strongly convex surrogate $\hat{g}_k$ (as discussed in Sec. IV-A), the first term can bounded as follows. With step size $\eta_k \propto 1/k$, the surrogate regret scales as $\log K$. The switching cost at step k is bounded by $\|b_k - b_{k-1}\|^2 \le (G\eta_{k-1})^2 \propto 1/k^2$. Since the series $\sum k^{-2}$ converges, the cumulative switching cost is bounded by a constant.

2) Value approximation error: The remaining terms capture the value difference between the true and surrogate objectives. By definition, $g_k(b) - \hat{g}_k(b) = \beta(L_k^*(b) - \hat{L}_k(b))$. Using the optimism lemma from the lower-level analysis (Please see Appendix A for complete proof of Lemma 10), we can show that $\hat{L}_k(b)$ (the value of the optimistic surrogate) is close to $L_k^*(b)$ (the true optimal value). The cumulative sum of these differences/errors is bounded by the lower-level regret bound, which thus yields the $\tilde{\mathcal{O}}(\sqrt{K})$ term in the regret. □

D. Final Regret Bound and Safety Guarantee

Combining (25)-(27), Lemma 1, and Lemma 2, we obtain the final performance and safety guarantee.

Theorem 1 (Total regret of our bi-level algorithm BLOL)**.** *With probability at least $1-3\delta$, the regret of the proposed bi-level algorithm BLOL (Algorithm 1) satisfies*

$$\mathit{Reg}(K, BLOL) \le \tilde{\mathcal{O}}\left(ST^3\sqrt{AK}/\gamma\right) + \tilde{\mathcal{O}}(\log K). \quad (28)$$

Theorem 1 establishes a near-optimal $\tilde{\mathcal{O}}(\sqrt{K})$ regret, implying the average regret vanishes. Importantly, the bound separates the roles of the two levels: the upper level contributes only logarithmic regret (and bounded switching overhead) thanks to strong convexity and stable updates, while the dominant $\tilde{\mathcal{O}}(\sqrt{K})$ term comes from learning the unknown lower-level dynamics safely even under time-varying budgets. This clean separation clarifies the benefit of our primal-dual interface. The extended LP provides a dual variable signal that converts lower-level learning progress into correct sensitivity feedback for upper-level subgradient in budget updating. Please see Appendix C for the complete proof of Theorem 1.

Theorem 2 (Safety Guarantee)**.** *With probability at least $1-3\delta$, the algorithm BALDE (Algorithm 2) satisfies the budget constraint for all episodes $k \in [K]$, i.e., $V_d(\textit{BALDE}; P) \le b_k$.*

Theorem 2 establishes the safety of our BALDE subroutine under dynamic/time-varying budget and constraint threshold. It guarantees that the lower-level scheduler strictly adheres to the resource limits imposed by the upper level. Please see Appendix D for the complete proof of Theorem 2.

Proof sketch of Theorem 2. For the warm-up phase ($k \le K_0$), safety is guaranteed by baseline policy π_{base}. For $k > K_0$, applying the value difference lemma and Hölder's inequality, the difference between the cost under true transition kernel P and that under P_k can be bound as $|V_d(\pi_k; P) - V_d(\pi_k; P_k)| \le \mathbb{E}[\sum_{t=1}^T \|P_t - P_{k,t}\|_1 \|V_{d,t+1}^{\pi_k}\|_\infty] \le T\mathbb{E}_{\pi_k, P_k}[\sum_{t=1}^T \bar{\beta}_{t,k}^p] = \epsilon_k^{\pi_k}(P_k)$, where $V_{d,t}^{\pi} \triangleq \mathbb{E}_{\pi,P}[\sum_{\tau=t}^T d_t(s_t, a_t)]$. Thus, we can always get $V_d(\pi_k; P) \le V_d(\pi_k; P_k) + \epsilon_k^{\pi_k}(P_k) = V_{\bar{d}}(\pi_k; P_k) \le b_k$. Hence, the budget constraint is satisfied for all episodes with high probability. □

VI. Numerical Results

We evaluate the proposed bi-level algorithm (Algorithm 1) on a queueing-based slice scheduling model under both Poisson arrivals (with mean $\lambda = 1.12$ packets/slot) and real traffic traces MAWI [16] (aggregated into 10ms bins and linearly scaled to mean 1.12 while preserving burstiness). The goal is to jointly provision an episode-level bandwidth budget with reconfiguration overhead and schedule packet service within each episode under an episode-wise budget constraint.

A. Experimental Configurations

1) Lower level (queue scheduling learning): The state $s_t \in \{0, 1, \ldots, S_{\max}\}$ is the queue backlog with $S_{\max} = 10$. The action $a_t \in \{0, 1, 2\}$ is the number of RBs allocated in slot t. The queue evolves as $s_{t+1} = \min\{S_{\max}, [s_t - a_t + A_t]^+\}$, where $[x]^+ = \max\{x, 0\}$ and A_t is the packet arrival. We use the latency/backlog stage loss $l(s_t, a_t) = \mu + (1-\mu)(s_t/S_{\max})^2$ and the per-slot resource consumption $d(s_t, a_t) = a_t/2$, where $\mu = 0.1$. We let the horizon $T = 10$.

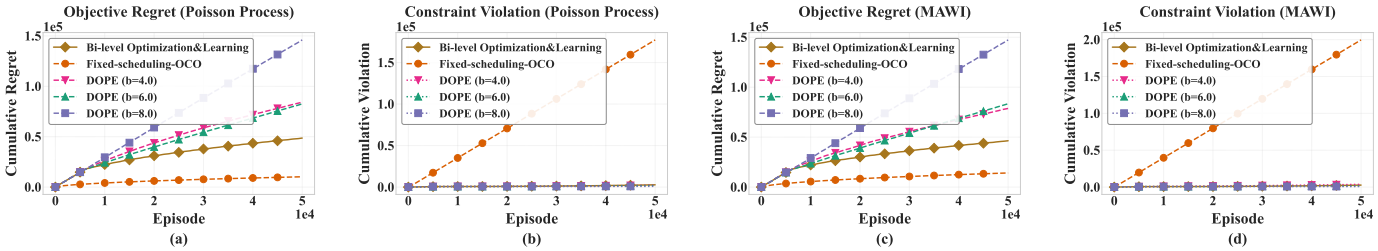

Fig. 3. Cumulative objective gap and cumulative budget violation under (a,b) Poisson arrivals and (c,d) MAWI traces. Here, $\rho_k = 5.0+0.5\sin(2\pi k/2000)+\epsilon_k$, $\epsilon_k \sim \mathcal{N}(0, 0.1^2)$, $M_0 = 0.25$, $M_1 = 0.01$, $M_2 = 0.05$, $\alpha = 0.5$, $\beta = 1.0$.

2) Upper level (budget provisioning optimization): At the beginning of episode k, the budget $b_k \in [B_0, T]$ is provisioned. The cost $f_k(b) = M_1 b + M_2(b - \rho_k)^2 + M_0(b - \rho_0)^2$ accounts for linear resource expenditures (M_1) and Service Level Agreement tracking (M_2, M_0), where sinusoidal ρ_k emulates diurnal seasonality for non-stationary evaluation. Across episodes we charge the switching cost $\alpha\|b_k - b_{k-1}\|^2$.

B. Evaluation Metrics and Baselines

We report the cumulative *objective gap* relative to a optimal static solution and cumulative budget violation $\mathrm{Viol}(K) = \sum_{k=1}^{K}[V_d(\pi_k; P) - b_k]_+$ for all methods. We compare our algorithm with two baselines, fixed-budget DOPE [17] with $b \in \{4, 6, 8\}$ (directly using DOPE with given fixed budget b and without adaptive upper-level provisioning) and fixed-scheduling OCO (directly optimizing $f_k(\cdot) + \alpha\|b_k - b_{k-1}\|^2$ via OCO, while the lower level runs standard RL with Q-learning [7] and without dual feedback coupling.

C. Results

We run $K = 5 \times 10^4$ episodes. Fig. 3 reports the cumulative objective gap (a,c) and cumulative budget violation (b,d) under Poisson arrivals and MAWI traces. In both traffic settings, our bi-level optimization-learning method achieves the *smallest objective gap among all budget-feasible algorithms* while keeping the cumulative violation essentially zero, indicating that dual-coupled provisioning can reduce long-run cost without sacrificing safety. In contrast, fixed-budget DOPE remains safe but incurs a substantially larger objective gap because the provisioned budget is not adapted across episodes. Finally, the fixed-scheduling OCO baseline attains a smaller objective gap by directly optimizing provisioning costs, but it persistently violates the episode budget, demonstrating that ignoring the cross-level coupling and lower-level safety/sensitivity feedback can lead to systematically infeasible provisioning under stateful queue dynamics and switching overhead.

