Lyapunov-Driven Optimistic Learning for Online Scheduling with Multi-Stage Tasks

Yutong Huang
ShanghaiTech University
huangyt2025@shanghaitech.edu.cn

Xin Liu*
ShanghaiTech University
liuxin7@shanghaitech.edu.cn

Abstract

We study online scheduling in a shared-resource system where tasks of different types arrive sequentially and compete for limited resources. Each task is modeled as an H-stage episodic Markov decision process, which captures the sequential resource-allocation decisions made within the task. At the system level, the learner must coordinate resource sharing across tasks to maximize the long-term ratio of cumulative reward to cumulative cost over K tasks, without knowledge of the task distribution. The key challenges are coupled intra-task Markovian dynamics and inter-task ratio maximization, further complicated by unknown models and large or continuous state spaces. To address these challenges, we propose *Contextual Optimistic Lyapunov Decomposition* (COLD), a Lyapunov-driven optimistic learning algorithm. COLD converts the long-term ratio objective into a queue-adjusted per-task surrogate, thereby decoupling reward-cost optimization from statistical value estimation. It uses double-optimistic least-squares value iteration to estimate reward and cost values under linear function approximation, and a softmax policy to smooth the greedy update with only a controlled approximation error. We prove a regret bound of $\tilde{\mathcal{O}}(\sqrt{Md^3H^3T})$, where M is the number of task types, d is the feature dimension, H is the episode horizon, and $T = KH$ is the number of interaction steps. The $\sqrt{T}$ dependence is order-optimal up to logarithmic factors, and the bound scales with the feature dimension rather than the size of the state space. Experiments on synthetic environments and an empirical KuaiRand-based MDP demonstrate the effectiveness of the proposed method.

1 Introduction

Online task scheduling studies how to allocate limited resources to a sequence of arriving tasks, and arises in applications such as video recommendation, cloud computing, and crowdsourcing (Gao et al., 2022; Ho and Vaughan, 2012). A natural objective is to maximize the *reward-to-cost ratio*, which captures overall system efficiency (Neely, 2011, 2013; Xu et al., 2025). In practice, arriving tasks often differ in their types, resource demands, and reward-cost profiles. Each task may also unfold over multiple decision stages, where early decisions affect later states and outcomes through the underlying system dynamics. For example, in a shared cloud computing platform, a task may proceed through data loading, distributed computation, and result aggregation, with each stage incurring different costs and contributing differently to the final reward. Across tasks, they may have different resource demands and compete for a shared pool of limited computation and memory resources, so allocating more resources to the current task may improve its performance but also incur resource costs, thereby affecting the long-term reward-to-cost tradeoff over the task sequence. These features motivate a multi-task episodic Markov decision process (MDP) formulation, where the objective is to

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

maximize the long-term ratio of cumulative reward to cumulative cost over a sequence of K tasks (Neely, 2021; Xu et al., 2025).

This problem involves two coupled challenges: *intra-task dependence* and *inter-task dependence*. The first arises within each task, where decisions interact through unknown Markovian transitions, rewards, and costs, making learning and planning difficult, especially in large or continuous state spaces (Jin et al., 2020; Xu et al., 2025). The second arises across tasks, where tasks of different types compete for shared resources, and the learner must optimize a long-term ratio of cumulative reward to cumulative cost without knowing the task distribution (Neely, 2021, 2025; Xu et al., 2025).

However, most existing literature addresses intra-task dynamics and inter-task resource competition in isolation. For *intra-task dependence*, prior work in stochastic control and reinforcement learning has studied ratio or fractional objectives in Markovian systems, including ratio MDPs, fractional-cost MDPs, and cost-aware MDPs (Aggarwal et al., 1977; Ren and Krogh, 2005; Tanaka, 2017, 2019; Sinha and Mondal, 2017; Suttle et al., 2021). These works focus on a single fixed Markovian environment and therefore do not capture online arrivals of different task types. For *inter-task* dependence, another line of work studies renewal systems and online task scheduling under long-term efficiency objectives. These works address how to optimize reward-per-cost performance across a sequence of tasks when task statistics are unknown, using tools such as Lyapunov-based control, stochastic approximation, and adaptive renewal optimization (Neely, 2011, 2013, 2021, 2025). A related line of scheduling work further considers unknown rewards and costs under bandit feedback, but still focuses on single-stage tasks rather than multi-stage task dynamics (Xu et al., 2024).

The closest prior work is RM^{2}VI (Xu et al., 2025), which studies online scheduling of multi-stage MDP tasks under a long-term reward-to-cost objective. However, its regret bound is of order $T^{3/4}$ over T interaction steps, and its tabular analysis scales explicitly with the state and action space sizes.

To overcome these limitations, we propose **COLD** (Contextual Optimistic Lyapunov Decomposition), which decouples intra-task learning from inter-task ratio optimization and extends the framework beyond the tabular setting to large or structured state spaces. Our contributions are:

Algorithm design. We propose COLD, a Lyapunov-driven optimistic learning framework for online reward-to-cost optimization in multi-task linear MDPs. Motivated by Neely (2025), COLD introduces a virtual queue and an auxiliary denominator-scale variable to convert the long-term fractional objective into an episode-level drift-plus-penalty surrogate. This surrogate separates the cross-task ratio-control term from the task-level value-estimation errors, allowing the two sources of regret to be controlled independently. We use double-optimistic least-squares value learning to learn reward and cost values within each multi-stage task, enabling exploration and long-horizon planning in a shared feature space. We further smooth the greedy decision rule with a softmax policy, avoiding discontinuous policy updates induced by small estimation errors.

Theoretical analysis. We show that COLD achieves regret $\tilde{\mathcal{O}}(\sqrt{Md^3H^3T})$ under the multi-task linear MDP assumption, where M is the number of task types, d is the feature dimension, H is the episode length, and T is the total number of steps. Its $\sqrt{T}$ dependence is order-optimal up to logarithmic factors, since our formulation contains standard finite-horizon linear MDPs as a special case and thus inherits their $\Omega(\sqrt{T})$ information-theoretic lower bound (Azar et al., 2017; Hu et al., 2022). Moreover, the bound scales with the feature dimension rather than the size of the state space. To the best of our knowledge, this gives the first order-optimal-in-T regret guarantee for online scheduling of multi-stage MDP tasks with unknown reward, cost, and transition models under a reward-to-cost objective. This improvement comes from a decoupling perspective: rather than intertwining ratio optimization with value estimation, our analysis controls the Lyapunov optimization error and the estimation error separately. A softmax relaxation makes the induced value-function class amenable to uniform concentration, with its approximation error absorbed into the regret bound.

Experiments. We evaluate COLD on both synthetic and empirical environments. The synthetic experiment validates the method under our theoretical setting, while the KuaiRand experiment (Gao et al., 2022) demonstrates its practical effectiveness for engagement-efficient recommendation. In both cases, COLD achieves the best reward-to-cost ratio among the baselines.

2 Related work

Task-level resource-allocation models. Online resource allocation has been widely studied through contextual bandits with knapsack constraints (CBwK), where each arriving request reveals a context, the learner selects a context-dependent allocation action, receives a reward, and consumes limited resources (Badanidiyuru et al., 2014; Agrawal et al., 2016; Agrawal and Devanur, 2016). Recent work extends this framework to richer resource-allocation settings, including conversion-based reward-cost coupling, regression-oracle methods, packing and covering constraints, Lyapunov-based constrained learning, and small-budget regimes (Li and Stoltz, 2022; Han et al., 2023; Slivkins et al., 2024; Guo and Liu, 2024; Chzhen et al., 2023; Guo and Liu, 2025). These formulations capture contextual arrivals, rewards, costs, and resource constraints. However, they typically treat each allocation as a stateless context-dependent decision and do not explicitly model how current decisions affect subsequent states within a task.

Markov decision processes (MDPs) and reinforcement learning provide a natural framework for capturing such within-task temporal dependence in multi-stage decision making. Constrained MDPs incorporate long-term resource or safety constraints into stateful decision models (Altman, 1999; Bhatnagar, 2010; Wei et al., 2022; Wachi et al., 2024). More closely related to our reward-cost objective, Markov ratio, fractional-cost, and cost-aware MDP formulations study reward-cost tradeoffs in Markovian systems (Aggarwal et al., 1977; Ren and Krogh, 2005; Tanaka, 2017, 2019; Sinha and Mondal, 2017; Suttle et al., 2021). Orthogonally, optimistic RL with linear function approximation provides dimension-dependent guarantees for large state spaces (Yang and Wang, 2019; Cai et al., 2020; Jin et al., 2020). However, these lines of work typically study a single decision environment governed by fixed reward-cost and dynamical laws. In contrast, our setting consists of a sequence of arriving episodic tasks: different task types may have distinct reward, cost, and transition laws, while their realized rewards and costs jointly determine a system-level long-term reward-to-cost ratio. This cross-task coupling motivates a system-level view of online scheduling.

System-level online scheduling. Learning-based dispatching, queueing, and matching systems highlight the role of system-level operational dynamics, where decisions interact with evolving server workloads, queue backlogs, or matching markets under stochastic arrivals (Liu et al., 2020; Johari et al., 2021; Yang et al., 2023; Freund et al., 2023; Kim and Oh, 2024; Walton and Xu, 2021; Huang et al., 2024). These models capture cross-task operational coupling at the system level, but typically optimize service-level or constraint-based criteria, such as stability, delay, holding cost, matching quality, or constrained cumulative reward. In many online scheduling systems, however, one also seeks to optimize long-term operational efficiency. This perspective is captured by renewal-optimization methods, which exploit renewal-frame structure to optimize reward-per-time or reward-per-cost objectives (Neely, 2011, 2013, 2021, 2025). Building on this line of work, Xu et al. (2024) studied online task scheduling with unknown rewards, costs, and arrivals under bandit feedback. Unlike budgeted or constrained formulations, ratio objectives require no pre-specified budget or constraint threshold; instead, the tradeoff between reward and cost is determined endogenously by the long-term efficiency criterion. Existing renewal-based methods address system-level efficiency optimization across frames or tasks, but usually abstract each task as a single-stage decision and therefore miss the within-task temporal dependence from multi-stage execution.

Combining this system-level efficiency perspective with task-level Markovian modeling, recent work has begun to study online task scheduling where each arriving task is modeled as an episodic MDP and the objective is a long-term reward-to-cost ratio (Xu et al., 2025). This setting simultaneously captures within-task Markovian dynamics and cross-task efficiency optimization. Existing results in this direction are established in the tabular setting, with regret bounds that scale explicitly with the state-action space and exhibit $\mathcal{O}(T^{3/4})$ horizon dependence.

3 Preliminaries

Multi-task Markov decision processes. We consider an online scheduling problem modeled as a multi-task episodic Markov decision process $\mathrm{MDP}(\mathcal{M}, \mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c)$, where $\mathcal{M}$ is the task-type set with $M := |\mathcal{M}|$, $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite action space with $A := |\mathcal{A}|$, and H is the episode length. The state space may be large or continuous, while the action space is assumed finite throughout. All task types share the same state and action spaces, while the transition kernel, reward function, and cost function may depend on the task type. At the beginning of episode $k \in [K]$, a task type $m_k \in \mathcal{M}$

is drawn i.i.d. from an unknown distribution $\mathbb{P}_{\mathcal{M}}$, and the initial state is sampled as $s_1^k \sim \nu(m_k)$. At each step $h \in [H]$, the agent chooses a_h^k, receives reward $r_h(m_k, s_h^k, a_h^k) \in [0, r_{\max}]$, incurs cost $c_h(m_k, s_h^k, a_h^k) \in [c_{\min}, c_{\max}]$, and transitions to $s_{h+1}^k \sim \mathbb{P}_h(\cdot \mid m_k, s_h^k, a_h^k)$.

We use $\pi = \{\pi_h\}_{h=1}^H$ to denote a (possibly randomized) Markov policy, where $\pi_h(\cdot \mid m, s) \in \mathcal{P}(\mathcal{A})$. For any policy π, define the reward value functions $V_h^\pi(m, s) := \mathbb{E}_\pi[\sum_{t=h}^H r_t \mid s_h = s, m]$ and $Q_h^\pi(m, s, a) := \mathbb{E}_\pi[\sum_{t=h}^H r_t \mid s_h = s, a_h = a, m]$, where r_t is shorthand for $r_t(m, s_t, a_t)$. Similarly, define the cost value functions $W_h^\pi(m, s) := \mathbb{E}_\pi[\sum_{t=h}^H c_t \mid s_h = s, m]$ and $C_h^\pi(m, s, a) := \mathbb{E}_\pi[\sum_{t=h}^H c_t \mid s_h = s, a_h = a, m]$, where c_t is shorthand for $c_t(m, s_t, a_t)$.

For any function $V : \mathcal{M} \times \mathcal{S} \to \mathbb{R}$, define $[\mathbb{P}_h V](m, s, a) := \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot \mid m, s, a)}[V(m, s')]$. Then the Bellman equations are

$$\begin{aligned} Q_h^\pi(m, s, a) &= (r_h + \mathbb{P}_h V_{h+1}^\pi)(m, s, a), \qquad V_h^\pi(m, s) = \sum_{a \in \mathcal{A}} \pi_h(a \mid m, s)\, Q_h^\pi(m, s, a), \\ C_h^\pi(m, s, a) &= (c_h + \mathbb{P}_h W_{h+1}^\pi)(m, s, a), \qquad W_h^\pi(m, s) = \sum_{a \in \mathcal{A}} \pi_h(a \mid m, s)\, C_h^\pi(m, s, a), \end{aligned} \tag{1}$$

with terminal conditions $V_{H+1}^\pi(m, s) = W_{H+1}^\pi(m, s) = 0$.

In the online setting, the learning algorithm produces a sequence of policies $\Pi := \{\pi^k\}_{k=1}^K$, where π^k is the policy used in episode k. We evaluate Π by its cumulative reward-to-cost ratio

$$\theta^\Pi := \frac{\mathbb{E}\Big[\sum_{k=1}^K V_1^{\pi^k}(m_k, s_1^k)\Big]}{\mathbb{E}\Big[\sum_{k=1}^K W_1^{\pi^k}(m_k, s_1^k)\Big]}. \tag{2}$$

Here the expectation is taken over the task arrivals, the initial states, and the trajectory induced by the policy sequence Π.

The benchmark $\theta^\star$ is defined as the optimal cumulative reward-to-cost ratio achieved by any fixed Markov policy. We define the convergence gap and regret of Π by

$$\mathrm{Gap}(K) = |\theta^\star - \theta^\Pi|, \qquad \mathcal{R}(K) = K\,|\theta^\star - \theta^\Pi|. \tag{3}$$

Multi-task linear Markov decision processes. We consider a multi-task linear MDP with a shared feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. The feature representation is common across tasks, while the transition and reward/cost parameters remain task-specific so that each task admits its own linear value representation.

Assumption 1. *$\mathrm{MDP}(\mathcal{M}, \mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c)$ is a multi-task linear MDP if, for every $h \in [H]$ and $m \in \mathcal{M}$, there exist d unknown task-specific signed measures $\mu_h(m, \cdot) = \big(\mu_h^{(1)}(m, \cdot), \ldots, \mu_h^{(d)}(m, \cdot)\big)$ over $\mathcal{S}$ and unknown task-specific vectors $\theta_{r,h}(m), \theta_{c,h}(m) \in \mathbb{R}^d$ such that, for all $(m, s, a) \in \mathcal{M} \times \mathcal{S} \times \mathcal{A}$,*

$$\begin{aligned} &\mathbb{P}_h(\cdot \mid m, s, a) = \langle \phi(s, a), \mu_h(m, \cdot) \rangle, \\ &r_h(m, s, a) = \langle \phi(s, a), \theta_{r,h}(m) \rangle, c_h(m, s, a) = \langle \phi(s, a), \theta_{c,h}(m) \rangle. \end{aligned} \tag{4}$$

Under Assumption 1, the one-step transition, reward, and cost models are linear in the shared feature map $\phi(s, a)$. We intentionally keep the task dependence in the coefficients rather than in the feature map: this preserves a common state-action representation across tasks while allowing each task type to induce its own linear Q- and C-function representation.

Proposition 1. *For any policy π, task $m \in \mathcal{M}$, and step $h \in [H]$, there exist task-specific vectors $w_{r,h}^\pi(m), w_{c,h}^\pi(m) \in \mathbb{R}^d$ such that, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$Q_h^\pi(m, s, a) = \langle \phi(s, a), w_{r,h}^\pi(m) \rangle, \qquad C_h^\pi(m, s, a) = \langle \phi(s, a), w_{c,h}^\pi(m) \rangle.$$

The proof is deferred to Appendix B.

We also impose the following boundedness conditions to fix the scales in the analysis.

Assumption 2 (Boundedness and normalization). *For all* $(m,h,s,a) \in \mathcal{M} \times [H] \times \mathcal{S} \times \mathcal{A}$, $0 \le r_h(m,s,a) \le r_{\max}, 0 < c_{\min} \le c_h(m,s,a) \le c_{\max}, \|\phi(s,a)\|_2 \le 1$, *and for all* $(m,h) \in \mathcal{M} \times [H]$, $\max\Big\{\|\mu_h(m,\mathcal{S})\|_2, \|\theta_{r,h}(m)\|_2, \|\theta_{c,h}(m)\|_2\Big\} \le \sqrt{d}$.

Under Assumption 2, define $H_{\max} := \max\{Hr_{\max}, Hc_{\max}\}$, $\gamma_{\min} := 1/(Hc_{\max})$, and $\gamma_{\max} := 1/(Hc_{\min})$. Here $H_{\max}$ uniformly bounds the reward and cost values, while $[\gamma_{\min}, \gamma_{\max}]$ ensures that the auxiliary target $1/\gamma$ lies in the feasible range $[Hc_{\min}, Hc_{\max}]$ of cumulative episode costs.

4 Algorithm design

In this section, we present COLD for online reward-to-cost optimization in multi-task linear MDPs. For any episode τ and stage h, write $\phi_h^\tau := \phi(s_h^\tau, a_h^\tau)$, $r_h^\tau := r_h(m_\tau, s_h^\tau, a_h^\tau)$, and $c_h^\tau := c_h(m_\tau, s_h^\tau, a_h^\tau)$. For episode k, define $\mathcal{I}_k(m_k) := \{\tau < k : m_\tau = m_k\}$ as the set of past episodes of the same task type. For $x \in \mathbb{R}$, let $[x]_\ell^u := \min\{\max\{x, \ell\}, u\}$.