VII. Conclusion and Future Work

We studied a bi-level online provisioning and scheduling problem with switching costs and cross-level budget constraints. The proposed framework integrates slow-time-scale resource provisioning (modeled as OCO with switching penalties) with fast-time-scale, state-dependent scheduling under *varying* constraint thresholds (modeled as CMDP under unknown dynamics). Our primal-dual bi-level algorithm couples the two levels through a dual sensitivity signal (the budget multiplier) returned by the lower level, enabling stable budget updates while maintaining safety during exploration. We established near-optimal static regret and high-probability satisfaction of the episode-wise budget constraints.

Future work includes the following directions. First, while the $\tilde{\mathcal{O}}(\sqrt{K})$ dependence is order-optimal, tightening the dependence on other problem parameters (e.g., T, S, A) remains an important objective. Second, we will extend the analysis to dynamic benchmarks to better capture non-stationary traffic and network conditions. Third, motivated by network slicing deployments, we plan to study multi-slice and distributed settings where provisioning decisions are coupled across slices.

References

[1] S. Lin, M. Shi, A. Arora, R. Bassily, E. Bertino, C. Caramanis, K. Chowdhury, E. Ekici, A. Eryilmaz, S. Ioannidis *et al.*, "Leveraging synergies between ai and networking to build next generation edge networks," in *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2022, pp. 16–25.

[2] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1452–1463, 2006.

[3] L. B. Le, K. Jagannathan, and E. Modiano, "Delay analysis of maximum weight scheduling in wireless ad hoc networks," in *Annual Conference on Information Sciences and Systems*. IEEE, 2009, pp. 389–394.

[4] M. Shi, X. Lin, and S. Fahmy, "Competitive online convex optimization with switching costs and ramp constraints," *IEEE/ACM Transactions on Networking*, vol. 29, no. 2, pp. 876–889, 2021.

[5] Y. Li, G. Qu, and N. Li, "Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit," *IEEE Transactions on Automatic Control*, vol. 66, no. 10, 2020.

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

[7] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[8] M. Shi, Y. Liang, and N. Shroff, "A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints," in *International Conference on Machine Learning*. PMLR, 2023, pp. 31 243–31 268.

[9] A. Roknilamouki, A. Ghosh, M. Shi, F. Nourzad, E. Ekici, and N. B. Shroff, "Provably efficient rl for linear mdps under instantaneous safety constraints in non-convex feature spaces," *arXiv preprint arXiv:2502.18655*, 2025.

[10] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete event dynamic systems*, vol. 13, no. 4, pp. 341–379, 2003.

[11] E. Hazan *et al.*, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[12] A. Gupta, X. Lin, and R. Srikant, "Low-complexity distributed scheduling algorithms for wireless networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 6, pp. 1846–1859, 2009.

[13] B. Ji, C. Joo, and N. B. Shroff, "Delay-based back-pressure scheduling in multihop wireless networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1539–1552, 2012.
[14] A. Maurer and M. Pontil, "Empirical bernstein bounds and sample variance penalization," *arXiv preprint arXiv:0907.3740*, 2009.
[15] M. Shi, Y. Liang, and N. Shroff, "Near-optimal adversarial reinforcement learning with switching costs," *arXiv preprint arXiv:2302.04374*, 2023.
[16] "Mawi working group traffic archive," https://mawi.wide.ad.jp/mawi/.
[17] A. Bura, A. HasanzadeZonuzy, D. Kalathil, S. Shakkottai, and J.-F. Chamberland, "Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning," *Advances in neural information processing systems*, vol. 35, pp. 1047–1059, 2022.
[18] C. Dann, T. Lattimore, and E. Brunskill, "Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[19] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained mdps," *arXiv preprint arXiv:2003.02189*, 2020.

APPENDIX A
PROOF OF LEMMA 1

A. *High probability events*

In this subsection, we define a high-probability event $\mathcal{G}$, which ensures that our estimations of the transition dynamics and state-action visitations are highly accurate. Conditioned on $\mathcal{G}$, we show that the regret of the lower-level algorithm is bounded.

To construct variance-aware confidence intervals for the unknown transition dynamics P, we bound the estimation error of the empirical model using the empirical Bernstein inequality.

Lemma 3 (Empirical Bernstein Inequality [14])**.** *Let $X = (X_1, \dots, X_n)$ be an i.i.d. sequence of random variables with values in $[0,1]$, and let $\delta \in (0,1)$. Then, with probability at least $1-\delta$:*

$$\Big| \mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} X_i \Big| \le \sqrt{\frac{2V_n(X)\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)}.$$

where $V_n(X)$ denotes the empirical variance of X.

Building upon Lemma 3, we construct the confidence set $\mathcal{P}_k$ centered around the empirical transition model $\hat{P}_{k,t}$ as: $\mathcal{P}_k = \bigcap_{t\in[T],(s,a)\in\mathbb{S}\times\mathbb{A}} \mathcal{P}_{k,t}(s,a)$, where the component set is defined as:

$$\mathcal{P}_{k,t}(s,a) = \Big\{P' : |P'_t(s'|s,a) - \hat{P}_{k,t}(s'|s,a)| \le \beta^p_{k,t}(s,a,s')\Big\}$$

The confidence radius $\beta^p_{k,t}(s,a,s')$ is given by:

$$\beta^p_{k,t}(s,a,s') = \sqrt{\frac{4\,\mathrm{Var}(\hat{P}_{k,t}(s'|s,a))\,L'}{n_{k,t}(s,a)\vee 1}} + \frac{14L'}{3\,(n_{k,t}(s,a)\vee 1)}$$

where $L' = \log\Big(\frac{2SATK}{\delta}\Big)$, and the empirical variance is $\mathrm{Var}(\hat{P}_{k,t}(s'|s,a)) = \hat{P}_{k,t}(s'|s,a)\big(1-\hat{P}_{k,t}(s'|s,a)\big)$.

First, we define the event F^p where the true dynamics fall within the confidence intervals:

$$F^p = \big\{P \in \mathcal{P}_k,\ \forall k \in [K]\big\}.$$

By invoking Lemma 3 and applying a union bound, we have $\Pr(F^p) \ge 1-2\delta$.

Next, we address the concentration of state-action visitation counters. We define the event F^w regarding the visitation frequency of each (s,a) pair:

$$F^w = \left\{ n_{k,t}(s,a) \ge \frac{1}{2}\sum_{j<k} w_{j,t}(s,a) - T\log\Big(\frac{SAT}{\delta}\Big)\right\}.$$

where $\forall(t,s,a,k) \in [T]\times\mathbb{S}\times\mathbb{A}\times[K]$, $w_{j,t}(s,a)$ denotes the occupancy measure induced by the policy in episode j.

The following lemma establishes that F^w holds with high probability. Its proof directly follows from [18, Corollary E.4.].

Lemma 4. *Let F^w be defined as above. Then, with probability at least $1-\delta$, the event holds, i.e., $\Pr(F^w) \ge 1-\delta$.*

Finally, we define the *good event* as the intersection of the F^p and F^w:

$$\mathcal{G} = F^p \cap F^w.$$

Since $\Pr(F^p) \ge 1-2\delta$ and $\Pr(F^w) \ge 1-\delta$, a standard union bound over F^p and F^w yields the yields the result formally stated in the following lemma.

Lemma 5 (Good Event)**.** *Let $\mathcal{G}$ be defined as above. Then, with probability at least $1-3\delta$, the event $\mathcal{G} = F^p \cap F^w$ holds, i.e., $\Pr(\mathcal{G}) \ge 1-3\delta$.*

Consequently, all subsequent theoretical analysis and regret bounds for the BALDE algorithm are derived conditioned on this high-probability event $\mathcal{G}$.

B. *Auxiliary Lemmas*

In this subsection, we present several technical lemmas that serve as the building blocks for our regret analysis.