Algorithm 1 COLD (Contextual Optimistic Lyapunov Decomposition)

Require: $v, \eta, \lambda, \beta, \gamma_{\min}, \gamma_{\max}, r_{\max}, c_{\min}$

Initialize $J_1 = 0$, $\gamma_0 = \gamma_{\min}$; for all (m,h), set $\Lambda_h^1(m) = \lambda I$, $w_{r,h}^1(m) = 0$, $w_{c,h}^1(m) = 0$

for $k = 1, 2, \ldots, K$ **do**

  Observe task type m_k and initial state s_1^k

  Learning and planning:

  for $h = H, H-1, \ldots, 1$ **do**

$$w_{r,h}^k(m_k) \leftarrow (\Lambda_h^k(m_k))^{-1} \textstyle\sum_{\tau \in \mathcal{I}_k(m_k)} \phi_h^\tau \big[r_h^\tau + \hat{V}_{h+1}^k(m_k, s_{h+1}^\tau)\big] \tag{5}$$

$$w_{c,h}^k(m_k) \leftarrow (\Lambda_h^k(m_k))^{-1} \textstyle\sum_{\tau \in \mathcal{I}_k(m_k)} \phi_h^\tau \big[c_h^\tau + \check{W}_{h+1}^k(m_k, s_{h+1}^\tau)\big] \tag{6}$$

$$B_h^k(m_k,s,a) \leftarrow \beta\sqrt{\phi(s,a)^\top (\Lambda_h^k(m_k))^{-1}\phi(s,a)} \tag{7}$$

$$\hat{Q}_h^k(m_k,s,a) \leftarrow \big[\phi(s,a)^\top w_{r,h}^k(m_k) + B_h^k(m_k,s,a)\big]_0^{(H-h+1)r_{\max}} \tag{8}$$

$$\check{C}_h^k(m_k,s,a) \leftarrow \Big[\phi(s,a)^\top w_{c,h}^k(m_k) - B_h^k(m_k,s,a)\Big]_{(H-h+1)c_{\min}}^{(H-h+1)c_{\max}} \tag{9}$$

$$\pi_h^k(\cdot \mid m_k, s) \propto \exp\Big(\eta\big[v\hat{Q}_h^k(m_k,s,\cdot) - J_k\check{C}_h^k(m_k,s,\cdot)\big]\Big) \tag{10}$$

$$\hat{V}_h^k(m_k,s) \leftarrow \langle \pi_h^k, \hat{Q}_h^k(m_k,s,\cdot)\rangle_{\mathcal{A}}, \quad \check{W}_h^k(m_k,s) \leftarrow \langle \pi_h^k, \check{C}_h^k(m_k,s,\cdot)\rangle_{\mathcal{A}}$$

  Execution and data update:

  for $h = 1, 2, \ldots, H$ **do**

    Execute $a_h^k \sim \pi_h^k(\cdot \mid m_k, s_h^k)$ and observe $(r_h^k, c_h^k, s_{h+1}^k)$

    $\Lambda_h^{k+1}(m_k) \leftarrow \Lambda_h^k(m_k) + \phi_h^k(\phi_h^k)^\top$, all other task-specific estimates are carried over.

  Queue and auxiliary update:

$$\gamma_k \leftarrow \arg\min_{\gamma \in [\gamma_{\min}, \gamma_{\max}]} \left\{\frac{\gamma}{\gamma_{k-1}}\big[J_k\check{W}_1^k(m_k,s_1^k) - v\hat{V}_1^k(m_k,s_1^k)\big] + \frac{v^2}{2}(\gamma - \gamma_{k-1})^2\right\} \tag{11}$$

$$J_{k+1} \leftarrow \big[J_k + \check{W}_1^k(m_k,s_1^k) - 1/\gamma_k\big]_+ \tag{12}$$

COLD is organized around three components: Lyapunov-based ratio control across tasks, double-optimistic value learning within each task, and softmax-smoothed decision making. We describe these components and show how the Lyapunov layer separates ratio regulation from value estimation.

Adaptive Lyapunov optimization for inter-task dependence: Rather than tracking the reward-to-cost ratio directly, COLD uses a Lyapunov drift-plus-penalty mechanism to convert the non-additive fractional objective into a queue-adjusted per-episode surrogate.

The main difficulty is that the accumulated cost appears in the denominator of the objective, rather than as an additive penalty that can be directly handled by the standard drift-plus-penalty framework. Following the adaptive renewal-optimization construction of Neely (2025), COLD introduces an auxiliary denominator-scale variable γ_k and a virtual queue J_k. Here $1/\gamma_k$ serves as a dynamically adjusted episode-level cost target, and the queue update in (12) accumulates the deviation between the estimated episode cost and this target. When the estimated cost is persistently above the target,

the queue grows, making future decisions more cost-sensitive through the queue-adjusted policy update in (10). This queue construction yields the one-step drift-plus-penalty bound

$$\Delta_k - v\hat{V}_1^k(m_k, s_1^k) \le b - v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right), \tag{13}$$

where $b := \frac{1}{2}(Hc_{\max} - Hc_{\min})^2$ is a constant. Thus, for a fixed denominator target, the fractional reward-cost tradeoff is reduced to an additive per-episode surrogate involving the reward term $-v\hat{V}_1^k(m_k, s_1^k)$ and the queue-weighted cost term $J_k\check{W}_1^k(m_k, s_1^k)$.

It remains to choose the denominator scale γ_k. The update in (11) minimizes a regularized auxiliary objective derived from (13): up to terms independent of γ, its linear term is obtained by multiplying the queue-adjusted reward-cost term in (13) by γ/γ_{k-1}. This term relates to the desired reward-to-cost objective in the adaptive renewal formulation. The second term in (11) is a proximal regularizer that keeps γ_k close to γ_{k-1}, so that the adaptive denominator target changes smoothly rather than distorting the original reward-to-cost objective. Thus, γ_k is not a direct estimate of the optimal ratio; it is an adaptive denominator-scale variable selected by a regularized drift-plus-penalty step.

The bound in (13) enables the desired error decomposition. The Lyapunov drift term controls the structural error from long-term reward-cost optimization. The remaining statistical error is isolated by adding and subtracting the true policy values $V_1^{\pi^k}(m_k, s_1^k)$ and $W_1^{\pi^k}(m_k, s_1^k)$, which leaves only the gaps between the optimistic/pessimistic estimates $\hat{V}_1^k(m_k, s_1^k), \check{W}_1^k(m_k, s_1^k)$ and their true counterparts under π^k. This separation provides the basis for the order-optimal-in-T regret analysis.

Double-optimistic LSVI-UCB for intra-task dependence: To address the intra-task dependence induced by the multi-stage decision structure within each task, COLD performs backward value estimation via a task-specific linear LSVI-UCB procedure under the shared feature map $\phi(s, a)$ (Jin et al., 2020). For each task type m_k and step h, the algorithm computes the reward and cost estimators according to (5) and (6), together with the confidence bonus (7), and then constructs the reward and cost value estimates (8) and (9). Following Xu et al. (2025), we adopt a double-optimistic design, where a positive bonus is added to reward estimation and a negative bonus is applied to cost estimation, so that the surrogate objective $v\hat{Q}_h^k(m_k, s, \cdot) - J_k\check{C}_h^k(m_k, s, \cdot)$ is aligned with ratio optimization. Moreover, compared with the tabular value-iteration procedure in Xu et al. (2025), this linear instantiation carries out the same double-optimistic construction in a shared feature space, so that the statistical complexity is governed by the feature dimension d rather than by the size of the state space, while the finite action space enters only through logarithmic factors.

Softmax-smoothed decision making: Given the queue-adjusted score $v\hat{Q}_h^k(m_k, s, \cdot) - J_k\check{C}_h^k(m_k, s, \cdot)$, a natural alternative is to select the action with the largest score. However, this hard greedy rule induces a discontinuous policy map with respect to the estimated reward and cost values: when two actions have similar scores, a small estimation error can completely change the selected action. Such discontinuity makes the algorithm-generated policies and their induced value estimates difficult to control uniformly in the regret analysis. COLD therefore uses the softmax policy in (10) as a smooth approximation to greedy maximization. Unlike hard argmax, the softmax map varies continuously with the queue-adjusted scores and smooths decisions among near-tie actions. This smoothing makes the induced policy-dependent value estimates more amenable to the concentration analysis, while introducing only a controlled approximation gap relative to hard maximization.

In summary, COLD views online scheduling with multi-stage tasks through two coupled layers. At the cross-task level, a Lyapunov layer uses the queue–auxiliary-variable pair (J_k, γ_k) to turn the long-term fractional objective into a queue-adjusted per-episode surrogate (Neely, 2025). This surrogate separates ratio regulation from value estimation, providing the structural basis for the improved regret guarantee. At the task level, a double-optimistic linear LSVI-UCB layer estimates reward and cost values under the shared feature map, extending tabular RM^{2}VI to the linear setting (Xu et al., 2025; Jin et al., 2020). These estimates induce a softmax-smoothed policy, which avoids discontinuous greedy updates while adding only a controlled approximation term.

5 Main results and proof sketch

In this section, we state the main regret guarantee for COLD in the following theorem.

Theorem 1. *Under Assumptions 1 and 2, suppose that the action space $\mathcal{A}$ is finite. There exists an absolute constant $c > 0$ such that, if we set $\lambda = 1$, $v = H^2\sqrt{K}$, $\eta = \sqrt{K}$, $p = \frac{1}{K}$ and the exploration radius $\beta = c \cdot dH_{\max}\sqrt{\iota}$ with $\iota := \log(2MAdT/p)$, then the cumulative regret of COLD over K episodes is at most $\mathcal{O}\Big(\sqrt{Md^3H^3T\,\iota^2}\Big)$, where $\mathcal{O}(\cdot)$ hides only absolute constants.*

Theorem 1 shows that, under the parameter choices $v = H^2\sqrt{K}$, $\eta = \sqrt{K}$, and suitable β and λ, COLD achieves a regret bound of $\tilde{\mathcal{O}}(\sqrt{Md^3H^3T})$. This improves over the prior RM^{2}VI guarantee $\mathcal{O}(\sqrt{HMAS}T^{3/4})$ (Xu et al., 2025). RM^{2}VI uses a Robbins–Monro-type ratio update driven by learned reward and cost estimates, which couples ratio optimization with value-estimation errors. In contrast, COLD decouples these two components through a Lyapunov-based design and improves the dependence on the total interaction length to the $\sqrt{T}$ scale. This $\sqrt{T}$ dependence is order-optimal up to logarithmic factors. Indeed, our formulation contains standard finite-horizon linear MDPs as a special case: by taking a single task type with deterministic unit costs, the reward-to-cost objective reduces to the standard cumulative-reward objective up to a constant scaling. Hence, the known $\Omega(\sqrt{T})$ information-theoretic lower bound for online reinforcement learning applies to our setting as well (Azar et al., 2017; Hu et al., 2022). In addition to attaining the optimal dependence on T, our bound replaces the state-space dependence of tabular guarantees with feature-dimension dependence, making COLD applicable to large or structured state spaces where tabular methods are impractical. Extensions to continuous action spaces, nonstationary task distributions, or adversarial arrivals remain interesting future directions. We next illustrate the key techniques used to prove Theorem 1.

5.1 Proof sketch of Theorem 1

At a high level, the proof proceeds in three steps: we first convert the reward-to-cost regret into a centered reward-cost residual, then decompose this residual into a Lyapunov structural term and linear value-estimation errors, and finally bound the two parts separately.

Before stating the proof sketch, we introduce the attainable performance set

$$\Gamma \subseteq [0, Hr_{\max}] \times [Hc_{\min}, Hc_{\max}],$$

defined as the set of all pairs $\big(\mathbb{E}[V_1^\pi(m_1, s_1^1)],\ \mathbb{E}[W_1^\pi(m_1, s_1^1)]\big)$ induced by possibly randomized within-episode policies π, where the expectation is taken over the task type, the initial state, and the trajectory randomness. Let $\bar{\Gamma}$ denote the closure of Γ. Since every $(r, c) \in \bar{\Gamma}$ satisfies $c \geq Hc_{\min} > 0$, the optimal ratio is well defined as $\theta^\star := \sup_{(r,c)\in\bar{\Gamma}} \frac{r}{c}$. See Appendix D.3 for further properties of Γ and the existence of an optimizer.

Let $(r^\star, c^\star) \in \bar{\Gamma}$ be an optimal pair with $\theta^\star = r^\star/c^\star$. Recall from (2) that θ^Π denotes the cumulative reward-to-cost ratio of Π, and that $\mathrm{Gap}(K) = |\theta^\star - \theta^\Pi|$. By Lemma 14, $\theta^\Pi \leq \theta^\star$, we have

$$\mathrm{Gap}(K) = \frac{-\sum_{k=1}^K \mathbb{E}\Big[V_1^{\pi^k}(m_k, s_1^k) - \theta^\star W_1^{\pi^k}(m_k, s_1^k)\Big]}{\sum_{k=1}^K \mathbb{E}\big[W_1^{\pi^k}(m_k, s_1^k)\big]}.$$

Moreover, $W_1^{\pi^k}(m_k, s_1^k) \geq Hc_{\min}$, so the denominator is at least $KHc_{\min}$. Therefore, bounding $\mathrm{Gap}(K)$ reduces to lower bounding the centered reward-cost residual

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\Big[V_1^{\pi^k}(m_k, s_1^k) - \theta^\star W_1^{\pi^k}(m_k, s_1^k)\Big].$$

For brevity, let $V_1^{\pi^k} := V_1^{\pi^k}(m_k, s_1^k)$, $W_1^{\pi^k} := W_1^{\pi^k}(m_k, s_1^k)$, $\hat{V}_1^k := \hat{V}_1^k(m_k, s_1^k)$, and $\check{W}_1^k := \check{W}_1^k(m_k, s_1^k)$. We then decompose this centered residual by introducing the optimistic reward estimate $\hat{V}_1^k$ and the pessimistic cost estimate $\check{W}_1^k$:

$$\sum_{k=1}^K \mathbb{E}\Big[V_1^{\pi^k} - \theta^\star W_1^{\pi^k}\Big] = \underbrace{\sum_{k=1}^K \mathbb{E}\Big[\hat{V}_1^k - \theta^\star \check{W}_1^k\Big]}_{\text{structural surrogate term}} - \underbrace{\sum_{k=1}^K \mathbb{E}\Big[\hat{V}_1^k - V_1^{\pi^k}\Big]}_{\text{reward estimation error}} - \theta^\star \underbrace{\sum_{k=1}^K \mathbb{E}\Big[W_1^{\pi^k} - \check{W}_1^k\Big]}_{\text{cost estimation error}}.$$

This decomposition separates the proof into two parts: controlling the queue-adjusted surrogate term $\sum_k(\hat{V}_1^k - \theta^\star\check{W}_1^k)$ and controlling the cumulative estimation errors.

For the first part, the analysis hinges on a one-step drift-plus-penalty inequality established in Appendix D Lemma 16. It shows that, for every episode $k \geq 2$,

$$\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k\Big] \leq -v\theta^\star \mathbb{E}[\check{W}_1^{k-1}] + v\theta^\star \mathbb{E}[J_k - J_{k-1}] + \frac{v^2}{2}\,\mathbb{E}\left[\left(\frac{1}{c^\star} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c^\star} - \gamma_k\right)^2\right] + \mathcal{O}\left(H^4 + \frac{H\log A}{\eta} + Hpv\right),$$

where $L_k := \frac{1}{2}J_k^2$ is the Lyapunov function and $\Delta_k := L_{k+1} - L_k$ denotes the corresponding Lyapunov drift. After summing over k, all evolving terms telescope: the drift Δ_k, the queue increment $J_k - J_{k-1}$, and the quadratic auxiliary term involving γ_k. Consequently, the Lyapunov control layer contributes only boundary terms. Using the uniform queue bound from Appendix D Lemma 11, namely $J_k \leq \mathcal{O}(v)$, these boundary terms are further bounded by $\mathcal{O}\Big(\frac{H^4}{v} + \frac{v}{K}\Big)$. The Softmax approximation error together with the failure-event contribution gives rise to an additional structural term of order $\mathcal{O}\Big(\frac{H\log A}{v\eta} + Hp\Big)$. Therefore,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\hat{V}_1^k] \geq \theta^\star \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\check{W}_1^k] - \mathcal{O}\left(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\right).$$

For the second part, the statistical error is controlled by the linear-MDP concentration argument. Appendix C Lemma 7 implies that the cumulative optimistic reward and pessimistic cost errors satisfy

$$\sum_{k=1}^{K}\mathbb{E}[\hat{V}_1^k - V_1^{\pi^k}] \leq \mathcal{O}\Big(\sqrt{Md^3H^3T\,\iota^2}\Big), \qquad -\sum_{k=1}^{K}\mathbb{E}[\check{W}_1^k - W_1^{\pi^k}] \leq \mathcal{O}\Big(\sqrt{Md^3H^3T\,\iota^2}\Big),$$

where $\iota = \log(2MAdT/p)$. Substituting these bounds into the previous display yields

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[V_1^{\pi^k}] \geq \theta^\star \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}] - \mathcal{O}\left(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\right) - \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Since $W_1^{\pi^k} \geq Hc_{\min} > 0$, dividing both sides by the average cumulative cost gives

$$\mathrm{Gap}(K) \leq \mathcal{O}\left(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\right) + \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Finally, we set $v = H^2\sqrt{K}$ and $\eta = \sqrt{K}$. Although the structural bound above would already remain of order $K^{-1/2}$ for any $p \lesssim K^{-1/2}$, the cumulative estimation error bound in the Appendix C Lemma 7 is instantiated with $p = 1/K$ in order to control the bad-event contribution uniformly. With this choice, and using $\log A \leq \iota$ under $\iota = \log(2MdTA/p)$, we obtain

$$\mathcal{R}(K) = K\,\mathrm{Gap}(K) \leq \mathcal{O}\Big(\sqrt{Md^3H^3T\,\iota^2}\Big).$$

The proof highlights the main advantage of the Lyapunov formulation: the optimization dynamics across episodes are isolated into a telescoping structural term, while the statistical uncertainty is handled separately through standard linear value-estimation arguments. This decoupling is what enables the order-optimal $\sqrt{T}$ dependence in the final regret bound.