Lemma 6 (Lemma 36, [19])**.** *Conditioned on the event F^w, the visitation frequencies satisfy:*

$$\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\sqrt{n_{k,t}(s_{k,t},a_{k,t})\vee 1}}\right] \le \tilde{\mathcal{O}}\big(\sqrt{SAT^2K} + SAT\big).$$

Lemma 7 (Lemma 37, [19])**.** *Conditioned on the event F^w, the following bound also holds:*

$$\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{n_{k,t}(s_{k,t},a_{k,t})\vee 1}\right] \le \tilde{\mathcal{O}}(SAT^2).$$

Lemma 8 (Value Difference Lemma)**.** *Consider two MDPs $M = (\mathbb{S},\mathbb{A},l,P)$ and $M' = (\mathbb{S},\mathbb{A},l,P')$ sharing the same state-action space and loss function l, but with different transition dynamics P and P'. For any policy π, any state $s\in\mathbb{S}$, and any time step $t\in[T]$, let $V^\pi_{l,t}(s;P) \triangleq \mathbb{E}_{\pi,P}[\sum_{\tau=t}^{T} l_\tau(s_\tau,a_\tau) \mid s_t = s]$ denote the value function from step t. Then, the following equality holds:*

$$\begin{aligned} &V^\pi_{l,t}(s;P) - V^\pi_{l,t}(s;P') \\ &= \mathbb{E}_{\pi,P'}\left[\sum_{\tau=t}^{T}\big((P_\tau - P'_\tau)(\cdot|s_\tau,a_\tau)\big)^\top V^\pi_{l,\tau+1}(\cdot;P) \;\Big|\; s_t = s\right] \\ &= \mathbb{E}_{\pi,P}\left[\sum_{\tau=t}^{T}\big((P'_\tau - P_\tau)(\cdot|s_\tau,a_\tau)\big)^\top V^\pi_{l,\tau+1}(\cdot;P') \;\Big|\; s_t = s\right]. \end{aligned}$$

The proof of lemma 8 is below:

Proof. Fix a policy π, a state $s \in \mathcal{S}$, and a time step $t \in [T]$. To simplify notation, we define:

$$V_t(s;P) \triangleq V_{l,t}^{\pi}(s;P), \quad V_t(s;P') \triangleq V_{l,t}^{\pi}(s;P').$$

and the value difference

$$\Delta_t(s) \triangleq V_t(s;P) - V_t(s;P').$$

We use the finite-horizon convention $V_{T+1}(\cdot;P) = V_{T+1}(\cdot;P') \equiv 0$.

By definition, the value function satisfies the Bellman equation. For model P, expanding the expectation over next states yields:

$$\begin{aligned} V_t(s;P) &= \mathbb{E}_{\pi}\left[l_t(s,a) + \sum_{s' \in \mathbb{S}} P_t(s'|s,a) V_{t+1}(s';P) \,\middle|\, s_t = s \right] \\ &= \mathbb{E}_{\pi}\left[l_t(s,a) + \big(P_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P) \,\middle|\, s_t = s \right]. \end{aligned} \tag{29}$$

Similarly, for model P', the Bellman equation is:

$$V_t(s;P') = \mathbb{E}_{\pi}\left[l_t(s,a) + \big(P'_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P') | s_t = s \right]. \tag{30}$$

Here, $\mathbb{E}_{\pi}[\cdot \mid s_t = s]$ denotes expectation w.r.t. $a \sim \pi_t(\cdot \mid s)$ only; the next-state expectation under $P_t(\cdot \mid s,a)$ (resp. $P'_t(\cdot \mid s,a)$) is already taken inside the inner product $\big(P_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P)$ (resp. $\big(P'_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P')$).

Subtracting (30) from (29), and noting that the l_t terms cancel, we obtain:

$$\begin{aligned} \Delta_t(s) = \mathbb{E}_{\pi}\Big[&\big(P_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P) \\ &- \big(P'_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P') \,\Big|\, s_t = s \Big]. \end{aligned} \tag{31}$$

Next, we add and subtract the term $\big(P'_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P)$ inside the expectation:

$$\begin{aligned} \Delta_t(s) = \mathbb{E}_{\pi}\Big[&\underbrace{\Big(\big(P_t(\cdot|s,a)\big)^{\top} - \big(P'_t(\cdot|s,a)\big)^{\top}\Big) V_{t+1}(\cdot;P)}_{\text{(a)}} \\ &+ \underbrace{\big(P'_t(\cdot|s,a)\big)^{\top} \big(V_{t+1}(\cdot;P) - V_{t+1}(\cdot;P')\big)}_{\text{(b)}} | s_t = s \Big]. \end{aligned} \tag{32}$$

Let us analyze term (b) in (32). Note that the term $\big(P'_t(\cdot|s,a)\big)^{\top}(V_{t+1}(\cdot;P) - V_{t+1}(\cdot;P'))$ is exactly the expected value of the next-step difference $\Delta_{t+1}(s_{t+1})$, where the next state is sampled from P'. Specifically:

$$\text{(b)} = \mathbb{E}_{s' \sim P'_t(\cdot|s,a)}[\Delta_{t+1}(s')].$$

Substituting this back into (32), we can rewrite the difference as an expectation over the trajectory generated by P':

$$\begin{aligned} \Delta_t(s) = \mathbb{E}_{\pi,P'}\Big[&\big((P_t - P'_t)(\cdot|s_t,a_t)\big)^{\top} V_{t+1}(\cdot;P) \\ &+ \Delta_{t+1}(s_{t+1}) | s_t = s \Big]. \end{aligned} \tag{33}$$

Equation (33) establishes a recursive relation for $\Delta_t(s)$. Since $V_{T+1} \equiv 0$ implies $\Delta_{T+1} \equiv 0$, we can unroll this recurrence forward in time from t to T:

$$\Delta_t(s) = \mathbb{E}_{\pi,P'}\left[\sum_{\tau=t}^{T} \big((P_\tau - P'_\tau)(\cdot|s_\tau,a_\tau)\big)^{\top} V_{\tau+1}(\cdot;P) \,\middle|\, s_t = s \right]$$

This proves the first equality of the lemma. The second equality follows from a symmetric argument by adding and subtracting $\big(P_t(\cdot|s,a)\big)^{\top} V_{t+1}(\cdot;P')$ in (31). □

We define the cumulative transition bonus $\epsilon_k^{\pi}(P)$ for an episode k under policy π and model P with initial state s_1 as:

$$\epsilon_k^{\pi}(P) \triangleq T \sum_{t=1}^{T} \mathbb{E}_{\pi,P}\Big[\sum_{s' \in \mathbb{S}} \beta_{k,t}^{p}(s_{k,t}, a_{k,t}, s') \,\Big|\, s_1 \Big].$$

We now derive a high-probability upper bound for the cumulative term $\epsilon_k^{\pi}(P)$.

Lemma 9. *Consider the sequence of policies $\{\pi_k\}_{k=1}^{K}$ generated by BALDE. For any $K' \le K$, with probability at least $1 - 3\delta$, the following bound holds:*

$$\sum_{k=1}^{K'} \epsilon_k^{\pi_k}(P) \le \tilde{\mathcal{O}}\Big(S\sqrt{A\,T^4 K'}\Big).$$

Proof. Recall the definition of the confidence radius $\beta_{k,t}^{p}(s,a,s')$:

$$\beta_{k,t}^{p}(s,a,s') = \sqrt{\frac{4\,\mathrm{Var}(\hat{P}_{k,t}(s'|s,a))\,L'}{n_{k,t}(s,a) \vee 1}} + \frac{14L'}{3\,(n_{k,t}(s,a) \vee 1)}.$$

Using the property $\mathrm{Var}(X) \le \mathbb{E}[X]$ for Bernoulli variables, we have $\mathrm{Var}(\hat{P}_{k,t}(s'|s,a)) \le \hat{P}_{k,t}(s'|s,a)$. Summing $\beta_{k,t}^{p}$ over all next states s' yields:

$$\begin{aligned} \sum_{s'} \beta_{k,t}^{p}(s,a,s') &\le \sum_{s'} \sqrt{\frac{4L'\hat{P}_{k,t}(s'|s,a)}{n_{k,t}(s,a) \vee 1}} + \sum_{s'} \frac{14L'}{3\,(n_{k,t}(s,a) \vee 1)} \\ &= \sqrt{\frac{4L'}{n_{k,t}(s,a) \vee 1}} \sum_{s'} \sqrt{\hat{P}_{k,t}(s'|s,a)} + \frac{14SL'}{3\,(n_{k,t}(s,a) \vee 1)}. \end{aligned}$$

By the Cauchy-Schwarz inequality, $\sum_{s'} \sqrt{\hat{P}_{k,t}(s'|s,a)} \leq \sqrt{S \sum_{s'} \hat{P}_{k,t}(s'|s,a)} = \sqrt{S}$. Substituting this back into the definition of $\epsilon_k^{\pi_k}(P)$, we have:

$$\begin{aligned}\sum_{k=1}^{K'} \epsilon_k^{\pi_k}(P) &= T \sum_{k=1}^{K'} \sum_{t=1}^{T} \mathbb{E}_{\pi_k,P}\left[\sum_{s'} \beta_{k,t}^p(s_{k,t}, a_{k,t}, s')\right] \\ &\leq T\sqrt{S} \sum_{k=1}^{K'} \sum_{t=1}^{T} \mathbb{E}_{\pi_k,P}\left[\sqrt{\frac{4L'}{n_{k,t}(s_{k,t},a_{k,t}) \vee 1}}\right] \\ &\quad + \frac{14SL'T}{3} \sum_{k=1}^{K'} \sum_{t=1}^{T} \mathbb{E}_{\pi_k,P}\left[\frac{1}{n_{k,t}(s_{k,t},a_{k,t}) \vee 1}\right].\end{aligned}$$

Conditioned on the event F^w, we invoke Lemma 6 and Lemma 7 to bound the summation terms:

$$\begin{aligned}\sum_{k=1}^{K'} \epsilon_k^{\pi_k}(P) &\leq T\sqrt{S} \cdot \tilde{\mathcal{O}}\left(\sqrt{SAT^2K'} + SAT\right) + \tilde{\mathcal{O}}(S^2AT^3) \\ &= \tilde{\mathcal{O}}\left(S\sqrt{AT^4K'}\right).\end{aligned}$$

This concludes the proof. □

C. *Proof of the Feasibility of the Extended LP Problem*

Before analyzing the regret bound , we first establish that the lower-level optimization in BALDE is well-defined at every episode. In the online learning phase ($k > K_0$), BALDE solves the extended LP problem (18) at each episode. Because the pessimistic constraint costs are heavily inflated by uncertainty bonuses early in learning, the LP might be infeasible in early learning phase. Thus, we need to show that this optimization problem is feasible after the K_0 warm-up phase. The proposition below shows that, on the good event $\mathcal{G}$, the baseline policy π_{base} remains feasible under the pessimistic constraint budget after sufficiently many warm-up episodes. Consequently, the extended LP admits at least one feasible solution for every episode $k > K_0$.

Proposition 1 (Feasibility of the Extended LP problem)**.** *Under the BALDE algorithm, let B_0 denote the initial budget used in the warm-up phase, and let π_{base} be a safe baseline policy satisfying $V_d(\pi_{\text{base}}; P) = b_{\text{base}}$ with slack $\gamma \triangleq B_0 - b_{\text{base}} > 0$. Suppose $\pi_k = \pi_{\text{base}}$ for all $k \leq K_0$. Then, with probability at least $1 - 3\delta$, the pair (π_{base}, P) is a feasible solution to the extended LP problem (18) for all episodes $k > K_0$, where $K_0 = \tilde{\mathcal{O}}\left(\frac{S^2AT^4}{\gamma^2}\right)$.*

Proof. Conditioned on the event $\mathcal{G}$, the model constraint $P \in \mathcal{P}_k$ is satisfied. Recall that the pessimistic constraint cost is

$$\bar{d}_{k,t}(s,a) = d_t(s,a) + T\beta_{k,t}^p(s,a),$$

where $\beta_{k,t}^p(s,a) = \sum_{s' \in \mathcal{S}} \beta_{k,t}^p(s,a,s')$. By linearity of expectation:

$$\begin{aligned}V_{\bar{d}}(\pi; P) &= \mathbb{E}_{\pi,P}\left[\sum_{t=1}^{T}\left(d_t(s_t,a_t) + T\bar{\beta}_{k,t}^p(s_t,a_t)\right)\right] \\ &= V_d(\pi; P) + \epsilon_k^{\pi}(P). \qquad (34)\end{aligned}$$

To prove feasibility of the extended LP, it suffices to show that for any episode $k > K_0$, the baseline pair (π_{base}, P) satisfies the pessimistic budget constraint:

$$V_{\bar{d}}(\pi_{\text{base}}; P) \leq b_k.$$

Since B_0 is the lower bound of the budgets (i.e., $b_k \geq B_0$ for all k), a sufficient condition for feasibility is:

$$V_{\bar{d}}(\pi_{\text{base}}; P) \leq B_0.$$

Using equation (34) and the property $V_d(\pi_{\text{base}}; P) = b_{\text{base}}$, this condition is equivalent to:

$$b_{\text{base}} + \epsilon_k^{\pi_{\text{base}}}(P) \leq B_0 \iff \epsilon_k^{\pi_{\text{base}}}(P) \leq \gamma. \qquad (35)$$

We bound the number of episodes required to satisfy (35) by contradiction. Suppose the condition is violated for the first K' episodes, i.e., $\epsilon_k^{\pi_{\text{base}}}(P) > \gamma$ for all $k \leq K'$, while running $\pi_k = \pi_{\text{base}}$. Summing this inequality over k yields:

$$K'\gamma < \sum_{k=1}^{K'} \epsilon_k^{\pi_{\text{base}}}(P).$$

Applying Lemma 9 with the sequence of policies $\pi_k = \pi_{\text{base}}$, we have:

$$\sum_{k=1}^{K'} \epsilon_k^{\pi_{\text{base}}}(P) \leq \tilde{\mathcal{O}}\left(S\sqrt{AT^4K'}\right).$$

Combining these inequalities implies $K'\gamma < \tilde{\mathcal{O}}(S\sqrt{AT^4K'})$, which simplifies to:

$$\sqrt{K'} < \tilde{\mathcal{O}}\left(\frac{S\sqrt{AT^4}}{\gamma}\right) \implies K' < \tilde{\mathcal{O}}\left(\frac{S^2AT^4}{\gamma^2}\right).$$

This leads to a contradiction provided that $K' \geq \tilde{\mathcal{O}}(S^2AT^4/\gamma^2)$. Thus, we can select $K_0 = \tilde{\mathcal{O}}(S^2AT^4/\gamma^2)$ to ensure the condition is met.

Finally, since the algorithm executes a fixed policy π_{base} during the warm-up phase, the visitation counts $n_{k,t}$ are non-decreasing in k. Consequently, the confidence radius and the cumulative bonus $\epsilon_k^{\pi_{\text{base}}}(P)$ are monotonically non-increasing. Therefore, the condition $\epsilon_k^{\pi_{\text{base}}}(P) \leq \gamma$ holds for all $k \geq K_0$. This completes the proof. □

D. *Proof of the CMDP Regret Bounds*

We first define the regret metric for the lower-level BALDE algorithm. Let π_k^* denote the optimal policy for the true CMDP under the episode budget b_k, and let π_k be the policy executed by BALDE algorithm in episode k. The cumulative regret after K episodes is defined as: $R(K) \triangleq \sum_{k=1}^{K}\big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big)$.

The cumulative regret can be decomposed into two parts: the warm-up phase and the online learning phase.

$$\begin{aligned} R(K) &= \sum_{k=1}^{K} \big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big) \\ &= \sum_{k=1}^{K_0} \big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big) \\ &\quad + \sum_{k=K_0+1}^{K} \big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big). \end{aligned} \tag{36}$$

a) Warm-up Phase ($k \le K_0$): During this phase, the algorithm executes the baseline policy $\pi_k = \pi_{\text{base}}$. Since the cumulative cost is bounded by T (i.e., $|V_l(\pi_k; P) - V_l(\pi_k^*; P)| \le T$), we can bound the regret using the warm-up length K_0 derived in Proposition 1:

$$\sum_{k=1}^{K_0} \big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big) \le TK_0 \le \tilde{\mathcal{O}}\left(\frac{S^2AT^5}{\gamma^2}\right). \tag{37}$$

where $\gamma \triangleq B_0 - b_{\text{base}}$ is the safety slack.

b) Online Phase ($k > K_0$): For the subsequent episodes, we decompose the episode-wise regret:

$$\begin{aligned} &V_l(\pi_k; P) - V_l(\pi_k^*; P) \\ &= \underbrace{V_l(\pi_k; P) - V_{\tilde{l}}(\pi_k; P_k)}_{\Delta_k^{(1)}} + \underbrace{V_{\tilde{l}}(\pi_k; P_k) - V_l(\pi_k^*; P)}_{\Delta_k^{(2)}}. \end{aligned}$$

where $V_{\tilde{l}}(\pi_k; P_k)$ denotes the optimistic value function constructed with the bonus. Summing over $k > K_0$, we obtain:

$$\begin{aligned} &\sum_{k=K_0+1}^{K} \big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big) \\ &= \sum_{k=K_0+1}^{K} \Delta_k^{(1)} + \sum_{k=K_0+1}^{K} \Delta_k^{(2)}. \end{aligned} \tag{38}$$

We now verify the optimism of the algorithm, which ensures that the second term is non-positive.