6 Numerical experiments

In this section, we empirically evaluate the effectiveness of COLD on online reward-to-cost ratio optimization in multi-task episodic linear MDPs. We primarily compare against RM^{2}VI (Xu et al., 2025), the most closely related baseline, as well as a natural adaptation of LSVI-UCB (Jin et al., 2020). Since LSVI-UCB is not originally designed for cost-aware ratio objectives, we adapt it following the surrogate approach in Xu et al. (2025): we use the same shared feature map and task-specific dynamics as in our setting, define an instantaneous surrogate reward r_h/c_h, and let the algorithm act greedily with respect to this surrogate reward. We design two sets of experiments and evaluate

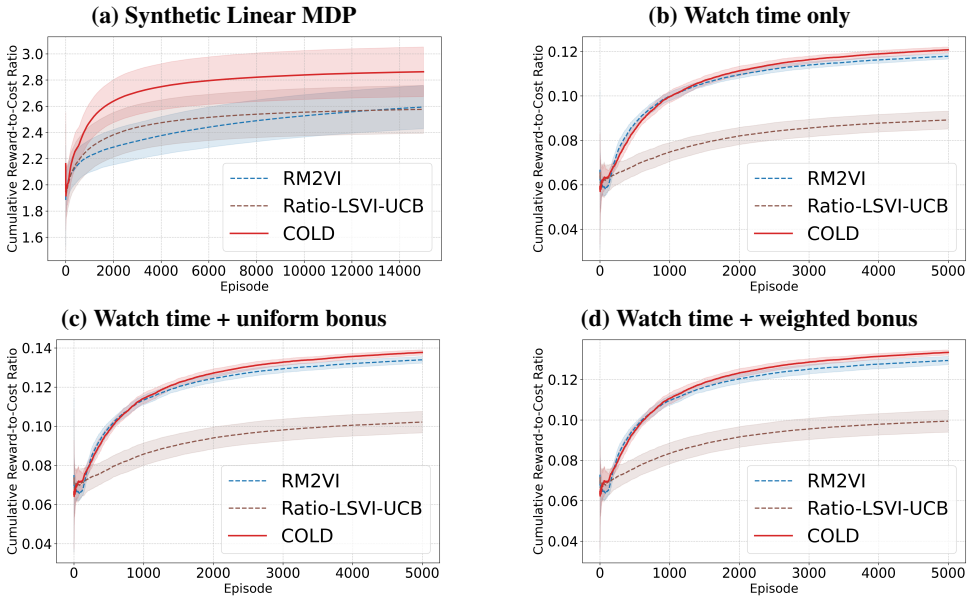


Figure 1: Comparison results on the synthetic and KuaiRand-based empirical environments.

performance by the cumulative reward-to-cost ratio $\sum_{k=1}^{K}\sum_{h=1}^{H} r_h / \sum_{k=1}^{K}\sum_{h=1}^{H} c_h$. All results are averaged over 20 trials, with light-shaded regions indicating one standard deviation.

Synthetic experiment. We construct a synthetic multi-task episodic linear MDP by sampling a shared feature map and task-dependent linear reward, cost, and transition parameters. In each episode, a task index is drawn at random, the initial state is fixed to $s_1 = 0$, and the interaction lasts for $H = 5$ stages. We set $M = 4$, $|\mathcal{S}| = 20$, $A = 10$, $d = 6$, and run each method for $K = 15000$ episodes. Construction details are provided in Appendix F. Figure 1a shows that COLD achieves the best cumulative reward-to-cost ratio throughout training. Ratio-LSVI-UCB outperforms RM^{2}VI in the early stage, illustrating the benefit of linear generalization, while RM^{2}VI eventually slightly surpasses it, suggesting that explicit ratio optimization is stronger in the long term than greedy value-based selection. By combining both advantages, COLD attains the best performance.

KuaiRand experiments. We evaluate all methods on an empirical multi-task finite-horizon MDP constructed from KuaiRand interaction logs (Gao et al., 2022); details are deferred to Appendix F. At a high level, task types are defined by user activity groups, states by discretized engagement signals, and actions by video content categories. The goal is to assess whether an algorithm can optimize *engagement efficiency*, rather than raw watchtime, by defining reward from normalized play time and cost from normalized video duration. Across all settings, the cost is fixed as normalized video duration, and we consider three reward specifications that capture progressively richer notions of user value. The first, *watch time only*, uses normalized watchtime as the sole reward signal. The second, *watch time plus uniform engagement bonus*, augments the same watchtime base with bonuses from active engagement events, namely `click`, `like`, `comment`, and `follow`, with the total bonus budget distributed uniformly across these events. The third, *watch time plus weighted engagement bonus*, uses the same watchtime base and bonus budget, but assigns larger coefficients to rarer engagement events, thereby placing greater emphasis on sparse yet high-value user feedback. For each reward specification, we construct an empirical multi-task finite-horizon MDP $(\nu^{\mathrm{emp}}, P^{\mathrm{emp}}, R^{\mathrm{emp}}, C^{\mathrm{emp}})$, where ν^{emp} denotes the empirical initial-state distribution, and compare the three algorithms in that environment. Each method is run for $K = 5000$ episodes with horizon $H = 5$.

Figures 1b, 1c, and 1d show a consistent pattern across all three reward settings. COLD achieves the highest cumulative reward-to-cost ratio in every case. RM^{2}VI is consistently the second-best method, remaining competitive but always below COLD. By contrast, Ratio-LSVI-UCB performs worst throughout, indicating that linear value generalization alone is insufficient here: without explicit long-term ratio control, greedy optimization of a ratio surrogate leads to clearly inferior performance.

Overall, these results show that the advantage of COLD is not limited to the synthetic linear MDP, but also carries over to the empirical KuaiRand-based environments.

7 Conclusion

We study online scheduling for multi-stage tasks and propose COLD, a Lyapunov-driven optimistic learning algorithm. By using a virtual queue and an auxiliary denominator-scale variable, COLD converts the long-term fractional objective into a queue-adjusted per-episode surrogate, separating ratio regulation from task-level value estimation. Together with double-optimistic least-squares value iteration and a softmax-smoothed policy update, this yields a sharper regret guarantee without explicit state-space dependence and only logarithmic action-space dependence. Experiments on synthetic and KuaiRand-based environments further validate the effectiveness of the proposed method.

Beyond the linear-MDP assumption. Our framework is modular in that its regret analysis separates the estimator-dependent statistical error from the Lyapunov optimization error, which does not directly depend on the specific function approximation architecture. This suggests that COLD can potentially be extended to richer function classes, including generalized-linear models and neural function approximation, under suitable estimation and approximation guarantees. We leave the development and formal analysis of these extensions to future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 62672321, Key Laboratory of Interdisciplinary Research of Computation and Economics (Shanghai University of Finance and Economics), Ministry of Education.

The first author gratefully acknowledges support from the NeurIPS 2026 Financial Assistance Program.

References

Aggarwal, V., Chandrasekaran, R. & Nair, K.P.K. (1977) Markov ratio decision processes. *Journal of Optimization Theory and Applications* **21**(1):27–37.

Agrawal, S. & Devanur, N.R. (2016) Linear contextual bandits with knapsacks. In D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3450–3458. Curran Associates, Inc.

Agrawal, S., Devanur, N.R. & Li, L. (2016) An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In V. Feldman, A. Rakhlin, and O. Shamir (eds.), *Proceedings of the 29th Annual Conference on Learning Theory*, **49**, pp. 4–18. PMLR.

Altman, E. (1999) *Constrained Markov Decision Processes*. Chapman & Hall/CRC.

Azar, M.G., Osband, I. & Munos, R. (2017) Minimax regret bounds for reinforcement learning. In D. Precup and Y.W. Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, **70**, pp. 263–272. PMLR.

Badanidiyuru, A., Langford, J. & Slivkins, A. (2014) Resourceful contextual bandits. In M.F. Balcan, V. Feldman, and C. Szepesvári (eds.), *Proceedings of the 27th Conference on Learning Theory*, **35**, pp. 1109–1134. PMLR.

Bhatnagar, S. (2010) An actor-critic algorithm with function approximation for discounted cost constrained Markov decision processes. *Systems & Control Letters* **59**(12):760–766.

Cai, Q., Yang, Z., Jin, C. & Wang, Z. (2020) Provably efficient exploration in policy optimization. In H. Daumé III and A. Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, **119**, pp. 1283–1294. PMLR.

Chzhen, E., Giraud, C., Li, Z. & Stoltz, G. (2023) Small total-cost constraints in contextual bandits with knapsacks, with application to fairness. In *Advances in Neural Information Processing Systems*, **36**, pp. 46737–46772.

Freund, D., Lykouris, T. & Weng, W. (2023) Quantifying the cost of learning in queueing systems. In *Advances in Neural Information Processing Systems*, **36**, pp. 6532–6544.

Gao, C., Li, S., Zhang, Y., Chen, J., Li, B., Lei, W., Jiang, P. & He, X. (2022) KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 3953–3957. ACM.

Guo, H. & Liu, X. (2024) Stochastic constrained contextual bandits via Lyapunov optimization based estimation to decision framework. In S. Agrawal and A. Roth (eds.), *Proceedings of the Thirty Seventh Conference on Learning Theory*, **247**, pp. 2204–2231. PMLR.

Guo, H. & Liu, X. (2025) On stochastic contextual bandits with knapsacks in small budget regime. In *The Thirteenth International Conference on Learning Representations*.

Han, Y., Zeng, J., Wang, Y., Xiang, Y. & Zhang, J. (2023) Optimal contextual bandits with knapsacks under realizability via regression oracles. In F. Ruiz, J. Dy, and J.-W. van de Meent (eds.), *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, **206**, pp. 5011–5035. PMLR.

Ho, C.-J. & Vaughan, J.W. (2012) Online task assignment in crowdsourcing markets. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp. 45–51. AAAI Press.

Hu, P., Chen, Y. & Huang, L. (2022) Nearly minimax optimal reinforcement learning with linear function approximation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, **162**, pp. 8971–9019. PMLR.

Huang, J., Golubchik, L. & Huang, L. (2024) When Lyapunov drift based queue scheduling meets adversarial bandit learning. *IEEE/ACM Transactions on Networking* **32**(4):3034–3044.

Jin, C., Yang, Z., Wang, Z. & Jordan, M.I. (2020) Provably efficient reinforcement learning with linear function approximation. In J. Abernethy and S. Agarwal (eds.), *Proceedings of the Thirty Third Conference on Learning Theory*, **125**, pp. 2137–2143. PMLR.

Johari, R., Kamble, V. & Kanoria, Y. (2021) Matching while learning. *Operations Research* **69**(2):655–681.

Kim, J. & Oh, M. (2024) Queueing matching bandits with preference feedback. In *Advances in Neural Information Processing Systems*, **37**, pp. 64649–64702.

Li, Z. & Stoltz, G. (2022) Contextual bandits with knapsacks for a conversion model. In *Advances in Neural Information Processing Systems*, **35**, pp. 35590–35602.

Liu, X., Li, B., Shi, P. & Ying, L. (2020) POND: Pessimistic-optimistic online dispatching. *arXiv preprint arXiv:2010.09995*.

Neely, M.J. (2011) Online fractional programming for Markov decision systems. In *49th Annual Allerton Conference on Communication, Control, and Computing*, pp. 353–360. IEEE.

Neely, M.J. (2013) Dynamic optimization and learning for renewal systems. *IEEE Transactions on Automatic Control* **58**(1):32–46.

Neely, M.J. (2021) Fast learning for renewal optimization in online task scheduling. *Journal of Machine Learning Research* **22**(279):1–44.

Neely, M.J. (2025) Adaptive optimization for stochastic renewal systems. *Mathematics* **13**(23):3850.

Ren, Z. & Krogh, B.H. (2005) Markov decision processes with fractional costs. *IEEE Transactions on Automatic Control* **50**(5):646–650.

Sinha, S. & Mondal, P. (2017) Semi-Markov decision processes with limiting ratio average rewards. *Journal of Mathematical Analysis and Applications* **455**(1):864–871.

Slivkins, A., Zhou, X., Sankararaman, K.A. & Foster, D.J. (2024) Contextual bandits with packing and covering constraints: A modular Lagrangian approach via regression. *Journal of Machine Learning Research* **25**(394):1–37.

Suttle, W., Zhang, K., Yang, Z., Liu, J. & Kraemer, D.N. (2021) Reinforcement learning for cost-aware Markov decision processes. In M. Meila and T. Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, **139**, pp. 9989–9999. PMLR.

Tanaka, T. (2017) A partially observable discrete time Markov decision process with a fractional discounted reward. *Journal of Information and Optimization Sciences* **38**(1):21–37.

Tanaka, T. (2019) Saddle–point type optimality criteria and dualities in fractional Markov decision processes with constraints. *Journal of Information and Optimization Sciences* **40**:957–972.

Wachi, A., Shen, X. & Sui, Y. (2024) A survey of constraint formulations in safe reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8262–8271. International Joint Conferences on Artificial Intelligence Organization.

Walton, N. & Xu, K. (2021) Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pp. 161–198. INFORMS.

Wei, H., Liu, X. & Ying, L. (2022) Triple-Q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In G. Camps-Valls, F.J.R. Ruiz, and I. Valera (eds.), *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, **151**, pp. 3274–3307. PMLR.

Xu, Y., Wang, S., Guo, H., Liu, X. & Shao, Z. (2024) Learning to schedule online tasks with bandit feedback. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1975–1983. IFAAMAS.

Xu, Y., Guo, H., Shao, Z. & Liu, X. (2025) No regret reinforcement learning algorithms for online scheduling with multi-stage tasks. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6767–6775. International Joint Conferences on Artificial Intelligence Organization.

Yang, L. & Wang, M. (2019) Sample-optimal parametric Q-learning using linearly additive features. In K. Chaudhuri and R. Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, **97**, pp. 6995–7004. PMLR.

Yang, Z., Srikant, R. & Ying, L. (2023) Learning while scheduling in multi-server systems with unknown statistics: MaxWeight with discounted UCB. In F. Ruiz, J. Dy, and J.-W. van de Meent (eds.), *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, **206**, pp. 4275–4312. PMLR.

A Roadmap of the proof

This appendix provides a roadmap for the proof of Theorem 1. We first present the main proof decomposition, and then summarize the organization of the subsequent technical appendices.

A.1 Proof decomposition

Let $\theta^\star := \sup_{(r,c)\in\bar{\Gamma}} \frac{r}{c}$, and let $(r^\star, c^\star) \in \bar{\Gamma}$ be an optimal pair such that $\theta^\star = \frac{r^\star}{c^\star}$. To prove the regret bound, we study the cumulative excess value

$$\sum_{k=1}^{K} \mathbb{E}\Big[V_1^{\pi^k}(m_k, s_1^k) - \theta^\star W_1^{\pi^k}(m_k, s_1^k)\Big].$$

For brevity, let $V_1^{\pi^k} := V_1^{\pi^k}(m_k, s_1^k)$, $W_1^{\pi^k} := W_1^{\pi^k}(m_k, s_1^k)$, $\hat{V}_1^k := \hat{V}_1^k(m_k, s_1^k)$, and $\check{W}_1^k := \check{W}_1^k(m_k, s_1^k)$.

By adding and subtracting the optimistic reward estimate $\hat{V}_1^k$ and the pessimistic cost estimate $\check{W}_1^k$, we obtain the decomposition

$$\sum_{k=1}^{K} \mathbb{E}\Big[V_1^{\pi^k} - \theta^\star W_1^{\pi^k}\Big] = \underbrace{\sum_{k=1}^{K} \mathbb{E}\Big[\hat{V}_1^k - \theta^\star \check{W}_1^k\Big]}_{\text{structural surrogate term}} - \underbrace{\sum_{k=1}^{K} \mathbb{E}\Big[\hat{V}_1^k - V_1^{\pi^k}\Big]}_{\text{reward estimation error}} - \theta^\star \underbrace{\sum_{k=1}^{K} \mathbb{E}\Big[W_1^{\pi^k} - \check{W}_1^k\Big]}_{\text{cost estimation error}}.$$

The proof therefore separates into two parts. The first part controls the queue-adjusted surrogate term

$$\sum_{k=1}^{K} \mathbb{E}\Big[\hat{V}_1^k - \theta^\star \check{W}_1^k\Big],$$

which is handled through the Softmax/optimism argument, the Lyapunov drift analysis, and a telescoping-sum argument. The second part controls the reward and cost replacement errors, which follow from the concentration analysis of the double-optimistic least-squares value iteration.

Finally, combining these two ingredients and using the positivity of the cumulative cost yield the ratio bound in Theorem 1.

A.2 Organization of the appendices

The technical appendices are organized as follows:

- Appendix B establishes basic structural properties of the multi-task linear MDP model;
- Appendix C develops the concentration argument and derives the cumulative estimation error bound;
- Appendix D controls the queue-adjusted surrogate term through the Softmax/optimism and Lyapunov analyses;
- Appendix E combines the preceding ingredients and completes the proof of Theorem 1.

In addition, Appendix F provides the construction details of the experimental environments used in our empirical evaluation, including both the synthetic linear MDP and the empirical KuaiRand MDP.

B Structural properties of multi-task linear MDPs

In this section, we establish the basic structural properties of the multi-task linear MDP model used throughout the appendix. In particular, we show that the reward and cost action-value functions admit task-specific linear representations under the shared feature map $\phi(s, a)$, and we derive uniform bounds for both the true and estimated weight vectors. These results serve as the foundation for the subsequent concentration and structural analyses.