Lemma 10. *Let (π_k, P_k) be the optimal solution to the DOP problem (18) at episode k. By definition:*

$$(\pi_k, P_k) \in \arg\min_{\pi,\, P' \in \mathcal{P}_k} \Big\{ V_{\tilde{l}}(\pi; P') \text{ s.t. } V_{\bar{d}}(\pi; P') \le b_k \Big\}.$$

This implies $V_{\tilde{l}}(\pi_k; P_k) \le V_l(\pi_k^*; P)$, and hence $\Delta_k^{(2)} = V_{\tilde{l}}(\pi_k; P_k) - V_l(\pi_k^*; P) \le 0$, provided that the bonus scaling parameter m is sufficiently large.

Proof. First, we define the bonus-augmented loss (used only in the proof) by

$$\begin{aligned} \tilde{l}_{k,t}(s,a) &= l_{k,t}(s,a) - mT\,\beta_{k,t}^p(s,a), \\ \beta_{k,t}^p(s,a) &= \sum_{s'} \beta_{k,t}^p(s,a,s'). \end{aligned}$$

The optimal solution $(\tilde{\pi}_k, \tilde{P}_k)$ can be formulated as:

$$(\tilde{\pi}_k, \tilde{P}_k) = \arg\min_{\pi',\, P' \in \mathcal{P}_k} V_{\tilde{r}_k}(\pi'; P') \quad \text{subject to } V_{\bar{b}_k}(\pi'; P') \le b_k. \tag{39}$$

We show that if $m \ge \frac{T}{b_k - b_{\text{base}}}$, then $V_{\tilde{l}}(\tilde{\pi}_k; \tilde{P}_k) \le V_l(\pi_k^*; P)$.

Let $w^\pi(P)$ denote the occupancy measure of policy π in model P. Specifically, let $w^{\pi_{\text{base}}}(\cdot; P)$ and $w^{\pi_k^*}(\cdot; P)$ denote the occupancy measures of the baseline policy π_{base} and the optimal policy π_k^*, respectively. For episode k, consider a scalar $\alpha_k \in [0,1]$ and construct the mixed occupancy measure:

$$\tilde{w}_t(s,a) \;=\; (1-\alpha_k)\, w_t^{\pi_k^*}(s,a;P) \;+\; \alpha_k\, w_t^{\pi_{\text{base}}}(s,a;P).$$

and let $\tilde{\pi}$ be the policy induced by $\tilde{w}$. Since the value function is linear in the occupancy measure, we have:

$$\begin{aligned} &V_{\bar{d}}(\tilde{\pi}; P) = (1-\alpha_k) V_{\bar{d}}(\pi_k^*; P) + \alpha_k V_{\bar{d}}(\pi_{\text{base}}; P) \\ &= (1-\alpha_k)\big(V_d(\pi_k^*; P) + \epsilon_k^{\pi_k^*}\big) + \alpha_k\big(V_d(\pi_{\text{base}}; P) + \epsilon_k^{\pi_{\text{base}}}\big) \\ &\le (1-\alpha_k)\Big(b_k + \epsilon_k^{\pi_k^*}\Big) + \alpha_k(b_{\text{base}} + \epsilon_k^{\pi_{\text{base}}}) \\ &= b_k + \epsilon_k^{\pi_k^*} - \alpha_k \epsilon_k^{\pi_k^*} + \alpha_k(\epsilon_k^{\pi_{\text{base}}} - (b_k - b_{\text{base}})). \end{aligned}$$

From Proposition 1, when $k \ge K_0$, we have $\epsilon_k^{\pi_{\text{base}}} \le B_0 - b_{\text{base}} = \gamma$. Since $b_k \ge B_0$, it follows that $\epsilon_k^{\pi_{\text{base}}} - (b_k - b_{\text{base}}) \le 0$. Therefore, if

$$\alpha_k \;\ge\; \frac{\epsilon_k^{\pi_k^*}}{b_k - b_{\text{base}} - \epsilon_k^{\pi_{\text{base}}} + \epsilon_k^{\pi_k^*}}.$$

ensures $V_{\bar{d}}(\tilde{\pi}; P) \le b_k$. Thus, $(\tilde{\pi}, P)$ is a feasible solution for the problem (39) at episode k.

Since $(\tilde{\pi}_k, \tilde{P}_k)$ is the optimal solution minimizing $V_{\tilde{l}}$, we have $V_{\tilde{l}}(\tilde{\pi}_k; \tilde{P}_k) \le V_{\tilde{l}}(\tilde{\pi}_k; P)$. Thus, it suffices to find m such that $V_{\tilde{l}}(\tilde{\pi}_k; P) \le V_l(\pi_k^*; P)$. By linearity of the occupancy measure:

$$\begin{aligned} V_{\tilde{l}}(\tilde{\pi}_k; P) &= (1-\alpha_k) V_{\tilde{l}}(\pi_k^*; P) + \alpha_k V_{\tilde{l}}(\pi_{\text{base}}; P) \\ &= (1-\alpha_k)\Big(V_l(\pi_k^*; P) - m\,\epsilon_k^{\pi_k^*}\Big) \\ &\quad + \alpha_k(V_l(\pi_{\text{base}}; P) - m\,\epsilon_k^{\pi_{\text{base}}}) \\ &\le V_l(\pi_k^*; P). \end{aligned} \tag{40}$$

This inequality holds if:

$$m(1-\alpha_k)\epsilon_k^{\pi_k^*} + m\alpha_k\epsilon_k^{\pi_{\text{base}}} \ge \alpha_k(V_l(\pi_{\text{base}}; P) - V_l(\pi_k^*; P)).$$

We just setting $\alpha_k = \frac{\epsilon_k^{\pi_k^*}}{b_k - b_{\text{base}} - \epsilon_k^{\pi_{\text{base}}} + \epsilon_k^{\pi_k^*}}$, solving for m:

$$\begin{aligned} m &\ge \frac{\alpha_k(V_l(\pi_{\text{base}}; P) - V_l(\pi_k^*; P))}{(1-\alpha_k)\epsilon_k^{\pi_k^*} + \alpha_k\epsilon_k^{\pi_{\text{base}}}} \\ &= \frac{V_l(\pi_{\text{base}}; P) - V_l(\pi_k^*; P)}{\frac{1-\alpha_k}{\alpha_k}\epsilon_k^{\pi_k^*} + \epsilon_k^{\pi_{\text{base}}}} \\ &= \frac{V_l(\pi_{\text{base}}; P) - V_l(\pi_k^*; P)}{b_k - b_{\text{base}}}. \end{aligned} \tag{41}$$

where the simplification uses the definition of α_k. Since $l \in [0,1]$, we have $V_l(\pi_{\text{base}}; P) \le T$ and $V_l(\pi_k^*; P) \ge 0$. Thus,

$m = \frac{T}{b_k - b_{\text{base}}}$ satisfies the condition. This concludes the proof of Lemma 10. □

Using Lemma 10, we can bound (38) as follows:

$$\sum_{k=K_0+1}^{K} \big(V_l(\pi_k; P) - V_l(\pi_k^*; P)\big) \le \sum_{k=K_0+1}^{K} \Delta_k^{(1)}. \tag{42}$$

Recall that $\Delta_k^{(1)} = V_l(\pi_k; P) - V_{\bar{l}}(\pi_k; P_k)$. Using Lemma 9, we have:

$$\begin{aligned}
&\sum_{k=K_0+1}^{K} \Delta_k^{(1)} \\
&= \sum_{k=K_0+1}^{K} \left(V_l(\pi_k; P) - V_l(\pi_k; P_k) + \frac{T}{b_k - b_{\text{base}}} \epsilon_k^{\pi_k}(P_k) \right) \\
&\le \sum_{k=K_0+1}^{K} \big(V_l(\pi_k; P) - V_l(\pi_k; P_k)\big) + \tilde{\mathcal{O}}\left(\frac{ST^3\sqrt{AK}}{\gamma} \right).
\end{aligned} \tag{43}$$

where $\gamma = B_0 - b_{\text{base}}$ is the safety slack.

Next, we bound the estimation error term $\sum_{k=K_0+1}^{K}(V_l(\pi_k; P) - V_l(\pi_k; P_k))$. Applying the Value Difference Lemma (Lemma 8) and Hölder's inequality:

$$\begin{aligned}
&\sum_{k=K_0+1}^{K} (V_l(\pi_k; P) - V_l(\pi_k; P_k)) \\
&= \sum_{k=K_0+1}^{K} \sum_{t=1}^{T} \mathbb{E}_{\pi_k, P}\Big[\big((P_t - P_{t,k})(\cdot|s_t, a_t)\big)^\top V_{l,t+1}^{\pi_k}(\cdot; P_k) \Big] \\
&\le \sum_{k=K_0+1}^{K} \sum_{t=1}^{T} \mathbb{E}_{\pi_k, P}\Big[\|P_t - P_{t,k}\|_1 \cdot \|V_{l,t+1}^{\pi_k}\|_\infty \Big] \\
&\le \sum_{k=K_0+1}^{K} \epsilon_k^{\pi_k}(P) \le \tilde{\mathcal{O}}\big(ST^2\sqrt{AK}\big).
\end{aligned} \tag{44}$$

Finally, combining Eq. (36), Eq. (37), Eq. (42), Eq. (43), and Eq. (44), we obtain the total regret bound:

$$\begin{aligned}
R(K) &= \sum_{k=1}^{K} \Big(V_l(\pi_k; P) - V_l(\pi_k^*; P) \Big) \\
&\le \tilde{\mathcal{O}}\left(\frac{S^2AT^5}{\gamma^2} \right) + \tilde{\mathcal{O}}\big(ST^2\sqrt{AK}\big) + \tilde{\mathcal{O}}\left(\frac{ST^3\sqrt{AK}}{\gamma} \right) \\
&\le \tilde{\mathcal{O}}\left(\frac{ST^3\sqrt{AK}}{\gamma} \right).
\end{aligned} \tag{45}$$

The bound (45) holds with probability at least $1 - 3\delta$. □

APPENDIX B
PROOF OF LEMMA 2

A. *Convexity and Subgradients of the Lower-Level Value Function* $L_k^*(b)$

Lemma 11. *$L_k^*(b)$ is convex for any $b > 0$.*

Proof. To prove convexity, we show that for any $b_1, b_2 > 0$ and any $\theta \in [0, 1]$,

$$L_k^*(\theta b_1 + (1-\theta)b_2) \le \theta L_k^*(b_1) + (1-\theta)L_k^*(b_2).$$

Recall the definitions of expected loss and consumption:

$$J(\pi) \triangleq \mathbb{E}_\pi\Big[\sum_{t=1}^{T} l_t(s_{k,t}, a_{k,t}) \Big],$$
$$D(\pi) \triangleq \mathbb{E}_\pi\Big[\sum_{t=1}^{T} d_t(s_{k,t}, a_{k,t}) \Big].$$

Let $\pi^{A,1}$ and $\pi^{A,2}$ be optimal policies under budgets b_1 and b_2, respectively. By definition, we have:

$$\begin{aligned}
L_k^*(b_1) &= J(\pi^{A,1}), \quad D(\pi^{A,1}) \le b_1, \\
L_k^*(b_2) &= J(\pi^{A,2}), \quad D(\pi^{A,2}) \le b_2.
\end{aligned}$$

Consider a mixed budget $\bar{b} = \theta b_1 + (1-\theta)b_2$. We construct a composite policy $\pi^{\bar{A}}$ as follows: before the episode starts, draw a Bernoulli random variable U with $\Pr(U = 1) = \theta$ and $\Pr(U = 0) = 1 - \theta$. If $U = 1$, execute $\pi^{A,1}$; otherwise, execute $\pi^{A,2}$. By the law of total expectation, the resource consumption of $\pi^{\bar{A}}$ is:

$$\begin{aligned}
D(\pi^{\bar{A}}) &= \theta D(\pi^{A,1}) + (1-\theta) D(\pi^{A,2}) \\
&\le \theta b_1 + (1-\theta) b_2 = \bar{b}.
\end{aligned}$$

Thus, $\pi^{\bar{A}}$ is feasible for budget $\bar{b}$. Since $L_k^*(\bar{b})$ is the infimum over all feasible policies, we have $L_k^*(\bar{b}) \le J(\pi^{\bar{A}})$. Similarly, the expected loss is:

$$\begin{aligned}
J(\pi^{\bar{A}}) &= \theta J(\pi^{A,1}) + (1-\theta) J(\pi^{A,2}) \\
&= \theta L_k^*(b_1) + (1-\theta) L_k^*(b_2).
\end{aligned}$$

Combining these yields $L_k^*(\bar{b}) \le \theta L_k^*(b_1) + (1-\theta)L_k^*(b_2)$, which completes the proof of convexity. □

Lemma 12. *(Subgradient of the value function $L_k^*(b)$) For each episode k, any optimal dual multiplier $\lambda^*(b)$ associated with the budget constraint satisfies $-\lambda^*(b) \in \partial L_k(b)$.*

Proof. For brevity, we omit the episode index k. The primal problem is defined as $L(b) = \inf_{\pi: D(\pi) \le b} J(\pi)$. We introduce a Lagrange multiplier $\lambda \ge 0$ and define the Lagrangian:

$$\mathcal{L}(\pi, \lambda; b) = J(\pi) + \lambda(D(\pi) - b).$$

The dual function is $g(\lambda; b) = \inf_\pi \mathcal{L}(\pi, \lambda; b) = \psi(\lambda) - \lambda b$, where $\psi(\lambda) \triangleq \inf_\pi (J(\pi) + \lambda D(\pi))$. Assuming Slater's condition holds, strong duality implies:

$$L(b) = \max_{\lambda \ge 0} (\psi(\lambda) - \lambda b). \tag{46}$$

Let $\lambda^*(b)$ be an optimal dual multiplier for a fixed budget b. Then, $L(b) = \psi(\lambda^*(b)) - \lambda^*(b)b$. Now, consider any arbitrary budget $\tilde{b} \in \mathcal{B}$. By (46), we have:

$$\begin{aligned}
L(\tilde{b}) &= \max_{\lambda \ge 0} (\psi(\lambda) - \lambda \tilde{b}) \\
&\ge \psi(\lambda^*(b)) - \lambda^*(b)\tilde{b}.
\end{aligned}$$

Substituting $\psi(\lambda^*(b)) = L(b) + \lambda^*(b)b$ into the inequality:

$$\begin{aligned} L(\tilde{b}) &\geq L(b) + \lambda^*(b)b - \lambda^*(b)\tilde{b} \\ &= L(b) + (-\lambda^*(b))(\tilde{b} - b). \end{aligned}$$

By the definition of the subgradient, this inequality implies that $-\lambda^*(b) \in \partial L(b)$. Restoring the episode index, we have $-\lambda_k^*(b_k) \in \partial L_k^*(b_k)$. □

B. Proof of the OCO with Switching Cost Regret Bounds

Proof. It suffices to prove the regret bound of OCO with switching costs. Recall that the regret is given by:

$$\begin{aligned} \mathrm{Reg}_{\mathrm{UL}}(K) &= \sum_{k=1}^{K} \Big(\big(f_k(b_k) + \alpha \|b_k - b_{k-1}\|^2 + \beta L_k^*(b_k) \big) \\ &\quad - \big(f_k(b^*) + \beta L_k^*(b^*) \big) \Big) \\ &= \sum_{k=1}^{K} (g_k(b_k) - g_k(b^*)) + \sum_{k=1}^{K} \alpha \|b_k - b_{k-1}\|^2. \end{aligned}$$

where $g_k(b) \triangleq f_k(b) + \beta L_k^*(b)$. Recall that each service cost f_k is θ_f-strongly convex by Assumption 1. By Lemma 11, $L_k^*(b)$ is convex in b. Since the sum of a strongly convex function and a convex function preserves strong convexity, the induced upper-level loss $g_k(b)$ is θ_g-strongly convex over $\mathcal{B} = [B_0, T]$ with $\theta_g = \theta_f$.