Proof of Proposition 1. We prove the claim for the reward action-value function; the cost case is identical. By the Bellman equation,

$$Q_h^\pi(m, s, a) = r_h(m, s, a) + (\mathbb{P}_h V_{h+1}^\pi)(m, s, a).$$

Using Assumption 1,

$$r_h(m, s, a) = \langle \phi(s, a), \theta_{r,h}(m) \rangle$$

and

$$(\mathbb{P}_h V_{h+1}^\pi)(m, s, a) = \int_{\mathcal{S}} V_{h+1}^\pi(m, s') \, \langle \phi(s, a), d\mu_h(m, s') \rangle.$$

Hence

$$Q_h^\pi(m, s, a) = \left\langle \phi(s, a), \theta_{r,h}(m) + \int_{\mathcal{S}} V_{h+1}^\pi(m, s') \, d\mu_h(m, s') \right\rangle.$$

Define

$$w^{\pi}_{r,h}(m) := \theta_{r,h}(m) + \int_{\mathcal{S}} V^{\pi}_{h+1}(m, s')\, d\mu_h(m, s').$$

Then

$$Q^{\pi}_h(m, s, a) = \langle \phi(s, a), w^{\pi}_{r,h}(m) \rangle.$$

Similarly, defining

$$w^{\pi}_{c,h}(m) := \theta_{c,h}(m) + \int_{\mathcal{S}} W^{\pi}_{h+1}(m, s')\, d\mu_h(m, s'),$$

we obtain

$$C^{\pi}_h(m, s, a) = \langle \phi(s, a), w^{\pi}_{c,h}(m) \rangle.$$

□

Lemma 1 (Boundedness of true weights)**.** *Let $w^{\pi}_{r,h}(m)$ and $w^{\pi}_{c,h}(m)$ be the vectors defined in Proposition 1. Under Assumptions 1 and 2, for all $m \in \mathcal{M}$ and $h \in [H]$,*

$$\|w^{\pi}_{r,h}(m)\|_2 \le (1 + H_{\max})\sqrt{d}, \qquad \|w^{\pi}_{c,h}(m)\|_2 \le (1 + H_{\max})\sqrt{d}.$$

Proof. We prove the bound for $w^{\pi}_{r,h}(m)$; the proof for $w^{\pi}_{c,h}(m)$ is the same. By definition,

$$w^{\pi}_{r,h}(m) = \theta_{r,h}(m) + \int_{\mathcal{S}} V^{\pi}_{h+1}(m, s')\, d\mu_h(m, s').$$

Applying the triangle inequality gives

$$\|w^{\pi}_{r,h}(m)\|_2 \le \|\theta_{r,h}(m)\|_2 + \left\| \int_{\mathcal{S}} V^{\pi}_{h+1}(m, s')\, d\mu_h(m, s') \right\|_2.$$

By Assumption 2, $\|\theta_{r,h}(m)\|_2 \le \sqrt{d}$. Moreover, since the cumulative value functions are upper-bounded by $H_{\max}$, we have

$$\left\| \int_{\mathcal{S}} V^{\pi}_{h+1}(m, s')\, d\mu_h(m, s') \right\|_2 \le H_{\max}\, \|\mu_h(m, \mathcal{S})\|_2 \le H_{\max}\sqrt{d}.$$

Therefore,

$$\|w^{\pi}_{r,h}(m)\|_2 \le (1 + H_{\max})\sqrt{d}.$$

The same argument yields

$$\|w^{\pi}_{c,h}(m)\|_2 \le (1 + H_{\max})\sqrt{d}.$$

□

Lemma 2 (Boundedness of estimated weights)**.** *For any $(m, h, k) \in \mathcal{M} \times [H] \times [K]$, the ridge-regression estimates in COLD satisfy*

$$\|w^{k}_{r,h}(m)\|_2 \le 2H_{\max}\sqrt{\frac{kd}{\lambda}}, \qquad \|w^{k}_{c,h}(m)\|_2 \le 2H_{\max}\sqrt{\frac{kd}{\lambda}}.$$

Proof. We prove the claim for $w^{k}_{r,h}(m)$; the argument for $w^{k}_{c,h}(m)$ is identical. Recall that

$$w^{k}_{r,h}(m) = \big(\Lambda^{k}_h(m)\big)^{-1} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, \phi^{\tau}_h \big[r^{\tau}_h + \hat{V}^{k}_{h+1}(m, s^{\tau}_{h+1})\big],$$

where $\phi^{\tau}_h := \phi(s^{\tau}_h, a^{\tau}_h)$.

Fix any $u \in \mathbb{R}^d$. Then

$$\begin{aligned} |u^{\top} w^{k}_{r,h}(m)| &= \left| u^{\top} \big(\Lambda^{k}_h(m)\big)^{-1} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, \phi^{\tau}_h \big[r^{\tau}_h + \hat{V}^{k}_{h+1}(m, s^{\tau}_{h+1})\big] \right| \\ &\le \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\} \left| u^{\top} \big(\Lambda^{k}_h(m)\big)^{-1} \phi^{\tau}_h \right| \cdot \left| r^{\tau}_h + \hat{V}^{k}_{h+1}(m, s^{\tau}_{h+1}) \right|. \end{aligned}$$

Since both the single-step reward and the estimated continuation value are upper-bounded by $H_{\max}$, we have

$$\left| r_h^\tau + \hat{V}_{h+1}^k(m, s_{h+1}^\tau) \right| \le 2H_{\max}.$$

Hence

$$|u^\top w_{r,h}^k(m)| \le 2H_{\max} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\} \left| u^\top \big(\Lambda_h^k(m)\big)^{-1} \phi_h^\tau \right|.$$

Applying Cauchy–Schwarz,

$$|u^\top w_{r,h}^k(m)| \le 2H_{\max} \sqrt{\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, u^\top \big(\Lambda_h^k(m)\big)^{-1} u} \sqrt{\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, (\phi_h^\tau)^\top \big(\Lambda_h^k(m)\big)^{-1} \phi_h^\tau}.$$

For the first factor, since

$$\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\} \le k \qquad \text{and} \qquad \big(\Lambda_h^k(m)\big)^{-1} \preceq \lambda^{-1} I,$$

we obtain

$$\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, u^\top \big(\Lambda_h^k(m)\big)^{-1} u \le \frac{k}{\lambda} \|u\|_2^2.$$

For the second factor, write

$$\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, \phi_h^\tau (\phi_h^\tau)^\top = U \operatorname{diag}(\lambda_1, \dots, \lambda_d) U^\top.$$

Then

$$\Lambda_h^k(m) = U \operatorname{diag}(\lambda_1 + \lambda, \dots, \lambda_d + \lambda) U^\top,$$

and therefore

$$\begin{aligned} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, (\phi_h^\tau)^\top \big(\Lambda_h^k(m)\big)^{-1} \phi_h^\tau &= \operatorname{tr}\left[\big(\Lambda_h^k(m)\big)^{-1} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\, \phi_h^\tau (\phi_h^\tau)^\top \right] \\ &= \operatorname{tr}\left(\operatorname{diag}\left(\frac{\lambda_1}{\lambda_1 + \lambda}, \dots, \frac{\lambda_d}{\lambda_d + \lambda} \right) \right) \le d. \end{aligned}$$

Combining the two bounds yields

$$|u^\top w_{r,h}^k(m)| \le 2H_{\max} \|u\|_2 \sqrt{\frac{kd}{\lambda}}.$$

Taking the supremum over all $\|u\|_2 = 1$ gives

$$\|w_{r,h}^k(m)\|_2 \le 2H_{\max} \sqrt{\frac{kd}{\lambda}}.$$

The same argument applies to $w_{c,h}^k(m)$. □

C Statistical concentration and cumulative estimation error

In this section, we control the statistical error induced by the double-optimistic least-squares value iteration. We first define the relevant value-function classes induced by the COLD parameterization, then bound their covering numbers, establish a uniform concentration event for Bellman errors, and finally derive both one-step and cumulative estimation bounds. These results will later be used to control the reward and cost replacement errors in the proof of Theorem 1.

C.1 Value-function classes induced by COLD

Fix a task type $m \in \mathcal{M}$ and a step $h \in [H]$. For this fixed pair (m, h), define the admissible parameter tuple

$$\omega = (w_r, w_c, \Sigma, J),$$

where $w_r, w_c \in \mathbb{R}^d$, $\Sigma = \beta^2 \Lambda^{-1} \in \mathbb{R}^{d\times d}$, and $J \in [0, J_{\max}]$. Given such a tuple, define

$$\hat{Q}_\omega(s,a) := \left[\langle w_r, \phi(s,a)\rangle + \sqrt{\phi(s,a)^\top \Sigma \phi(s,a)}\right]_0^{(H-h+1)r_{\max}},$$

$$\check{C}_\omega(s,a) := \left[\langle w_c, \phi(s,a)\rangle - \sqrt{\phi(s,a)^\top \Sigma \phi(s,a)}\right]_{(H-h+1)c_{\min}}^{(H-h+1)c_{\max}}.$$

The associated Softmax policy is

$$\pi_\omega(a \mid m, s) \propto \exp\Big(\eta\big[v\hat{Q}_\omega(m,s,a) - J\check{C}_\omega(m,s,a)\big]\Big), \tag{14}$$

and the induced value functions are

$$\hat{V}_\omega(m,s) = \sum_{a\in\mathcal{A}} \pi_\omega(a \mid m, s)\hat{Q}_\omega(m,s,a), \qquad \check{W}_\omega(m,s) = \sum_{a\in\mathcal{A}} \pi_\omega(a \mid m, s)\check{C}_\omega(m,s,a). \tag{15}$$

For each fixed (m, h), let $\mathcal{V}_{m,h}$ be the class of all reward value functions $\hat{V}_\omega(m, \cdot)$ induced by admissible tuples ω, and let $\mathcal{W}_{m,h}$ be the class of all cost value functions $\check{W}_\omega(m, \cdot)$ induced by admissible tuples ω. We equip these classes with the supremum distance

$$\mathrm{dist}_m(U_1, U_2) := \sup_{s\in\mathcal{S}} |U_1(m,s) - U_2(m,s)|.$$

C.2 Covering number of the value-function classes

To control the concentration uniformly over all possible value functions generated by COLD, we first bound the covering number of the classes $\mathcal{V}$ and $\mathcal{W}$.

Lemma 3 (Covering number of the value-function classes)**.** *For every fixed (m, h), let N_ϵ denote the ϵ-covering number of $\mathcal{V}_{m,h} \cup \mathcal{W}_{m,h}$ with respect to the distance $\mathrm{dist}_m(\cdot,\cdot)$ defined above. Then, for any $\epsilon > 0$,*

$$\begin{aligned}\log N_\varepsilon \le\ & d\log\left(1 + \frac{8L_w e_r}{\varepsilon}\right) + d\log\left(1 + \frac{8L_w e_c}{\varepsilon}\right) \\ & + d^2\log\left(1 + \frac{32L_\Sigma (e_r + e_c)^2}{\varepsilon^2}\right) + \log\left(1 + \frac{4J_{\max} e_J}{\varepsilon}\right),\end{aligned} \tag{16}$$

where $e_r := 1 + 2\eta v H_{\max}$, $e_c := 1 + 2\eta H_{\max} J_{\max}$, $e_J := 2\eta H_{\max}^2$, $L_w := 2H_{\max}\sqrt{Kd/\lambda}$, and $L_\Sigma := \sqrt{d}\,\beta^2/\lambda$.

Proof. Fix two parameter tuples $\omega_1 = (w_{r,1}, w_{c,1}, \Sigma_1, J_1)$ and $\omega_2 = (w_{r,2}, w_{c,2}, \Sigma_2, J_2)$. Write $\hat{Q}_i := \hat{Q}_{\omega_i}$, $\check{C}_i := \check{C}_{\omega_i}$, $\pi_i := \pi_{\omega_i}$, $\hat{V}_i := \hat{V}_{\omega_i}$, and $\check{W}_i := \check{W}_{\omega_i}$ for $i \in \{1, 2\}$, and define $\|g\|_\infty := \sup_{m,s,a} |g(m,s,a)|$ for functions on $\mathcal{M} \times \mathcal{S} \times \mathcal{A}$. We first compare the reward value functions. Since $\hat{V}_i(m,s) = \sum_{a\in\mathcal{A}} \pi_i(a \mid m, s)\hat{Q}_i(m,s,a)$, we have

$$\begin{aligned}|\hat{V}_1(m,s) - \hat{V}_2(m,s)| &\le \left|\sum_a \pi_1(a \mid m,s)\big(\hat{Q}_1 - \hat{Q}_2\big)(m,s,a)\right| \\ &\quad + \left|\sum_a \big(\pi_1 - \pi_2\big)(a \mid m,s)\hat{Q}_2(m,s,a)\right| \\ &\le \|\hat{Q}_1 - \hat{Q}_2\|_\infty + \|\hat{Q}_2\|_\infty \|\pi_1(\cdot \mid m,s) - \pi_2(\cdot \mid m,s)\|_1.\end{aligned}$$

Because $\hat{Q}_\omega$ is clipped above by $(H-h+1)r_{\max} \le H_{\max}$, we have $\|\hat{Q}_2\|_\infty \le H_{\max}$. Next define the logits $Z_i(m,s,a) := \eta\big(v\hat{Q}_i(m,s,a) - J_i\check{C}_i(m,s,a)\big)$. By the standard 2-Lipschitz property of the Softmax map from ℓ_∞ to ℓ_1,

$$\|\pi_1(\cdot \mid m,s) - \pi_2(\cdot \mid m,s)\|_1 \le 2\|Z_1(m,s,\cdot) - Z_2(m,s,\cdot)\|_\infty.$$

Moreover,

$$\begin{aligned}\|Z_1(m,s,\cdot) - Z_2(m,s,\cdot)\|_\infty &\le \eta v\|\hat{Q}_1 - \hat{Q}_2\|_\infty + \eta\|J_1\check{C}_1 - J_2\check{C}_2\|_\infty \\ &\le \eta v\|\hat{Q}_1 - \hat{Q}_2\|_\infty + \eta J_{\max}\|\check{C}_1 - \check{C}_2\|_\infty + \eta H_{\max}|J_1 - J_2|,\end{aligned}$$

where we used $\|\check{C}_2\|_\infty \le H_{\max}$ and $J_i \in [0, J_{\max}]$. Combining the last two displays yields

$$\operatorname{dist}(\hat{V}_1, \hat{V}_2) \le e_r\|\hat{Q}_1 - \hat{Q}_2\|_\infty + e_c\|\check{C}_1 - \check{C}_2\|_\infty + e_J|J_1 - J_2|.$$

where $e_r := 1 + 2\eta v H_{\max}$, $e_c := 1 + 2\eta H_{\max} J_{\max}$, and $e_J := 2\eta H_{\max}^2$.

The same argument applied to $\check{W}_i(m,s) = \sum_a \pi_i(a \mid m,s)\check{C}_i(m,s,a)$ gives

$$\operatorname{dist}(\check{W}_1, \check{W}_2) \le e_r\|\hat{Q}_1 - \hat{Q}_2\|_\infty + e_c\|\check{C}_1 - \check{C}_2\|_\infty + e_J|J_1 - J_2|.$$

We now bound $\hat{Q}_1 - \hat{Q}_2$ and $\check{C}_1 - \check{C}_2$ in terms of the parameter tuples. Since clipping is 1-Lipschitz,

$$|\hat{Q}_1(m,s,a) - \hat{Q}_2(m,s,a)| \le |\langle\phi(s,a), w_{r,1} - w_{r,2}\rangle| + \left|\sqrt{\phi(s,a)^\top\Sigma_1\phi(s,a)} - \sqrt{\phi(s,a)^\top\Sigma_2\phi(s,a)}\right|.$$

Using $\|\phi(s,a)\|_2 \le 1$, Cauchy–Schwarz, and $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$ for $x, y \ge 0$, we obtain

$$|\hat{Q}_1(m,s,a) - \hat{Q}_2(m,s,a)| \le \|w_{r,1} - w_{r,2}\|_2 + \sqrt{\|\Sigma_1 - \Sigma_2\|_F}.$$

Hence $\|\hat{Q}_1 - \hat{Q}_2\|_\infty \le \|w_{r,1} - w_{r,2}\|_2 + \sqrt{\|\Sigma_1 - \Sigma_2\|_F}$. Similarly,

$$\|\check{C}_1 - \check{C}_2\|_\infty \le \|w_{c,1} - w_{c,2}\|_2 + \sqrt{\|\Sigma_1 - \Sigma_2\|_F}.$$

Therefore both $\operatorname{dist}(\hat{V}_1, \hat{V}_2)$ and $\operatorname{dist}(\check{W}_1, \check{W}_2)$ are bounded by

$$e_r\|w_{r,1} - w_{r,2}\|_2 + e_c\|w_{c,1} - w_{c,2}\|_2 + (e_r + e_c)\sqrt{\|\Sigma_1 - \Sigma_2\|_F} + e_J|J_1 - J_2|.$$

It thus suffices to construct separate covers satisfying

$$\|w_{r,1} - w_{r,2}\|_2 \le \frac{\varepsilon}{4e_r},\ \|w_{c,1} - w_{c,2}\|_2 \le \frac{\varepsilon}{4e_c},\ \sqrt{\|\Sigma_1 - \Sigma_2\|_F} \le \frac{\varepsilon}{4(e_r + e_c)},\ |J_1 - J_2| \le \frac{\varepsilon}{4e_J}.$$

By Lemma 2, for every episode $k \le K$, both w_r and w_c lie in an ℓ_2-ball of radius $L_w := 2H_{\max}\sqrt{Kd/\lambda}$. A standard Euclidean covering bound yields

$$|C_{w_r}| \le \left(1 + \frac{8L_w e_r}{\varepsilon}\right)^d, \qquad |C_{w_c}| \le \left(1 + \frac{8L_w e_c}{\varepsilon}\right)^d.$$

For the matrix parameter, since $\Sigma = \beta^2\Lambda^{-1}$, we have $\|\Sigma\|_F \le \sqrt{d}\,\beta^2/\lambda =: L_\Sigma$. To ensure $\sqrt{\|\Sigma_1 - \Sigma_2\|_F} \le \frac{\varepsilon}{4(e_r+e_c)}$., it suffices to build a Frobenius-norm cover with radius $\delta_\Sigma = \frac{\varepsilon^2}{16(e_r+e_c)^2}$., which gives

$$|C_\Sigma| \le \left(1 + \frac{32L_\Sigma(e_r + e_c)^2}{\varepsilon^2}\right)^{d^2}.$$

Finally, since $J \in [0, J_{\max}]$, an interval cover with radius $\varepsilon/(4e_J)$ has size at most

$$|C_J| \le 1 + \frac{4J_{\max}e_J}{\varepsilon}.$$

Multiplying the four covering numbers and taking logarithms yields

$$\begin{aligned}\log N_\varepsilon \le\ & d\log\left(1 + \frac{8L_w e_r}{\varepsilon}\right) + d\log\left(1 + \frac{8L_w e_c}{\varepsilon}\right) \\ &+ d^2\log\left(1 + \frac{32L_\Sigma(e_r + e_c)^2}{\varepsilon^2}\right) + \log\left(1 + \frac{4J_{\max}e_J}{\varepsilon}\right).\end{aligned}$$

which is exactly (16). □

Lemma 4 (Uniform concentration for value functions)**.** *Let N_ϵ denote the uniform upper bound on the ϵ-covering number of $\mathcal{V}_{m,h} \cup \mathcal{W}_{m,h}$ over all $(m,h) \in \mathcal{M} \times [H]$, with respect to $\mathrm{dist}_m(\cdot,\cdot)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds simultaneously for all episodes $k \in [K]$, steps $h \in [H]$, task types $m \in \mathcal{M}$, and all $\widetilde{U} \in \mathcal{V}_{m,h+1} \cup \mathcal{W}_{m,h+1}$:*

$$\begin{aligned}\left\| \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau \Big[\widetilde{U}(m, s_{h+1}^\tau) - (\mathbb{P}_h \widetilde{U})(m, s_h^\tau, a_h^\tau)\Big] \right\|^2_{(\Lambda_h^k(m))^{-1}} \\ \leq\ 4H_{\max}^2 \left[\frac{d}{2}\log\Big(\frac{\lambda+k}{\lambda}\Big) + \log\Big(\frac{MHN_\varepsilon}{\delta}\Big)\right] \quad + \frac{8k^2\varepsilon^2}{\lambda}.\end{aligned} \tag{17}$$

Proof. Let $\mathcal{N}_\varepsilon$ be an ε-net of $\mathcal{V} \cup \mathcal{W}$ with respect to $\mathrm{dist}(U_1, U_2) := \sup_{m\in\mathcal{M},\ s\in\mathcal{S}} |U_1(m,s) - U_2(m,s)|$, and let $|\mathcal{N}_\varepsilon| = N_\varepsilon$. Fix arbitrary $\widetilde{U} \in \mathcal{V} \cup \mathcal{W}$.

For fixed $(m,h) \in \mathcal{M} \times [H]$ and $\bar{U} \in \mathcal{N}_\varepsilon$, define $x_\tau := \mathbf{1}\{m_\tau = m\}\phi_h^\tau$ and $\chi_\tau^{\bar{U}} := \bar{U}(m, s_{h+1}^\tau) - (\mathbb{P}_h \bar{U})(m, s_h^\tau, a_h^\tau)$ for $\tau = 1, \ldots, k-1$. Then x_τ is predictable with respect to the natural filtration, $\mathbb{E}[\chi_\tau^{\bar{U}} \mid \mathcal{F}_{\tau-1}] = 0$, and $|\chi_\tau^{\bar{U}}| \leq H_{\max}$ since $\bar{U}(m,\cdot) \in [0, H_{\max}]$. Hence $\chi_\tau^{\bar{U}}$ is conditionally $H_{\max}$-sub-Gaussian.

Applying Theorem D.3 of Jin et al. (2020) to the predictable covariates $\{x_\tau\}$ and the scalar noise sequence $\{\chi_\tau^{\bar{U}}\}$ yields that, for any $\delta' > 0$, with probability at least $1-\delta'$, simultaneously for all $k \in [K]$,

$$\left\| \sum_{\tau=1}^{k-1} x_\tau \chi_\tau^{\bar{U}} \right\|^2_{(\Lambda_h^k(m))^{-1}} \leq 2H_{\max}^2 \left[\frac{d}{2}\log\Big(\frac{\lambda+k}{\lambda}\Big) + \log\frac{1}{\delta'}\right]. \tag{18}$$

Taking a union bound over all triples $(m, h, \bar{U}) \in \mathcal{M} \times [H] \times \mathcal{N}_\varepsilon$ and setting $\delta' = \delta/(MHN_\varepsilon)$, we conclude that with probability at least $1-\delta$,

$$\begin{aligned}\left\| \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau \Big[\bar{U}(m, s_{h+1}^\tau) - (\mathbb{P}_h \bar{U})(m, s_h^\tau, a_h^\tau)\Big] \right\|^2_{(\Lambda_h^k(m))^{-1}} \\ \leq 2H_{\max}^2 \left[\frac{d}{2}\log\Big(\frac{\lambda+k}{\lambda}\Big) + \log\Big(\frac{MHN_\varepsilon}{\delta}\Big)\right].\end{aligned} \tag{19}$$

Choose $\bar{U} \in \mathcal{N}_\varepsilon$ such that $\mathrm{dist}(\widetilde{U}, \bar{U}) \leq \varepsilon$, and define

$$\zeta_\tau := \mathbf{1}\{m_\tau = m\}\phi_h^\tau \Big[(\widetilde{U} - \bar{U})(m, s_{h+1}^\tau) - (\mathbb{P}_h(\widetilde{U} - \bar{U}))(m, s_h^\tau, a_h^\tau)\Big].$$

Then

$$\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau \Big[\widetilde{U}(m, s_{h+1}^\tau) - (\mathbb{P}_h \widetilde{U})(m, s_h^\tau, a_h^\tau)\Big] = \sum_{\tau=1}^{k-1} x_\tau \chi_\tau^{\bar{U}} + \sum_{\tau=1}^{k-1} \zeta_\tau.$$

Using $\|u+v\|^2_{\Lambda^{-1}} \leq 2\|u\|^2_{\Lambda^{-1}} + 2\|v\|^2_{\Lambda^{-1}}$, we obtain

$$\begin{aligned}\left\| \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau \Big[\widetilde{U}(m, s_{h+1}^\tau) - (\mathbb{P}_h \widetilde{U})(m, s_h^\tau, a_h^\tau)\Big] \right\|^2_{(\Lambda_h^k(m))^{-1}} \\ \leq\ 2\left\| \sum_{\tau=1}^{k-1} x_\tau \chi_\tau^{\bar{U}} \right\|^2_{(\Lambda_h^k(m))^{-1}} + 2\left\| \sum_{\tau=1}^{k-1} \zeta_\tau \right\|^2_{(\Lambda_h^k(m))^{-1}}.\end{aligned} \tag{20}$$

Since $\mathrm{dist}(\widetilde{U}, \bar{U}) \leq \varepsilon$, we have

$$2\left\| \sum_{\tau=1}^{k-1} \zeta_\tau \right\|^2_{(\Lambda_h^k(m))^{-1}} \leq \frac{8k^2\varepsilon^2}{\lambda}.$$

Combining this with (19) and (20) yields (17). □

Lemma 5 (Uniform concentration of Bellman error)**.** *Assume $\lambda = 1$, choose $v = H^2\sqrt{K}$, $\eta = \sqrt{K}$ and let $\beta = C_\beta dH_{\max}\sqrt{\iota}$. Then there exists an absolute constant $C_1 > 0$, independent of C_β, such that with probability at least $1 - p/2$, for all $(k, m, h) \in [K] \times \mathcal{M} \times [H]$ and all $\widetilde{U} \in \mathcal{V}_{m,h+1} \cup \mathcal{W}_{m,h+1}$,*

$$\left\|\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau\Big[\widetilde{U}_{h+1}(m, s_{h+1}^\tau) - (\mathbb{P}_h\widetilde{U}_{h+1})(m, s_h^\tau, a_h^\tau)\Big]\right\|_{(\Lambda_h^k(m))^{-1}} \le C_1 dH_{\max}\sqrt{X},$$

where $X = \log\big(2(C_\beta + 1)MdT/p\big)$. Denote this event by $\mathcal{E}$.

Proof. Fix $\delta = p/2$. By Lemma 4 and Lemma 3, with probability at least $1 - p/2$, for all $(k, m, h) \in [K] \times \mathcal{M} \times [H]$ and all $\widetilde{U} \in \mathcal{V} \cup \mathcal{W}$,

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau\Big[\widetilde{U}(m, s_{h+1}^\tau) - (\mathbb{P}_h\widetilde{U})(m, s_h^\tau, a_h^\tau)\Big]\right\|^2_{(\Lambda_h^k(m))^{-1}} \\
&\quad\le 4H_{\max}^2\left[\frac{d}{2}\log\left(\frac{\lambda + k}{\lambda}\right) + d\log\left(1 + \frac{8L_w e_r}{\varepsilon}\right) + d\log\left(1 + \frac{8L_w e_c}{\varepsilon}\right)\right. \\
&\qquad\left. + d^2\log\left(1 + \frac{32L_\Sigma(e_r + e_c)^2}{\varepsilon^2}\right) + \log\left(1 + \frac{4J_{\max}e_J}{\varepsilon}\right) + \log\left(\frac{2MH}{p}\right)\right] \\
&\qquad + \frac{8k^2\varepsilon^2}{\lambda}.
\end{aligned} \tag{21}$$

Now choose $\lambda = 1$ and $\varepsilon_k = dH_{\max}/K$. Then the approximation term becomes $8d^2H_{\max}^2$. Under the final parameter choice $v = H^2\sqrt{K}$ and $\eta = \sqrt{K}$, Lemma 11 implies $J_{\max} \lesssim \sqrt{K}$, and hence $e_r \lesssim K$, $e_c \lesssim K$, and $e_J \lesssim \sqrt{K}$. Therefore, there exists an absolute constant $C_1 > 0$, independent of C_β, such that

$$\left\|\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau\Big[\widetilde{U}_{h+1}(m, s_{h+1}^\tau) - (\mathbb{P}_h\widetilde{U}_{h+1})(m, s_h^\tau, a_h^\tau)\Big]\right\|^2_{(\Lambda_h^k(m))^{-1}} \le C_1^2 d^2 H_{\max}^2 X.$$

Taking square roots proves the lemma. □

C.3 One-step estimation error

On the confidence event $\mathcal{E}$, we can now compare the optimistic and pessimistic Bellman estimates with the true action-value functions.

Lemma 6 (One-step estimation error for reward and cost)**.** *There exists an absolute constant $C_\beta > 0$ such that, if $\beta = C_\beta dH_{\max}\sqrt{\iota}$ with $\iota := \log\Big(\frac{2MAdT}{p}\Big)$, then on the event $\mathcal{E}$ defined in Lemma 5, for any fixed policy π, all $(m, s, a, h, k) \in \mathcal{M} \times \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we have*

$$\langle\phi(s, a), w_{r,h}^k(m)\rangle - Q_h^\pi(m, s, a) = (\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m, s, a) + \Delta_{r,h}^k(m, s, a),$$

$$\langle\phi(s, a), w_{c,h}^k(m)\rangle - C_h^\pi(m, s, a) = (\mathbb{P}_h(\check{W}_{h+1}^k - W_{h+1}^\pi))(m, s, a) + \Delta_{c,h}^k(m, s, a),$$

where

$$\max\{|\Delta_{r,h}^k(m, s, a)|,\ |\Delta_{c,h}^k(m, s, a)|\} \le \beta\sqrt{\phi(s, a)^\top(\Lambda_h^k(m))^{-1}\phi(s, a)}.$$

Proof. We prove the reward statement; the cost statement follows identically. By Proposition 1, the true reward action-value function is linear:

$$Q_h^\pi(m, s, a) = \langle\phi(s, a), w_{r,h}^\pi(m)\rangle.$$

Using the Bellman equation and the ridge-regression update, we write

$$w_{r,h}^k(m) - w_{r,h}^\pi(m) = q_1 + q_2 + q_3,$$

where

$$q_1 := -\lambda(\Lambda_h^k(m))^{-1} w_{r,h}^\pi(m),$$

$$q_2 := (\Lambda_h^k(m))^{-1} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau \big[\hat{V}_{h+1}^k(m, s_{h+1}^\tau) - (\mathbb{P}_h \hat{V}_{h+1}^k)(m, s_h^\tau, a_h^\tau)\big],$$

and

$$q_3 := (\Lambda_h^k(m))^{-1} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau (\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m, s_h^\tau, a_h^\tau).$$

We now bound the inner product of $\phi(s,a)$ with each term separately. For the first term, by Cauchy–Schwarz, $(\Lambda_h^k(m))^{-1} \preceq \lambda^{-1} I$, and Lemma 1,

$$\begin{aligned} |\langle \phi(s,a), q_1 \rangle| &= \lambda \big| \langle (\Lambda_h^k(m))^{-1/2} \phi(s,a), (\Lambda_h^k(m))^{-1/2} w_{r,h}^\pi(m) \rangle \big| \\ &\le 2H_{\max}\sqrt{d\lambda}\, \sqrt{\phi(s,a)^\top (\Lambda_h^k(m))^{-1} \phi(s,a)}. \end{aligned}$$

For the second term, on the event $\mathcal{E}$ from Lemma 5,

$$\begin{aligned} &|\langle \phi(s,a), q_2 \rangle| \\ &\le \|\phi(s,a)\|_{(\Lambda_h^k(m))^{-1}} \left\| \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau \big[\hat{V}_{h+1}^k(m, s_{h+1}^\tau) - (\mathbb{P}_h \hat{V}_{h+1}^k)(m, s_h^\tau, a_h^\tau)\big] \right\|_{(\Lambda_h^k(m))^{-1}} \\ &\le C_1 d H_{\max} \sqrt{X}\, \|\phi(s,a)\|_{(\Lambda_h^k(m))^{-1}}. \end{aligned}$$

where $X = \log\big(2(C_\beta + 1)MdT/p\big)$ and C_1 an absolute constant independent of C_β.

For the third term, let

$$\mu_{r,h}^k(m) := \int_{\mathcal{S}} \big(\hat{V}_{h+1}^k(m, s') - V_{h+1}^\pi(m, s')\big)\, d\mu_h(m, s').$$

Then

$$(\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m, s, a) = \langle \phi(s,a), \mu_{r,h}^k(m) \rangle.$$

Hence

$$\langle \phi(s,a), q_3 \rangle = \Big\langle \phi(s,a), (\Lambda_h^k(m))^{-1} \sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau (\phi_h^\tau)^\top \mu_{r,h}^k(m) \Big\rangle.$$

Using $\sum_{\tau=1}^{k-1} \mathbf{1}\{m_\tau = m\}\phi_h^\tau (\phi_h^\tau)^\top = \Lambda_h^k(m) - \lambda I$, we decompose

$$\langle \phi(s,a), q_3 \rangle = p_1 + p_2,$$

where

$$p_1 := \langle \phi(s,a), \mu_{r,h}^k(m) \rangle = (\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m, s, a),$$

and

$$p_2 := -\lambda \langle \phi(s,a), (\Lambda_h^k(m))^{-1} \mu_{r,h}^k(m) \rangle.$$

Since both $\hat{V}_{h+1}^k$ and V_{h+1}^π are bounded by $H_{\max}$, we have $\|\mu_{r,h}^k(m)\|_2 \le H_{\max}\sqrt{d}$, and therefore

$$|p_2| \le H_{\max}\sqrt{d\lambda}\, \sqrt{\phi(s,a)^\top (\Lambda_h^k(m))^{-1} \phi(s,a)}.$$

Combining the above bounds and using

$$\langle \phi(s,a), w_{r,h}^k(m) \rangle - Q_h^\pi(m, s, a) = \langle \phi(s,a), q_1 + q_2 + q_3 \rangle,$$

we obtain

$$\begin{aligned} &\big| \langle \phi(s,a), w_{r,h}^k(m) \rangle - Q_h^\pi(m, s, a) - (\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m, s, a) \big| \\ &\le \big(3H_{\max}\sqrt{d\lambda} + C_1 d H_{\max}\sqrt{X}\big) \sqrt{\phi(s,a)^\top (\Lambda_h^k(m))^{-1} \phi(s,a)}. \end{aligned}$$

Thus the desired reward decomposition holds with

$$\Delta_{r,h}^k(m,s,a) := \langle \phi(s,a), q_1+q_2+q_3\rangle - (\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m,s,a).$$

Since $\lambda = 1$, $A \geq 1$, and $X = \log\Big(\frac{2(C_\beta+1)MdT}{p}\Big) \leq \iota + \log(C_\beta+1)$, with $\iota = \log(2MdTA/p)$, we can choose C_β as a sufficiently large absolute constant such that

$$3H_{\max}\sqrt{d\lambda} + C_1 dH_{\max}\sqrt{X} \leq C_\beta dH_{\max}\sqrt{\iota} = \beta.$$

which proves

$$|\Delta_{r,h}^k(m,s,a)| \leq \beta\sqrt{\phi(s,a)^\top(\Lambda_h^k(m))^{-1}\phi(s,a)}.$$

The proof for the cost function is identical. □

C.4 Cumulative estimation error control

We now sum the one-step errors along the executed trajectory to obtain the cumulative estimation error bound used in the main text.

Lemma 7 (Cumulative estimation error)**.** *Let π^k denote the policy executed at episode k, let $T = KH$, and define $\iota = \log(2MAdT/p)$. If we set $p = 1/K$ and $\beta = C_\beta dH_{\max}\sqrt{\iota}$ for a sufficiently large absolute constant $C_\beta > 0$, then*

$$\sum_{k=1}^K \mathbb{E}\Big[\hat{V}_1^k(m_k,s_1^k) - V_1^{\pi^k}(m_k,s_1^k)\Big] \leq O\Big(\sqrt{Md^3H^3T\,\iota^2}\Big),$$

and

$$-\sum_{k=1}^K \mathbb{E}\Big[\check{W}_1^k(m_k,s_1^k) - W_1^{\pi^k}(m_k,s_1^k)\Big] \leq O\Big(\sqrt{Md^3H^3T\,\iota^2}\Big).$$

Proof. We prove the reward bound; the cost bound follows identically. Let

$$\delta_h^k := \hat{V}_h^k(m_k,s_h^k) - V_h^{\pi^k}(m_k,s_h^k),$$

so that $\delta_h^k \in [-H_{\max}, H_{\max}]$.

We first decompose the expected cumulative error according to the good event $\mathcal{E}$ from Lemma 5:

$$\sum_{k=1}^K \mathbb{E}[\delta_1^k] = \mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\sum_{k=1}^K \delta_1^k\right] + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}^c\}\sum_{k=1}^K \delta_1^k\right].$$

Since $\mathbb{P}(\mathcal{E}) \geq 1 - p/2$, the bad-event term satisfies

$$\mathbb{E}\left[\mathbf{1}\{\mathcal{E}^c\}\sum_{k=1}^K \delta_1^k\right] \leq \frac{p}{2}KH_{\max}.$$

It remains to control the good-event term. On $\mathcal{E}$, fix an episode k and a step h. By the definition of $V_h^{\pi^k}$,

$$\delta_h^k = \sum_{a\in\mathcal{A}} \pi_h^k(a \mid m_k,s_h^k)\big(\hat{Q}_h^k(m_k,s_h^k,a) - Q_h^{\pi^k}(m_k,s_h^k,a)\big).$$

Adding and subtracting the term evaluated at the sampled action a_h^k, define

$$\zeta_{h,1}^k := \delta_h^k - \big(\hat{Q}_h^k(m_k,s_h^k,a_h^k) - Q_h^{\pi^k}(m_k,s_h^k,a_h^k)\big).$$

Then $\{\zeta_{h,1}^k\}$ is a martingale difference sequence.