Since the true transition model is unknown, the algorithm operates on a surrogate CMDP induced by the extended LP. Let $\hat{L}_k(b) \triangleq V_{\tilde{l}}(\pi_k; P_k)$ denote the optimistic surrogate value function. Similarly, $\hat{L}_k(b)$ is convex. Define the surrogate objective as $\hat{g}_k(b) \triangleq f_k(b) + \beta \hat{L}_k(b)$.

The per-episode upper-level regret can be decomposed as follows: (i) for the warm-up phase $1 \leq k \leq K_0$, the regret is $g_k(B_0) - g_k(b^*)$; (ii) for the online learning phase $k > K_0$, the regret term (including switching cost) is:

$$\begin{aligned} &g_k(b_k) - g_k(b^*) + \alpha \|b_k - b_{k-1}\|^2 \\ &= \underbrace{\hat{g}_k(b_k) - \hat{g}_k(b^*) + \alpha \|b_k - b_{k-1}\|^2}_{\text{Surrogate OCO Regret}} \\ &\quad + \underbrace{\beta \big(L(b_k) - \hat{L}(b_k) \big) - \beta \big(L(b^*) - \hat{L}(b^*) \big)}_{\text{Approximation Error}}. \end{aligned}$$

For $k > K_0$, the projected online subgradient update is

$$b_{k+1} = \Pi_{\mathcal{B}} \big(b_k - \eta_k \hat{h}_k \big)$$

where $\hat{h}_k = \nabla f_k(b_k) - \beta \lambda_k$ and $\eta_k = \frac{1}{\theta_g (k - K_0)}$, here, λ_k is the optimal dual multiplier associated with the budget constraint in the lower-level Extended LP (23b) at episode k. To establish the boundedness of the subgradient $\hat{h}_k$, we rely on the convexity and subgradient of the value function $\hat{L}_k(b)$ (Lemmas 11 and 12). Utilizing the safe baseline policy with $b_a < B_0$, and the convex function subgradient inequality $\hat{L}_k(b_a) - \hat{L}_k(b_k) \geq -\lambda_k (b_a - b_k)$, we obtain the bound:

$$\lambda_k \leq \frac{\hat{L}_k(b_a) - \hat{L}_k(b_k)}{b_k - b_a} < \frac{T}{\gamma}.$$

The feasible region $\mathcal{B}$ has diameter $D = T - B_0$. Since $\|\nabla f_k(b)\|$ is bounded by F, and $\|\hat{h}_k\| = \|\nabla f_k - \beta \lambda_k\| \leq \|\nabla f_k\| + \|\beta \lambda_k\|$ there exists a constant $G \triangleq F + \beta \frac{T}{\gamma}$ such that the surrogate subgradient is bounded by $\|\hat{h}_k\| \leq G$.

By the θ_g-strong convexity of $\hat{g}_k$, we have:

$$\hat{g}_k(b_k) - \hat{g}_k(b^*) \leq \langle \hat{h}_k, b_k - b^* \rangle - \frac{\theta_g}{2} \|b_k - b^*\|^2. \quad (47)$$

Using the update rule for b_{k+1} and the property of Euclidean projection:

$$\begin{aligned} \|b_{k+1} - b^*\|^2 &= \|\Pi_{[B_0,T]} \big(b_k - \eta_k \hat{h}_k \big) - b^*\|^2 \\ &\leq \|b_k - \eta_k \hat{h}_k - b^*\|^2 \\ &= \|b_k - b^*\|^2 - 2\eta_k \langle \hat{h}_k, b_k - b^* \rangle + \eta_k^2 \|\hat{h}_k\|^2. \end{aligned}$$

Rearranging the terms yields:

$$\langle \hat{h}_k, b_k - b^* \rangle \leq \frac{1}{2\eta_k} \big(\|b_k - b^*\|^2 - \|b_{k+1} - b^*\|^2 \big) + \frac{\eta_k}{2} \|\hat{h}_k\|^2. \quad (48)$$

Substituting (48) into (47) and using $\|\hat{h}_k\| \leq G$, $\|b_k - b^*\| \leq D$, we obtain:

$$\begin{aligned} \hat{g}_k(b_k) - \hat{g}_k(b^*) &\leq \frac{1}{2\eta_k} \big(\|b_k - b^*\|^2 - \|b_{k+1} - b^*\|^2 \big) \\ &\quad + \frac{\eta_k G^2}{2} - \frac{\theta_g}{2} \|b_k - b^*\|^2. \end{aligned}$$

For the switching cost:

$$\begin{aligned} \|b_k - b_{k-1}\| &= \|\Pi_{[B_0,T]}(b_{k-1} - \eta_{k-1} \hat{h}_{k-1}) - b_{k-1}\| \\ &\leq \|\eta_{k-1} \hat{h}_{k-1}\| \leq G \eta_{k-1}. \end{aligned}$$

Summing from $k = 1$ to K, and setting $\eta_{K_0} = 0$, we obtain:

$$\begin{aligned} &\mathrm{Reg}_{\mathrm{UL}}(K, b) = \sum_{k=1}^{K} \big(g_k(b_k) - g_k(b^*) \big) + \sum_{k=1}^{K} \alpha \|b_k - b_{k-1}\|^2 \\ &= \sum_{k=1}^{K_0} (g_k(B_0) - g_k(b^*)) \\ &\quad + \sum_{k=K_0+1}^{K} (g_k(b_k) - g_k(b^*)) + \sum_{k=K_0+1}^{K} \alpha \|b_k - b_{k-1}\|^2 \\ &\leq K_0 G D + \sum_{k=K_0+1}^{K} \left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}} - \frac{\theta_g}{2} \right) \|b_k - b^*\|^2 \\ &\quad + \sum_{k=K_0+1}^{K} \frac{\eta_k G^2}{2} + \sum_{k=K_0+1}^{K} (G \eta_k)^2 \\ &\quad + \sum_{k=K_0}^{K} \beta \big((L_k^*(b_k) - \hat{L}_k(b_k)) - (L_k^*(b^*) - \hat{L}_k(b^*)) \big). \end{aligned} \quad (49)$$

With the step size $\eta_k = \frac{1}{\theta_g (k - K_0)}$, the following bounds hold:

$$\sum_{k=K_0+1}^{K} \frac{\eta_k G^2}{2} = \sum_{k=1}^{K-K_0-1} \frac{G^2}{2\theta_g k} \leq \frac{G^2}{2\theta_g} (1 + \log K). \quad (50)$$

$$\sum_{k=K_0+1}^{K} \left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}} - \frac{\theta_g}{2}\right)\left(\|b_k - b^*\|^2\right) = 0. \tag{51}$$

$$\sum_{k=K_0+1}^{K} (G\eta_k)^2 = G^2 \sum_{k=1}^{K-K_0} \frac{1}{\theta_g^2 k^2} = \frac{G^2\pi^2}{6\theta_g^2}. \tag{52}$$

For the term $L(b_k) - \hat{L}(b_k)$, as $L(b_k) = V_{\bar{l}}(\pi_k^{\mathrm{A}}; P_k)$, $\hat{L}(b_k) = V_l(\pi_k^{\mathrm{A}^*}; P)$, then by lemma 10, we can get $V_{\bar{l}}(\pi_k^{\mathrm{A}}; P_k) \leq V_l(\pi_k^{\mathrm{A}^*}; P)$. By (43) and (44), we can get

$$\sum_{k=K_0+1}^{K} V_l(\pi_k^{\mathrm{A}}; P) - V_{\bar{l}}(\pi_k^{\mathrm{A}}; P_k) \leq \tilde{O}\left(\frac{T^3}{\gamma} S\sqrt{AK}\right)$$

Consequently, we obtain:

$$\begin{aligned}\sum_{k=K_0+1}^{K} \left(L(b_k) - \hat{L}(b_k)\right) &= \sum_{k=K_0+1}^{K} \left(V_l(\pi_k^{\mathrm{A}^*}; P) - V_{\bar{l}}(\pi_k^{\mathrm{A}}; P_k)\right) \\ &\leq \tilde{O}\left(\frac{T^3}{\gamma} S\sqrt{AK}\right). \end{aligned} \tag{53}$$