Next, Lemma 6 implies

$$\hat{Q}_h^k(m_k,s_h^k,a_h^k) - Q_h^{\pi^k}(m_k,s_h^k,a_h^k) \leq (\mathbb{P}_h\delta_{h+1}^k)(m_k,s_h^k,a_h^k) + 2\beta\, b_h^k,$$

where

$$b_h^k := \sqrt{(\phi_h^k)^\top (\Lambda_h^k(m_k))^{-1} \phi_h^k}, \qquad \phi_h^k := \phi(s_h^k, a_h^k).$$

Hence

$$\delta_h^k \le (\mathbb{P}_h \delta_{h+1}^k)(m_k, s_h^k, a_h^k) + 2\beta\, b_h^k + \zeta_{h,1}^k.$$

Now define

$$\zeta_{h,2}^k := (\mathbb{P}_h \delta_{h+1}^k)(m_k, s_h^k, a_h^k) - \delta_{h+1}^k.$$

Then $\{\zeta_{h,2}^k\}$ is also a martingale difference sequence. Let $\zeta_h^k := \zeta_{h,1}^k + \zeta_{h,2}^k$. We obtain

$$\delta_h^k \le \delta_{h+1}^k + \zeta_h^k + 2\beta\, b_h^k.$$

Unrolling over $h = 1, \dots, H$, $k = 1, \dots, K$ and using $\delta_{H+1}^k = 0$ gives

$$\sum_{k=1}^K \delta_1^k \le \sum_{k=1}^K \sum_{h=1}^H \zeta_h^k + 2\beta \sum_{k=1}^K \sum_{h=1}^H b_h^k.$$

We now bound the two terms on the right-hand side.

For the martingale term, define the filtration $\mathcal{F}_{k,h}$ as the sigma-algebra generated by all information up to the beginning of step h in episode k (including m_k and s_h^k), and let $\mathcal{F}_{k,h}^+$ be the filtration after additionally sampling the action a_h^k. Then $\mathbb{E}[\zeta_{h,1}^k \mid \mathcal{F}_{k,h}] = 0$ and $\mathbb{E}[\zeta_{h,2}^k \mid \mathcal{F}_{k,h}^+] = 0$. Since $\mathcal{F}_{k,h} \subset \mathcal{F}_{k,h}^+ \subset \mathcal{F}_{k,h+1}$, the interleaved sequence $\{\zeta_h^k\}_{k,h}$ (with $\zeta_h^k := \zeta_{h,1}^k + \zeta_{h,2}^k$) forms a martingale difference sequence with respect to the natural filtration. Moreover, $|\zeta_{h,1}^k| \le 2H_{\max}$ and $|\zeta_{h,2}^k| \le 2H_{\max}$ (as all value functions lie in $[0, H_{\max}]$), so $|\zeta_h^k| \le 4H_{\max}$. By the Azuma–Hoeffding inequality, with probability at least $1 - p/2$,

$$\sum_{k=1}^K \sum_{h=1}^H \zeta_h^k \le 8H_{\max}\sqrt{T\iota}.$$

Let $\mathcal{E}_{\mathrm{Az}}^V$ denote the event on which the above Azuma–Hoeffding bound holds. Then $\mathbb{P}\big((\mathcal{E}_{\mathrm{Az}}^V)^c\big) \le \frac{p}{2}$.

For the bonus term, fix a step h. We write

$$\sum_{k=1}^K (\phi_h^k)^\top (\Lambda_h^k(m_k))^{-1} \phi_h^k = \sum_{m \in \mathcal{M}} \sum_{k=1}^K \mathbf{1}\{m_k = m\}\, (\phi_h^k)^\top (\Lambda_h^k(m))^{-1} \phi_h^k.$$

By the elliptical potential lemma (see, e.g., Jin et al. (2020)), for each fixed $m \in \mathcal{M}$,

$$\sum_{k=1}^K \mathbf{1}\{m_k = m\}\, (\phi_h^k)^\top (\Lambda_h^k(m))^{-1} \phi_h^k \le 2d \log\Big(\frac{\lambda + \sum_{k=1}^K \mathbf{1}\{m_k = m\}}{\lambda}\Big) \le 2d \log\Big(\frac{\lambda + K}{\lambda}\Big).$$

Summing over $m \in \mathcal{M}$ yields

$$\sum_{k=1}^K (\phi_h^k)^\top (\Lambda_h^k(m_k))^{-1} \phi_h^k \le 2dM\iota.$$

Applying Cauchy–Schwarz over the K episodes gives

$$\sum_{k=1}^K b_h^k \le \sqrt{K} \left(\sum_{k=1}^K (\phi_h^k)^\top (\Lambda_h^k(m_k))^{-1} \phi_h^k \right)^{1/2} \le \sqrt{K}\sqrt{2dM\iota}.$$

Summing over $h = 1, \dots, H$, we obtain

$$\sum_{k=1}^K \sum_{h=1}^H b_h^k \le H\sqrt{2dMK\iota}.$$

On the event $\mathcal{E} \cap \mathcal{E}_{\mathrm{Az}}^V$, combining the Azuma–Hoeffding bound with the bonus bound gives

$$\sum_{k=1}^{K} \delta_1^k \le 8H_{\max}\sqrt{T\iota} + 2\beta H\sqrt{2dMK\iota}.$$

On the event $\mathcal{E} \cap (\mathcal{E}_{\mathrm{Az}}^V)^c$, we use the uniform bound $\delta_1^k \le H_{\max}$ for all k, and hence $\sum_{k=1}^{K} \delta_1^k \le KH_{\max}$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\{\mathcal{E}\}\sum_{k=1}^{K}\delta_1^k\right] &= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}\cap\mathcal{E}_{\mathrm{Az}}^V\}\sum_{k=1}^{K}\delta_1^k\right] + \mathbb{E}\left[\mathbf{1}\{\mathcal{E}\cap(\mathcal{E}_{\mathrm{Az}}^V)^c\}\sum_{k=1}^{K}\delta_1^k\right] \\ &\le 8H_{\max}\sqrt{T\iota} + 2\beta H\sqrt{2dMK\iota} + \frac{p}{2}KH_{\max}.\end{aligned}$$

Together with the bad-event bound for $\mathcal{E}^c$, this gives

$$\sum_{k=1}^{K}\mathbb{E}[\delta_1^k] \le 8H_{\max}\sqrt{T\iota} + 2\beta H\sqrt{2dMK\iota} + pKH_{\max}.$$

Substituting $p = 1/K$, $\beta = C_\beta dH_{\max}\sqrt{\iota}$, and $T = KH$, we conclude

$$\sum_{k=1}^{K}\mathbb{E}\Big[\hat{V}_1^k(m_k, s_1^k) - V_1^{\pi^k}(m_k, s_1^k)\Big] \le O\Big(\sqrt{Md^3H^3T\,\iota^2}\Big).$$

The proof for the cost value function is identical. □

Lemma 7 will later be combined with the structural surrogate bound to control the reward and cost replacement errors in the main proof.

D Structural control of the queue-adjusted surrogate

In this section, we control the queue-adjusted surrogate term

$$\sum_{k=1}^{K}\mathbb{E}\Big[\hat{V}_1^k - \theta^\star\check{W}_1^k\Big].$$

The argument has two components. First, we use the Softmax policy construction and the optimism of the combined action-value estimate to relate the executed policy to a suitable reference policy up to a controllable approximation term. Second, we combine the virtual-queue recursion with the auxiliary-variable update to derive a one-step drift-plus-penalty inequality, whose summation yields a telescoping structural bound.

D.1 Softmax and optimism chain

We begin with the Softmax approximation bound for the combined value and then establish the optimism of the queue-adjusted surrogate. These ingredients allow us to compare the executed policy against a suitable reference policy on the good event.

Lemma 8 (Value function bound via Softmax policy)**.** *For any episode $k \in [K]$, step $h \in [H]$, task $m \in \mathcal{M}$, and state $s \in \mathcal{S}$, define*

$$z_h^k(a) := v\hat{Q}_h^k(m, s, a) - J_k\check{C}_h^k(m, s, a), \qquad a \in \mathcal{A}.$$

Under the Softmax policy

$$\pi_h^k(a \mid m, s) = \frac{\exp(\eta z_h^k(a))}{\sum_{a'\in\mathcal{A}}\exp(\eta z_h^k(a'))},$$

the corresponding value functions satisfy

$$v\hat{V}_h^k(m, s) - J_k\check{W}_h^k(m, s) \ge \max_{a\in\mathcal{A}} z_h^k(a) - \frac{\log A}{\eta}.$$

Proof. By definition of $\hat{V}_h^k$ and $\check{W}_h^k$ under the policy π_h^k,

$$v\hat{V}_h^k(m,s) - J_k\check{W}_h^k(m,s) = \sum_{a\in\mathcal{A}} \pi_h^k(a \mid m,s)\, z_h^k(a).$$

Since π_h^k is the Gibbs distribution associated with z_h^k, we have the standard identity

$$\sum_{a\in\mathcal{A}} \pi_h^k(a \mid m,s)\, z_h^k(a) = \frac{1}{\eta}\log\Big(\sum_{a\in\mathcal{A}} e^{\eta z_h^k(a)}\Big) - \frac{1}{\eta}\sum_{a\in\mathcal{A}} \pi_h^k(a \mid m,s)\log\frac{1}{\pi_h^k(a \mid m,s)}.$$

The first term is lower bounded by $\max_{a\in\mathcal{A}} z_h^k(a)$, while the entropy term is upper bounded by $\log A$. Therefore,

$$\sum_{a\in\mathcal{A}} \pi_h^k(a \mid m,s)\, z_h^k(a) \geq \max_{a\in\mathcal{A}} z_h^k(a) - \frac{\log A}{\eta},$$

which proves the lemma. □

Lemma 9 (Optimism of the combined Q-function)**.** *Fix any reference policy π. On the event $\mathcal{E}$ defined in Lemma 5, for any episode $k \in [K]$, task $m \in \mathcal{M}$, step $h \in [H]$, and state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, define*

$$Z_h^k(m,s,a) := v\hat{Q}_h^k(m,s,a) - J_k\check{C}_h^k(m,s,a), \qquad Z_h^\pi(m,s,a) := vQ_h^\pi(m,s,a) - J_kC_h^\pi(m,s,a).$$

Then

$$Z_h^k(m,s,a) \geq Z_h^\pi(m,s,a) - (H-h)\frac{\log A}{\eta}.$$

Proof. We prove the claim by backward induction on h.

For the base case $h = H+1$, both the estimated and true value functions are zero by definition, and hence

$$Z_{H+1}^k(m,s,a) = Z_{H+1}^\pi(m,s,a) = 0.$$

Therefore the claim holds trivially.

Now assume the statement holds at step $h+1$, namely, for all $s' \in \mathcal{S}$ and $a' \in \mathcal{A}$,

$$Z_{h+1}^k(m,s',a') \geq Z_{h+1}^\pi(m,s',a') - (H-h-1)\frac{\log A}{\eta}.$$

We consider step h. By definition,

$$Z_h^k(m,s,a) - Z_h^\pi(m,s,a) = v\big(\hat{Q}_h^k - Q_h^\pi\big)(m,s,a) - J_k\big(\check{C}_h^k - C_h^\pi\big)(m,s,a).$$

By Lemma 6 and the constructions of $\hat{Q}_h^k$ and $\check{C}_h^k$, we have

$$\hat{Q}_h^k(m,s,a) \geq Q_h^\pi(m,s,a) + (\mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi))(m,s,a),$$

and

$$\check{C}_h^k(m,s,a) \leq C_h^\pi(m,s,a) + (\mathbb{P}_h(\check{W}_{h+1}^k - W_{h+1}^\pi))(m,s,a).$$

Substituting these two bounds into the difference above gives

$$Z_h^k(m,s,a) - Z_h^\pi(m,s,a) \geq (\mathbb{P}_h[(v\hat{V}_{h+1}^k - J_k\check{W}_{h+1}^k) - (vV_{h+1}^\pi - J_kW_{h+1}^\pi)])(m,s,a).$$

We now bound the term inside the transition expectation. Fix any next state s'. By Lemma 8,

$$v\hat{V}_{h+1}^k(m,s') - J_k\check{W}_{h+1}^k(m,s') \geq \max_{a'\in\mathcal{A}} Z_{h+1}^k(m,s',a') - \frac{\log A}{\eta}.$$

Applying the induction hypothesis to the maximum term yields

$$\max_{a'\in\mathcal{A}} Z_{h+1}^k(m,s',a') \geq \max_{a'\in\mathcal{A}} Z_{h+1}^\pi(m,s',a') - (H-h-1)\frac{\log A}{\eta}.$$

Moreover, the maximum over actions is always no smaller than the expectation under the reference policy $\pi_{h+1}(\cdot \mid m, s')$, so

$$\max_{a' \in \mathcal{A}} Z^{\pi}_{h+1}(m, s', a') \geq \sum_{a' \in \mathcal{A}} \pi_{h+1}(a' \mid m, s')\, Z^{\pi}_{h+1}(m, s', a') = vV^{\pi}_{h+1}(m, s') - J_k W^{\pi}_{h+1}(m, s').$$

Combining the last three displays, we obtain

$$(v\hat{V}^{k}_{h+1} - J_k \check{W}^{k}_{h+1})(m, s') \geq (vV^{\pi}_{h+1} - J_k W^{\pi}_{h+1})(m, s') - (H - h)\frac{\log A}{\eta}.$$

Since this bound holds for every s', taking the transition expectation preserves it. Therefore,

$$Z^{k}_{h}(m, s, a) - Z^{\pi}_{h}(m, s, a) \geq -(H - h)\frac{\log A}{\eta},$$

which proves the induction step and hence the lemma. □

Lemma 10 (Expected bound on reference policy evaluation)**.** *Let $\mathcal{E}$ be the event defined in Lemma 5, so that $\mathbb{P}(\mathcal{E}) \geq 1 - p/2$. Fix any episode $k \in [K]$ and any reference policy π. For the realized initial task-state pair (m_k, s^k_1), define*

$$\hat{V}^{k,\pi}_{1}(m_k, s^k_1) := \sum_{a \in \mathcal{A}} \pi(a \mid m_k, s^k_1)\hat{Q}^{k}_{1}(m_k, s^k_1, a),$$

and

$$\check{W}^{k,\pi}_{1}(m_k, s^k_1) := \sum_{a \in \mathcal{A}} \pi(a \mid m_k, s^k_1)\check{C}^{k}_{1}(m_k, s^k_1, a).$$

Then

$$\mathbb{E}\Big[-v\big(\hat{V}^{k,\pi}_{1} - V^{\pi}_{1}\big)(m_k, s^k_1) + J_k\big(\check{W}^{k,\pi}_{1} - W^{\pi}_{1}\big)(m_k, s^k_1)\Big] \leq H\frac{\log A}{\eta} + \frac{p}{2}(v + J_{\max})H_{\max}.$$

Proof. We decompose the expectation into two parts according to the good event $\mathcal{E}$ and the bad event $\mathcal{E}^c$:

$$\begin{aligned}
&\mathbb{E}\Big[-v\big(\hat{V}^{k,\pi}_{1} - V^{\pi}_{1}\big)(m_k, s^k_1) + J_k\big(\check{W}^{k,\pi}_{1} - W^{\pi}_{1}\big)(m_k, s^k_1)\Big] \\
&= \mathbb{E}\Big[\mathbf{1}\{\mathcal{E}\}\Big(-v\big(\hat{V}^{k,\pi}_{1} - V^{\pi}_{1}\big) + J_k\big(\check{W}^{k,\pi}_{1} - W^{\pi}_{1}\big)\Big)(m_k, s^k_1)\Big] \\
&\quad + \mathbb{E}\Big[\mathbf{1}\{\mathcal{E}^c\}\Big(-v\big(\hat{V}^{k,\pi}_{1} - V^{\pi}_{1}\big) + J_k\big(\check{W}^{k,\pi}_{1} - W^{\pi}_{1}\big)\Big)(m_k, s^k_1)\Big].
\end{aligned}$$

We first bound the good-event term. On $\mathcal{E}$, Lemma 9 implies that for any task $m \in \mathcal{M}$ and any state-action pair (s, a) at step $h = 1$,

$$-v\big(\hat{Q}^{k}_{1} - Q^{\pi}_{1}\big)(m, s, a) + J_k\big(\check{C}^{k}_{1} - C^{\pi}_{1}\big)(m, s, a) \leq (H - 1)\frac{\log A}{\eta}.$$

By the definitions of $\hat{V}^{k,\pi}_{1}$ and $\check{W}^{k,\pi}_{1}$, taking the expectation over the action distribution $\pi(\cdot \mid m_k, s^k_1)$ preserves the same bound:

$$-v\big(\hat{V}^{k,\pi}_{1} - V^{\pi}_{1}\big)(m_k, s^k_1) + J_k\big(\check{W}^{k,\pi}_{1} - W^{\pi}_{1}\big)(m_k, s^k_1) \leq (H - 1)\frac{\log A}{\eta}.$$

Since $\mathbb{P}(\mathcal{E}) \leq 1$, the contribution of the good event is bounded by

$$\mathbb{E}\Big[\mathbf{1}\{\mathcal{E}\}\Big(-v\big(\hat{V}^{k,\pi}_{1} - V^{\pi}_{1}\big) + J_k\big(\check{W}^{k,\pi}_{1} - W^{\pi}_{1}\big)\Big)(m_k, s^k_1)\Big] \leq H\frac{\log A}{\eta}.$$

We next bound the bad-event term. On $\mathcal{E}^c$, we use the boundedness of the value functions. Since both the estimated and true reward value functions lie in $[0, H_{\max}]$, we have

$$\big|\hat{V}^{k,\pi}_{1}(m_k, s^k_1) - V^{\pi}_{1}(m_k, s^k_1)\big| \leq H_{\max}.$$

Similarly, both the estimated and true cost value functions lie in $[0, H_{\max}]$, so

$$\left|\check{W}_1^{k,\pi}(m_k, s_1^k) - W_1^{\pi}(m_k, s_1^k)\right| \le H_{\max}.$$

Therefore,

$$\left|-v\big(\hat{V}_1^{k,\pi} - V_1^{\pi}\big)(m_k, s_1^k) + J_k\big(\check{W}_1^{k,\pi} - W_1^{\pi}\big)(m_k, s_1^k)\right| \le (v + J_{\max})H_{\max}.$$

Using $\mathbb{P}(\mathcal{E}^c) \le p/2$, we obtain

$$\mathbb{E}\Big[\mathbf{1}\{\mathcal{E}^c\}\Big(-v\big(\hat{V}_1^{k,\pi} - V_1^{\pi}\big) + J_k\big(\check{W}_1^{k,\pi} - W_1^{\pi}\big)\Big)(m_k, s_1^k)\Big] \le \frac{p}{2}(v + J_{\max})H_{\max}.$$

Combining the two bounds proves the lemma. □

D.2 Lyapunov drift analysis

We next analyze the virtual queue and the auxiliary update. We first establish a uniform bound on the queue length, then derive the basic Lyapunov drift inequality, and finally combine it with the auxiliary update to obtain the one-step drift-plus-penalty bound used in the final synthesis.

Lemma 11 (Bound on the virtual queue)**.** *Let the virtual queue be updated by*

$$J_{k+1} = \Big[J_k + \check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\Big]_+.$$

Then, for all $k \ge 1$,

$$0 \le J_k \le v(\alpha_1 + \alpha_2),$$

where

$$\alpha_1 := \frac{1 + Hr_{\max}}{Hc_{\min}}, \qquad \alpha_2 := \frac{n}{v}(Hc_{\max} - Hc_{\min}), \qquad n := \Big\lceil v\gamma_{\max}(\gamma_{\max} - \gamma_{\min})\Big\rceil.$$

Proof. We first establish two claims.

Claim 1. If $J_{k_0} \le v\alpha_1$ for some $k_0 \ge 1$, then

$$J_k \le v(\alpha_1 + \alpha_2), \qquad k \in \{k_0, k_0 + 1, \dots, k_0 + n\}.$$

Proof of Claim 1. Since $\check{W}_1^k(m_k, s_1^k) \le Hc_{\max}$ and $\gamma_k \le \gamma_{\max}$, we have

$$\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k} \le Hc_{\max} - \frac{1}{\gamma_{\max}} = Hc_{\max} - Hc_{\min}.$$

Hence, over any n consecutive episodes, the queue can increase by at most

$$n(Hc_{\max} - Hc_{\min}) = v\alpha_2.$$

Therefore, if $J_{k_0} \le v\alpha_1$, then

$$J_k \le v\alpha_1 + n(Hc_{\max} - Hc_{\min}) = v(\alpha_1 + \alpha_2)$$

for all $k \in \{k_0, \dots, k_0 + n\}$. This proves Claim 1.

Claim 2. If $J_k \ge v\alpha_1$, then

$$\gamma_k \le \max\left\{\gamma_{\min},\ \gamma_{k-1} - \frac{1}{\gamma_{\max} v}\right\}.$$

Proof of Claim 2. From the update rule of the auxiliary variable,

$$\gamma_k = \Pi_{[\gamma_{\min},\gamma_{\max}]}\left[\gamma_{k-1} + \frac{v\hat{V}_1^k(m_k, s_1^k) - J_k\check{W}_1^k(m_k, s_1^k)}{\gamma_{k-1}v^2}\right].$$

If $J_k \geq v\alpha_1$, then using $\hat{V}_1^k(m_k, s_1^k) \leq Hr_{\max}$ and $\check{W}_1^k(m_k, s_1^k) \geq Hc_{\min}$, we obtain

$$\frac{v\hat{V}_1^k(m_k, s_1^k) - J_k\check{W}_1^k(m_k, s_1^k)}{\gamma_{k-1}v^2} \leq \frac{vHr_{\max} - J_kHc_{\min}}{\gamma_{k-1}v^2} \leq \frac{vHr_{\max} - v\alpha_1Hc_{\min}}{\gamma_{k-1}v^2}.$$

By the definition of α_1,

$$vHr_{\max} - v\alpha_1Hc_{\min} = vHr_{\max} - v\Big(\frac{1 + Hr_{\max}}{Hc_{\min}}\Big)Hc_{\min} = -v,$$

and hence

$$\frac{v\hat{V}_1^k(m_k, s_1^k) - J_k\check{W}_1^k(m_k, s_1^k)}{\gamma_{k-1}v^2} \leq -\frac{1}{\gamma_{k-1}v} \leq -\frac{1}{\gamma_{\max}v}.$$

Applying the projection onto $[\gamma_{\min}, \gamma_{\max}]$ gives

$$\gamma_k \leq \max\left\{\gamma_{\min},\ \gamma_{k-1} - \frac{1}{\gamma_{\max}v}\right\},$$

which proves Claim 2.

We now prove the lemma by induction on k.

For the base case, $J_1 = 0 \leq v\alpha_1$. Hence Claim 1 implies

$$J_k \leq v(\alpha_1 + \alpha_2), \qquad k \in \{1, \ldots, 1 + n\}.$$

Now assume that

$$J_k \leq v(\alpha_1 + \alpha_2), \qquad k \in \{1, \ldots, k_0\}$$

for some $k_0 \geq 1 + n$. We prove that $J_{k_0+1} \leq v(\alpha_1 + \alpha_2)$.

If there exists some $k \in \{k_0 + 1 - n, \ldots, k_0\}$ such that $J_k \leq v\alpha_1$, then Claim 1 immediately gives

$$J_{k_0+1} \leq v(\alpha_1 + \alpha_2).$$

It remains to consider the case

$$J_k > v\alpha_1, \qquad k \in \{k_0 + 1 - n, \ldots, k_0\}.$$

We distinguish two subcases.

Case 1: If $\gamma_k = \gamma_{\min}$ for some $k \in \{k_0 + 1 - n, \ldots, k_0\}$, then by Claim 2 the sequence $\{\gamma_k\}$ is nonincreasing on this window, so in particular $\gamma_{k_0} = \gamma_{\min} = 1/Hc_{\max}$. Therefore,

$$J_{k_0+1} = \Big[J_{k_0} + \check{W}_1^{k_0}(m_{k_0}, s_1^{k_0}) - \frac{1}{\gamma_{k_0}}\Big]_+ \leq \big[J_{k_0} + Hc_{\max} - Hc_{\max}\big]_+ \leq J_{k_0} \leq v(\alpha_1 + \alpha_2).$$

Case 2: Suppose instead that

$$\gamma_k > \gamma_{\min}, \qquad k \in \{k_0 + 1 - n, \ldots, k_0\}.$$

Then Claim 2 implies that for every such k,

$$\gamma_k - \gamma_{k-1} \leq -\frac{1}{\gamma_{\max}v}.$$

Summing over the n steps in this window yields

$$\gamma_{k_0} \leq \gamma_{k_0-n} - \frac{n}{\gamma_{\max}v} \leq \gamma_{\max} - \frac{n}{\gamma_{\max}v}.$$

Since $n = \lceil v\gamma_{\max}(\gamma_{\max} - \gamma_{\min})\rceil$, we have

$$\gamma_{k_0} \leq \gamma_{\max} - (\gamma_{\max} - \gamma_{\min}) = \gamma_{\min},$$

which contradicts the assumption of Case 2. Hence this case is impossible.

Therefore, in all cases, $J_{k_0+1} \leq v(\alpha_1 + \alpha_2)$. By induction, the bound holds for all $k \geq 1$. The lower bound $J_k \geq 0$ follows directly from the update rule. □

Lemma 12 (Basic drift bound)**.** *Let the Lyapunov function be $L_k := \frac{1}{2}J_k^2$, $\Delta_k := L_{k+1} - L_k$. Then, for any episode $k \in [K]$,*

$$\Delta_k - v\hat{V}_1^k(m_k, s_1^k) \le b - v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right),$$

where

$$b := \frac{1}{2}(Hc_{\max} - Hc_{\min})^2.$$

Proof. Recall the virtual queue update $J_{k+1} = \left[J_k + \check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right]_+$. Using $[x]_+^2 \le x^2$ for any $x \in \mathbb{R}$, we have

$$J_{k+1}^2 \le \left(J_k + \check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)^2.$$

Expanding the square gives

$$J_{k+1}^2 \le J_k^2 + 2J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right) + \left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)^2.$$

Dividing by 2 and rearranging, we obtain

$$\Delta_k \le J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right) + \frac{1}{2}\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)^2.$$

By definition, both $\check{W}_1^k(m_k, s_1^k)$ and $1/\gamma_k$ lie in the interval $[Hc_{\min}, Hc_{\max}]$. Hence

$$\frac{1}{2}\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)^2 \le \frac{1}{2}(Hc_{\max} - Hc_{\min})^2 = b.$$

Substituting this bound into the previous display yields

$$\Delta_k \le b + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right).$$

Finally, subtracting $v\hat{V}_1^k(m_k, s_1^k)$ from both sides gives

$$\Delta_k - v\hat{V}_1^k(m_k, s_1^k) \le b - v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right),$$

which proves the lemma. □

Lemma 13 (Drift-plus-penalty via auxiliary optimization)**.** *Fix any episode $k \in [K]$ and any reference policy π. For the realized initial task-state pair (m_k, s_1^k), define*

$$\hat{V}_1^{k,\pi}(m_k, s_1^k) := \sum_{a \in \mathcal{A}} \pi(a \mid m_k, s_1^k)\hat{Q}_1^k(m_k, s_1^k, a),$$

and

$$\check{W}_1^{k,\pi}(m_k, s_1^k) := \sum_{a \in \mathcal{A}} \pi(a \mid m_k, s_1^k)\check{C}_1^k(m_k, s_1^k, a).$$

Then, for any $\gamma^\star \in [\gamma_{\min}, \gamma_{\max}]$, the one-step drift-plus-penalty satisfies

$$\begin{aligned}\Delta_k - v\hat{V}_1^k(m_k, s_1^k) \le b + \frac{\gamma^\star}{\gamma_{k-1}}\left[-v\hat{V}_1^{k,\pi}(m_k, s_1^k) + J_k\left(\check{W}_1^{k,\pi}(m_k, s_1^k) - \frac{1}{\gamma^\star}\right)\right] \\ + \frac{v^2}{2}(\gamma^\star - \gamma_{k-1})^2 - \frac{v^2}{2}(\gamma^\star - \gamma_k)^2 + \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} \\ + \frac{\left(vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k\right)^2}{2\gamma_{\min}^2 v^2}.\end{aligned}$$

Proof. By Lemma 8, the value constructed under the Softmax policy satisfies

$$-v\hat{V}_1^k(m_k, s_1^k) + J_k\check{W}_1^k(m_k, s_1^k) \le \min_{a\in\mathcal{A}}\big[-v\hat{Q}_1^k(m_k, s_1^k, a) + J_k\check{C}_1^k(m_k, s_1^k, a)\big] + \frac{\log A}{\eta}.$$

Since the minimum over actions is no larger than the expectation under any reference policy π, we obtain

$$-v\hat{V}_1^k(m_k, s_1^k) + J_k\check{W}_1^k(m_k, s_1^k) \le -v\hat{V}_1^{k,\pi}(m_k, s_1^k) + J_k\check{W}_1^{k,\pi}(m_k, s_1^k) + \frac{\log A}{\eta}.$$

Recall that γ_k is chosen to minimize the strongly convex objective

$$\gamma\Big[-v\hat{V}_1^k(m_k, s_1^k) + J_k\check{W}_1^k(m_k, s_1^k)\Big] + \frac{\gamma_{k-1}v^2}{2}(\gamma - \gamma_{k-1})^2$$

over $\gamma \in [\gamma_{\min}, \gamma_{\max}]$. By the standard pushback property of Euclidean projection for strongly convex functions, for any $\gamma^\star \in [\gamma_{\min}, \gamma_{\max}]$,

$$\begin{aligned}
&\gamma_k\Big[-v\hat{V}_1^k(m_k, s_1^k) + J_k\check{W}_1^k(m_k, s_1^k)\Big] + \frac{\gamma_{k-1}v^2}{2}(\gamma_k - \gamma_{k-1})^2 \\
&\le \gamma^\star\Big[-v\hat{V}_1^k(m_k, s_1^k) + J_k\check{W}_1^k(m_k, s_1^k)\Big] + \frac{\gamma_{k-1}v^2}{2}(\gamma^\star - \gamma_{k-1})^2 - \frac{\gamma_{k-1}v^2}{2}(\gamma^\star - \gamma_k)^2.
\end{aligned}$$

Substituting the previous policy bound into the right-hand side gives

$$\begin{aligned}
&\gamma_k\Big[-v\hat{V}_1^k(m_k, s_1^k) + J_k\check{W}_1^k(m_k, s_1^k)\Big] + \frac{\gamma_{k-1}v^2}{2}(\gamma_k - \gamma_{k-1})^2 \\
&\le \gamma^\star\Big[-v\hat{V}_1^{k,\pi}(m_k, s_1^k) + J_k\check{W}_1^{k,\pi}(m_k, s_1^k)\Big] + \frac{\gamma_{k-1}v^2}{2}(\gamma^\star - \gamma_{k-1})^2 - \frac{\gamma_{k-1}v^2}{2}(\gamma^\star - \gamma_k)^2 + \gamma^\star\frac{\log A}{\eta}.
\end{aligned}$$

Dividing both sides by γ_{k-1} and subtracting J_k/γ_{k-1} gives

$$\begin{aligned}
\mathrm{LHS}_k &:= \frac{\gamma_k}{\gamma_{k-1}}\left[-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)\right] + \frac{v^2}{2}(\gamma_k - \gamma_{k-1})^2 \\
&\le \frac{\gamma^\star}{\gamma_{k-1}}\left[-v\hat{V}_1^{k,\pi}(m_k, s_1^k) + J_k\left(\check{W}_1^{k,\pi}(m_k, s_1^k) - \frac{1}{\gamma^\star}\right)\right] \\
&\quad + \frac{v^2}{2}(\gamma^\star - \gamma_{k-1})^2 - \frac{v^2}{2}(\gamma^\star - \gamma_k)^2 + \frac{\gamma^\star}{\gamma_{k-1}}\frac{\log A}{\eta}.
\end{aligned} \tag{22}$$

Since $\gamma^\star/\gamma_{k-1} \le \gamma_{\max}/\gamma_{\min} = c_{\max}/c_{\min}$, the last term is bounded by

$$\frac{\gamma^\star}{\gamma_{k-1}}\frac{\log A}{\eta} \le \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta}.$$

Using $\frac{\gamma_k}{\gamma_{k-1}} = 1 + \frac{\gamma_k - \gamma_{k-1}}{\gamma_{k-1}}$, we rewrite LHS_k as

$$\begin{aligned}
\mathrm{LHS}_k = &-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right) \\
&+ \frac{\gamma_k - \gamma_{k-1}}{\gamma_{k-1}}\left[-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)\right] + \frac{v^2}{2}(\gamma_k - \gamma_{k-1})^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right) \\
&= \mathrm{LHS}_k - \frac{\gamma_k - \gamma_{k-1}}{\gamma_{k-1}}\left[-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)\right] - \frac{v^2}{2}(\gamma_k - \gamma_{k-1})^2.
\end{aligned}$$

Applying the inequality $yx + \frac{v^2}{2}x^2 \geq -\frac{y^2}{2v^2}$ for all $x, y \in \mathbb{R}$, we obtain

$$\begin{aligned}
&- v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right) \\
&\leq \mathrm{LHS}_k + \frac{\left(-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)\right)^2}{2\gamma_{k-1}^2 v^2}.
\end{aligned}$$

Since $\gamma_{k-1} \geq \gamma_{\min}$ and

$$\left|-v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right)\right| \leq vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k,$$

we further obtain

$$\begin{aligned}
&- v\hat{V}_1^k(m_k, s_1^k) + J_k\left(\check{W}_1^k(m_k, s_1^k) - \frac{1}{\gamma_k}\right) \\
&\leq \mathrm{LHS}_k + \frac{\left(vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k\right)^2}{2\gamma_{\min}^2 v^2}.
\end{aligned}$$

Finally, by Lemma 12 and Eq.(22), we have

$$\begin{aligned}
\Delta_k - v\hat{V}_1^k(m_k, s_1^k) \leq b &+ \frac{\gamma^\star}{\gamma_{k-1}}\left[-v\hat{V}_1^{k,\pi}(m_k, s_1^k) + J_k\left(\check{W}_1^{k,\pi}(m_k, s_1^k) - \frac{1}{\gamma^\star}\right)\right] \\
&+ \frac{v^2}{2}(\gamma^\star - \gamma_{k-1})^2 - \frac{v^2}{2}(\gamma^\star - \gamma_k)^2 + \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} \\
&+ \frac{\left(vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k\right)^2}{2\gamma_{\min}^2 v^2},
\end{aligned}$$

which proves the lemma. □

D.3 Final one-step drift-plus-penalty bound

To connect the Lyapunov analysis with the optimal ratio benchmark $\theta^\star$, we introduce the virtual target set Γ and show that every target pair in Γ can be realized by a history-independent randomized policy. This allows us to compare the queue-adjusted surrogate against an optimal reference pair $(r^\star, c^\star)$ and conclude the final one-step drift-plus-penalty bound.

Before proving Lemma 16, we introduce the history $\mathcal{H}_k$ and the virtual target set Γ.

History. For each episode $k \in [K]$, define the history $\mathcal{H}_k$ as the collection of the constructed optimistic reward and pessimistic cost values evaluated at the initial states up to episode $k-1$:

$$\mathcal{H}_k = \left\{(\hat{V}_1^1(m_1, s_1^1), \check{W}_1^1(m_1, s_1^1)), \ldots, (\hat{V}_1^{k-1}(m_{k-1}, s_1^{k-1}), \check{W}_1^{k-1}(m_{k-1}, s_1^{k-1}))\right\},$$

with the convention that $\mathcal{H}_1 = \emptyset$.

Virtual target set. Let

$$\Gamma \subseteq [0, Hr_{\max}] \times [Hc_{\min}, Hc_{\max}]$$

denote the set of attainable expected per-episode performance pairs. More precisely, Γ consists of all pairs

$$\left(\mathbb{E}[V_1^\pi(m_1, s_1^1)],\ \mathbb{E}[W_1^\pi(m_1, s_1^1)]\right)$$

induced by admissible, possibly randomized, within-episode policies π, where the expectation is taken over the randomness of the sampled task $m_1 \sim \mathbb{P}_\mathcal{M}$, the initial state $s_1^1 \sim \nu(m_1)$, and the trajectory induced by π and the transition kernels $\mathbb{P}_h$.

Since the task types and initial states are drawn independently and identically distributed across episodes, the set of achievable expectations is the same for every episode:

$$\Gamma_1 = \Gamma_2 = \cdots = \Gamma_K = \Gamma.$$

Moreover, it can be shown that the set Γ is convex. Let $\bar{\Gamma}$ denote its closure. Since every $(r, c) \in \bar{\Gamma}$ satisfies $c \geq Hc_{\min} > 0$, the optimal ratio is well defined, and

$$\theta^\star = \sup_{(r,c)\in\bar{\Gamma}} \frac{r}{c}.$$

By compactness and continuity, the supremum is attained at some $(r^\star, c^\star) \in \bar{\Gamma}$.