Since $L(b_k) \geq L(b^*)$, it follows that:

$$\sum_{k=K_0+1}^{K} \left(L(b^*) - \hat{L}(b^*)\right) \leq \tilde{O}\left(\frac{T^3}{\gamma} S\sqrt{AK}\right). \tag{54}$$

Adding (50) – (54) to (49), we obtain a bound of the form:

$$\begin{aligned}\mathrm{Reg}_{\mathrm{UL}}(K, b) \leq\ & \frac{G^2}{2\theta_g}\left(1 + \log K\right) + \tilde{O}\left(\frac{T^3}{\gamma} S\sqrt{AK}\right) \\ & + K_0 GD + \frac{G^2\pi^2}{6\theta_g^2}. \end{aligned} \tag{55}$$

□

Appendix C
Proof of Theorem 1

Proof. Recall that the total regret after K episodes is defined as

$$\mathrm{Reg}(K, b, \pi) \triangleq \sum\nolimits_{k=1}^{K} C_k(b_k, \pi_k^{\mathrm{A}}) - \sum\nolimits_{k=1}^{K} C_k(b^*, \pi^{\mathrm{A},*}).$$

For each episode k and a fixed upper-level decision b_k, define the *episode-wise optimal safe lower-level policy* under the true model P:

$$\pi_k^*(b_k) \in \arg\min\nolimits_{\pi \in \Pi(b_k)} V_l(\pi; P), \tag{56}$$

where $\Pi(b_k)$ denotes the feasible policy set induced by budget b_k. Using $\pi_k^*(b_k)$ as an intermediate comparator, we can decompose the total regret as follows:

$$\begin{aligned}\mathrm{Reg}(K, b, \pi) = & \underbrace{\sum\nolimits_{k=1}^{K} \left(C_k(b_k, \pi_k^{\mathrm{A}}) - C_k(b_k, \pi_k^*(b_k))\right)}_{\text{(I) Lower-level BALDE regret under budget } b_k\text{: } \mathrm{Reg}_{\mathrm{LL}}(K,\pi)} \\ + & \underbrace{\sum\nolimits_{k=1}^{K} \left(C_k(b_k, \pi_k^*(b_k)) - C_k(b^*, \pi^{\mathrm{A},*})\right)}_{\text{(II) Upper-level BLOL regret with switching costs: } \mathrm{Reg}_{\mathrm{UL}}(K,b)}. \end{aligned} \tag{57}$$

By the the definition of C_k in (4), Term (I) in (57) simplifies to:

$$\begin{aligned}&\sum\nolimits_{k=1}^{K} \left(C_k(b_k, \pi_k^{\mathrm{A}}) - C_k(b_k, \pi_k^*(b_k))\right) \\ &= \sum\nolimits_{k=1}^{K} \Big([f_k(b_k) + \alpha\left\|b_k - b_{k-1}\right\|^2 + \beta V_l(\pi_k^{\mathrm{A}}; P)] \\ &- [f_k(b_k) + \alpha\left\|b_k - b_{k-1}\right\|^2 + \beta V_l(\pi_k^*(b_k); P)]\Big) \\ &= \beta \sum\nolimits_{k=1}^{K} \left(V_l(\pi_k^{\mathrm{A}}; P) - V_l(\pi_k^*(b_k); P)\right) \\ &= \beta R(K). \end{aligned} \tag{58}$$

Term (II) in (57) represents the upper-level BLOL regret with switching costs, i.e., the provisioning regret. Using the oracle effective loss $g_k(b) \triangleq f_k(b) + \beta L_k^*(b)$, where $L_k^*(b) = \min_{\pi \in \Pi(b)} V_l(\pi; P)$ (see (9)), the upper-level regret term $\mathrm{Reg}_{\mathrm{UL}}(K, b)$ becomes:

$$\sum\nolimits_{k=1}^{K} \left(g_k(b_k) - g_k(b^*)\right) + \alpha \sum\nolimits_{k=1}^{K} \|b_k - b_{k-1}\|^2. \tag{59}$$

Combining (57)–(59), Lemma 1, and Lemma 2, we obtain

$$\begin{aligned}\mathrm{Reg}(K, b, \pi) &= \mathrm{Reg}_{\mathrm{LL}}(K, \pi) + \mathrm{Reg}_{\mathrm{UL}}(K, b) \\ &= \beta R(K) + \mathrm{Reg}_{\mathrm{UL}}(K, b) \\ &= \tilde{\mathcal{O}}\left(\frac{ST^3\sqrt{AK}}{\gamma}\right) + \tilde{\mathcal{O}}(\log K). \end{aligned} \tag{60}$$

□

Appendix D
Proof of Theorem 2

In this section, we provide a detailed proof of Theorem 2. The theorem establishes that, conditioned on the good event $\mathcal{G}$, the safety constraint $V_d(\pi_k; P) \leq b_k$ holds for all episodes under the BALDE algorithm.

Proof. The proof is conditioned on the good event $\mathcal{G}$. We analyze the two phases of the algorithm separately.

Case 1: Warm-up phase ($k \leq K_0$). During this phase, the algorithm executes $\pi_k = \pi_{\mathrm{base}}$ and the budget is fixed at $b_k = B_0$. Since the baseline is safe with respect to B_0, we have:

$$V_d(\pi_k; P) = V_d(\pi_{\mathrm{base}}; P) = b_{\mathrm{base}} < B_0 = b_k. \tag{61}$$

Thus, the constraint is trivally satisfied.

Case 2: Online learning phase ($k > K_0$). By Proposition 1, the DOP problem is feasible for $k > K_0$. Let (π_k, P_k) be the solution. We apply the Value Difference Lemma (Lemma 8) to the constriant cost function d:

$$\begin{aligned}&V_d(\pi_k; P) - V_d(\pi_k; P_k) \\ &= \mathbb{E}_{\pi_k, P_k}\left[\sum_{t=1}^{T} \left((P_t - P_{t,k})(\cdot \mid s_{k,t}, a_{k,t})\right)^\top V_{d,t+1}^{\pi_k}(\cdot; P)\right]. \end{aligned}$$

Using Hölder's inequality and the fact that $d \in [0,1]$ implies $\|V_{d,t+1}^{\pi_k}\|_\infty \leq T$, we bound the inner term:

$$\begin{aligned}
&\left|\left((P_t - P_{t,k})(\cdot \mid s_{k,t}, a_{k,t})\right)^\top V_{d,t+1}^{\pi_k}\right| \\
&\leq \|P_t(\cdot \mid s_{k,t}, a_{k,t}) - P_{t,k}(\cdot \mid s_{k,t}, a_{k,t})\|_1 \cdot \|V_{d,t+1}^{\pi_k}\|_\infty \\
&\leq \sum_{s'} \beta_{k,t}^p(s_{k,t}, a_{k,t}, s') \cdot T.
\end{aligned}$$

Taking the expectation, we obtain:

$$\begin{aligned}
V_d(\pi_k; P) - V_d(\pi_k; P_k) &\leq T\, \mathbb{E}_{\pi_k, P_k}\left[\sum_{t=1}^{T}\sum_{s'} \beta_{k,t}^p(s_{k,t}, a_{k,t}, s')\right] \\
&= \epsilon_k^{\pi_k}(P_k). \qquad (62)
\end{aligned}$$

Rearranging (62), we have $V_d(\pi_k; P) \leq V_d(\pi_k; P_k) + \epsilon_k^{\pi_k}(P_k)$. Recall that the DOP problem imposes the pessimistic constraint $V_{\bar{d}}(\pi_k; P_k) \leq b_k$. Using the decomposition property $V_{\bar{d}}(\pi; P') = V_d(\pi; P') + \epsilon_k^{\pi}(P')$, we have:

$$V_d(\pi_k; P_k) + \epsilon_k^{\pi_k}(P_k) = V_{\bar{d}}(\pi_k; P_k) \leq b_k. \qquad (63)$$

Combining these inequalities yields:

$$V_d(\pi_k; P) \leq V_d(\pi_k; P_k) + \epsilon_k^{\pi_k}(P_k) \leq b_k. \qquad (64)$$

This confirms that the safety constraint is satisfied in the true environment for all $k > K_0$.

Combining Case 1 and Case 2, the constraint is never violated with probability at least $1 - 3\delta$. □