Lemma 14 (Adaptive policy sequences are benchmark dominated)**.** *For any adaptive policy sequence $\Pi = \{\pi^k\}_{k=1}^K$ generated by the algorithm, define*

$$\bar{r}_\Pi := \frac{1}{K}\sum_{k=1}^K \mathbb{E}[V_1^{\pi^k}(m_k, s_1^k)], \qquad \bar{c}_\Pi := \frac{1}{K}\sum_{k=1}^K \mathbb{E}[W_1^{\pi^k}(m_k, s_1^k)].$$

Then $(\bar{r}_\Pi, \bar{c}_\Pi) \in \overline{\Gamma}$. Consequently,

$$\theta^\Pi = \frac{\sum_{k=1}^K \mathbb{E}[V_1^{\pi^k}(m_k, s_1^k)]}{\sum_{k=1}^K \mathbb{E}[W_1^{\pi^k}(m_k, s_1^k)]} \leq \theta^\star.$$

Proof. Conditioned on the history before episode k, the algorithm selects an admissible possibly randomized Markov policy π^k. Since the task type and initial state in episode k are drawn from the same distribution as in the definition of Γ, the conditional expected performance pair induced by π^k belongs to Γ. Taking expectation over the history gives a mixture of points in Γ, and therefore belongs to $\overline{\Gamma}$ by convexity and closure. Averaging over $k = 1, \ldots, K$ again gives a convex mixture, so $(\bar{r}_\Pi, \bar{c}_\Pi) \in \overline{\Gamma}$. Since every point in $\overline{\Gamma}$ has positive cost and $\theta^\star = \sup_{(r,c)\in\overline{\Gamma}} r/c$, we obtain $\theta^\Pi = \bar{r}_\Pi/\bar{c}_\Pi \leq \theta^\star$. □

Lemma 15 (Existence of an $\mathcal{H}_k$-independent realizing policy)**.** *For any episode $k \in [K]$ and any target pair $(r, c) \in \Gamma$, there exists an admissible randomized policy $\pi^{(r,c)}$, selected independently of $\mathcal{H}_k$, such that*

$$\mathbb{E}\Big[\big(V_1^{\pi^{(r,c)}}(m_k, s_1^k),\, W_1^{\pi^{(r,c)}}(m_k, s_1^k)\big) \,\Big|\, \mathcal{H}_k\Big] = (r, c) \qquad \textit{a.s.}$$

Proof. Fix any $(r, c) \in \Gamma$. By the definition of Γ, there exists an admissible randomized policy $\pi^{(r,c)}$ such that

$$\mathbb{E}\Big[\big(V_1^{\pi^{(r,c)}}(m_1, s_1^1),\, W_1^{\pi^{(r,c)}}(m_1, s_1^1)\big)\Big] = (r, c).$$

Since the admissible policy class is fixed a priori, we may choose one such realizing policy $\pi^{(r,c)}$ without using the realized history $\mathcal{H}_k$.

Moreover, the task type m_k and the initial state s_1^k are sampled independently of $\mathcal{H}_k$, and the internal randomization used to implement $\pi^{(r,c)}$ can be generated by an auxiliary random variable $U_k \sim \mathrm{Unif}(0, 1)$ independent of $\mathcal{H}_k$. Therefore, conditional on $\mathcal{H}_k$, the law of the episode generated by $\pi^{(r,c)}$ is the same as its unconditional law. Hence,

$$\mathbb{E}\Big[\big(V_1^{\pi^{(r,c)}}(m_k, s_1^k),\, W_1^{\pi^{(r,c)}}(m_k, s_1^k)\big) \,\Big|\, \mathcal{H}_k\Big] = \mathbb{E}\Big[\big(V_1^{\pi^{(r,c)}}(m_k, s_1^k),\, W_1^{\pi^{(r,c)}}(m_k, s_1^k)\big)\Big].$$

Since the task types and initial states are identically distributed across episodes, the achievable expectation is the same for every episode, and thus

$$\mathbb{E}\Big[\big(V_1^{\pi^{(r,c)}}(m_k, s_1^k),\, W_1^{\pi^{(r,c)}}(m_k, s_1^k)\big)\Big] = (r, c).$$

This completes the proof. □

Lemma 16 (Final expected one-step drift-plus-penalty bound)**.** *Fix every $k \geq 2$. Let $(r^\star, c^\star) \in \bar{\Gamma}$ be an optimal performance pair such that $\theta^\star := \frac{r^\star}{c^\star} = \sup_{(r,c)\in\bar{\Gamma}} \frac{r}{c}$. Then*

$$\begin{aligned}\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k(m_k, s_1^k)\Big] \leq b - v\theta^\star\,\mathbb{E}\big[\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1})\big] + v\theta^\star\,\mathbb{E}[J_k - J_{k-1}] \\ + \frac{v^2}{2}\,\mathbb{E}\left[\left(\frac{1}{c^\star} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c^\star} - \gamma_k\right)^2\right] \\ + \frac{c_{\max}}{c_{\min}}\left[H\frac{\log A}{\eta} + \frac{p}{2}v(1+\alpha_1+\alpha_2)H_{\max}\right] \\ + \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} + \frac{\big(Hr_{\max} + (Hc_{\max} - Hc_{\min})(\alpha_1+\alpha_2)\big)^2}{2\gamma_{\min}^2}.\end{aligned}$$

Proof. Since $(r^\star, c^\star) \in \bar{\Gamma}$, there exists a sequence $\{(r_j, c_j)\}_{j\geq 1} \subset \Gamma$ such that $(r_j, c_j) \to (r^\star, c^\star)$. For each $j \geq 1$, define $\theta_j := \frac{r_j}{c_j}$ and choose $\gamma_j^\star := \frac{1}{c_j}$. By Lemma 15, for each $j \geq 1$, there exists a history-independent randomized policy π_j such that

$$\mathbb{E}\big[\big(V_1^{\pi_j}(m_k, s_1^k), W_1^{\pi_j}(m_k, s_1^k)\big) \,\big|\, \mathcal{H}_k\big] = (r_j, c_j) \qquad \text{a.s.}$$

Apply Lemma 13 with reference policy π_j and auxiliary variable $\gamma_j^\star$. Since γ_{k-1} is $\mathcal{H}_k$-measurable, taking conditional expectation given $\mathcal{H}_k$ yields

$$\begin{aligned}\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k(m_k, s_1^k) \mid \mathcal{H}_k\Big] \leq b + \frac{1/c_j}{\gamma_{k-1}}\Big[-vr_j + J_k\Big(c_j - \frac{1}{\gamma_j^\star}\Big)\Big] \\ + \frac{v^2}{2}\mathbb{E}\left[\left(\frac{1}{c_j} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c_j} - \gamma_k\right)^2 \,\middle|\, \mathcal{H}_k\right] \\ + \frac{1/c_j}{\gamma_{k-1}}\mathbb{E}\Big[-v\big(\hat{V}_1^{k,\pi_j} - V_1^{\pi_j}\big)(m_k, s_1^k) \\ + J_k\big(\check{W}_1^{k,\pi_j} - W_1^{\pi_j}\big)(m_k, s_1^k) \,\Big|\, \mathcal{H}_k\Big] \\ + \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} + \frac{\big(vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k\big)^2}{2\gamma_{\min}^2 v^2}.\end{aligned} \tag{23}$$

Since $\gamma_j^\star = 1/c_j$, the queue term in the first line vanishes. Taking total expectation gives, for every $j \geq 1$,

$$\begin{aligned}\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k(m_k, s_1^k)\Big] \leq b - v\theta_j\,\mathbb{E}\left[\frac{1}{\gamma_{k-1}}\right] \\ + \frac{v^2}{2}\mathbb{E}\left[\left(\frac{1}{c_j} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c_j} - \gamma_k\right)^2\right] \\ + \mathbb{E}\left[\frac{1/c_j}{\gamma_{k-1}}\Big(-v\big(\hat{V}_1^{k,\pi_j} - V_1^{\pi_j}\big)(m_k, s_1^k) \\ + J_k\big(\check{W}_1^{k,\pi_j} - W_1^{\pi_j}\big)(m_k, s_1^k)\Big)\right] \\ + \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} + \mathbb{E}\left[\frac{\big(vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k\big)^2}{2\gamma_{\min}^2 v^2}\right].\end{aligned} \tag{24}$$

We now bound the terms in Eq.(24). First,

$$-v\theta_j\mathbb{E}\left[\frac{1}{\gamma_{k-1}}\right] = -v\theta_j\mathbb{E}\big[\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1})\big] + v\theta_j\mathbb{E}\left[\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1}) - \frac{1}{\gamma_{k-1}}\right].$$

By the queue update,

$$J_k = \Big[J_{k-1} + \check{W}_1^{k-1}(m_{k-1}, s_1^{k-1}) - \frac{1}{\gamma_{k-1}}\Big]_+,$$

and hence

$$\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1}) - \frac{1}{\gamma_{k-1}} \le J_k - J_{k-1}.$$

Therefore,

$$-v\theta_j \mathbb{E}\left[\frac{1}{\gamma_{k-1}}\right] \le -v\theta_j \mathbb{E}\big[\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1})\big] + v\theta_j \mathbb{E}[J_k - J_{k-1}].$$

Next, Lemma 10 applies to every history-independent reference policy π_j. Moreover, $\frac{1/c_j}{\gamma_{k-1}} \le \frac{c_{\max}}{c_{\min}}$, and Lemma 11 gives $J_k \le v(\alpha_1 + \alpha_2)$. Hence,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1/c_j}{\gamma_{k-1}}\Big(-v\big(\hat{V}_1^{k,\pi_j} - V_1^{\pi_j}\big) + J_k\big(\check{W}_1^{k,\pi_j} - W_1^{\pi_j}\big)\Big)(m_k, s_1^k)\right] \\
&\qquad \le \frac{c_{\max}}{c_{\min}}\left[H\frac{\log A}{\eta} + \frac{p}{2}v(1 + \alpha_1 + \alpha_2)H_{\max}\right].
\end{aligned}$$

Finally, using Lemma 11 again,

$$\frac{\big(vHr_{\max} + (Hc_{\max} - Hc_{\min})J_k\big)^2}{2\gamma_{\min}^2 v^2} \le \frac{\big(Hr_{\max} + (Hc_{\max} - Hc_{\min})(\alpha_1 + \alpha_2)\big)^2}{2\gamma_{\min}^2}.$$

Substituting the above bounds into Eq.(24), we obtain, for every $j \ge 1$,

$$\begin{aligned}
\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k(m_k, s_1^k)\Big] \le\ & b - v\theta_j\, \mathbb{E}\big[\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1})\big] + v\theta_j\, \mathbb{E}[J_k - J_{k-1}] \\
&+ \frac{v^2}{2}\mathbb{E}\left[\left(\frac{1}{c_j} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c_j} - \gamma_k\right)^2\right] \\
&+ \frac{c_{\max}}{c_{\min}}\left[H\frac{\log A}{\eta} + \frac{p}{2}v(1 + \alpha_1 + \alpha_2)H_{\max}\right] \\
&+ \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} + \frac{\big(Hr_{\max} + (Hc_{\max} - Hc_{\min})(\alpha_1 + \alpha_2)\big)^2}{2\gamma_{\min}^2}.
\end{aligned}$$

Since $(r_j, c_j) \to (r^\star, c^\star)$, we have $\theta_j \to \theta^\star$ and $1/c_j \to 1/c^\star$. All remaining terms are uniformly bounded in j, because $c_j \in [Hc_{\min}, Hc_{\max}]$ and J_k is uniformly bounded by Lemma 11. Letting $j \to \infty$ and applying bounded convergence yields

$$\begin{aligned}
\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k(m_k, s_1^k)\Big] \le\ & b - v\theta^\star\, \mathbb{E}\big[\check{W}_1^{k-1}(m_{k-1}, s_1^{k-1})\big] + v\theta^\star\, \mathbb{E}[J_k - J_{k-1}] \\
&+ \frac{v^2}{2}\,\mathbb{E}\left[\left(\frac{1}{c^\star} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c^\star} - \gamma_k\right)^2\right] \\
&+ \frac{c_{\max}}{c_{\min}}\left[H\frac{\log A}{\eta} + \frac{p}{2}v(1 + \alpha_1 + \alpha_2)H_{\max}\right] \\
&+ \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} + \frac{\big(Hr_{\max} + (Hc_{\max} - Hc_{\min})(\alpha_1 + \alpha_2)\big)^2}{2\gamma_{\min}^2},
\end{aligned}$$

which proves the lemma. □

Lemma 16 is the key structural inequality in the proof. After summing it over episodes, the Lyapunov drift, queue increment, and auxiliary quadratic term all telescope, leaving only boundary contributions together with the Softmax approximation and failure-event terms.

E Final synthesis and proof of Theorem 1

In this section, we combine the structural surrogate bound from the Lyapunov analysis with the cumulative estimation error bound from Appendix C and complete the proof of Theorem 1.

E.1 Structural surrogate bound

We first sum the one-step drift-plus-penalty inequality and isolate the queue-adjusted surrogate term. This yields the structural contribution of the control layer before replacing the optimistic and pessimistic estimates by the true value functions.

Lemma 17 (Structural surrogate bound)**.** *Let* $\theta^\star := \sup_{(r,c)\in\bar{\Gamma}} \frac{r}{c}$*, and let* $(r^\star, c^\star) \in \bar{\Gamma}$ *satisfy* $\theta^\star = r^\star/c^\star$*. Then*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\hat{V}_1^k(m_k, s_1^k)\Big] \geq \theta^\star \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big[\check{W}_1^k(m_k, s_1^k)\big] - \mathcal{O}\bigg(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\bigg).$$

where $A := |\mathcal{A}|$.

Proof. For brevity, throughout this proof we write

$$\hat{V}_1^k := \hat{V}_1^k(m_k, s_1^k), \qquad \check{W}_1^k := \check{W}_1^k(m_k, s_1^k).$$

We start from Lemma 16. For every episode $k = 2, 3, \ldots, K$, it gives

$$\begin{aligned}\mathbb{E}\Big[\Delta_k - v\hat{V}_1^k\Big] \leq\ & b - v\theta^\star\,\mathbb{E}\big[\check{W}_1^{k-1}\big] + v\theta^\star\,\mathbb{E}[J_k - J_{k-1}] \\ & + \frac{v^2}{2}\,\mathbb{E}\left[\left(\frac{1}{c^\star} - \gamma_{k-1}\right)^2 - \left(\frac{1}{c^\star} - \gamma_k\right)^2\right] \\ & + \frac{c_{\max}}{c_{\min}}\left[H\frac{\log A}{\eta} + \frac{p}{2}v(1+\alpha_1+\alpha_2)H_{\max}\right] \\ & + \frac{c_{\max}}{c_{\min}}\frac{\log A}{\eta} + \frac{\big(Hr_{\max} + (Hc_{\max} - Hc_{\min})(\alpha_1+\alpha_2)\big)^2}{2\gamma_{\min}^2}.\end{aligned}$$

Since the cost term is delayed, we sum the above inequality over $k = 2, 3, \ldots, K$ and divide both sides by Kv.

For the Lyapunov drift,

$$\frac{1}{Kv}\sum_{k=2}^{K}\mathbb{E}[\Delta_k] = \frac{1}{Kv}\mathbb{E}[L_{K+1} - L_2] \geq -\frac{1}{Kv}\mathbb{E}[L_2].$$

Since $L_2 = \frac{1}{2}J_2^2 \leq \frac{1}{2}v^2(\alpha_1+\alpha_2)^2$, we have

$$\frac{1}{Kv}\sum_{k=2}^{K}\mathbb{E}[\Delta_k] \geq -\frac{v}{2K}(\alpha_1+\alpha_2)^2.$$

For the optimistic reward term,

$$-\frac{1}{K}\sum_{k=2}^{K}\mathbb{E}[\hat{V}_1^k] = -\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\hat{V}_1^k] + \frac{1}{K}\mathbb{E}[\hat{V}_1^1] \geq -\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\hat{V}_1^k],$$

where we used $\hat{V}_1^1 \geq 0$.

For the delayed pessimistic cost term,

$$-\frac{\theta^\star}{K}\sum_{k=2}^{K}\mathbb{E}[\check{W}_1^{k-1}] = -\frac{\theta^\star}{K}\sum_{k=1}^{K-1}\mathbb{E}[\check{W}_1^k] = -\frac{\theta^\star}{K}\sum_{k=1}^{K}\mathbb{E}[\check{W}_1^k] + \frac{\theta^\star}{K}\mathbb{E}[\check{W}_1^K].$$

Since $0 \le \check{W}_1^K \le H_{\max}$, this implies

$$-\frac{\theta^\star}{K}\sum_{k=2}^{K}\mathbb{E}[\check{W}_1^{k-1}] \le -\frac{\theta^\star}{K}\sum_{k=1}^{K}\mathbb{E}[\check{W}_1^{k}] + \frac{\theta^\star H_{\max}}{K}.$$

For the queue difference term,

$$\frac{\theta^\star}{K}\sum_{k=2}^{K}\mathbb{E}[J_k - J_{k-1}] = \frac{\theta^\star}{K}\mathbb{E}[J_K - J_1] \le \frac{\theta^\star v(\alpha_1+\alpha_2)}{K},$$

since $J_1 = 0$ and $J_K \le v(\alpha_1+\alpha_2)$.

For the auxiliary-variable term,

$$\begin{aligned}\frac{v}{2K}\sum_{k=2}^{K}\mathbb{E}\left[\left(\frac{1}{c^\star}-\gamma_{k-1}\right)^2 - \left(\frac{1}{c^\star}-\gamma_k\right)^2\right] &= \frac{v}{2K}\mathbb{E}\left[\left(\frac{1}{c^\star}-\gamma_1\right)^2 - \left(\frac{1}{c^\star}-\gamma_K\right)^2\right] \\ &\le \frac{v}{2K}(\gamma_{\max}-\gamma_{\min})^2.\end{aligned}$$

Rearranging the summed inequality yields

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\hat{V}_1^k\right] \ge \theta^\star\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\check{W}_1^k\right] - \frac{\theta^\star H_{\max}}{K} - \frac{g_1}{v} - \frac{vg_2}{K} - g_3, \tag{25}$$

where

$$g_1 := b + \frac{\left(Hr_{\max} + (Hc_{\max} - Hc_{\min})(\alpha_1+\alpha_2)\right)^2}{2\gamma_{\min}^2},$$

$$g_2 := \frac{1}{2}(\gamma_{\max}-\gamma_{\min})^2 + \theta^\star(\alpha_1+\alpha_2) + \frac{1}{2}(\alpha_1+\alpha_2)^2,$$

and

$$g_3 := \frac{c_{\max}}{c_{\min}}\left[\frac{2H\log A}{v\eta} + \frac{p}{2}(1+\alpha_1+\alpha_2)H_{\max}\right].$$

Therefore,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\hat{V}_1^k\right] \ge \theta^\star\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\check{W}_1^k\right] - \mathcal{O}\left(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\right).$$

□

E.2 Final regret bound

We now replace the optimistic reward estimate and pessimistic cost estimate in Lemma 17 by the true value functions using the cumulative estimation error bound.

Lemma 18 (Final regret bound)**.** *Let the trade-off parameter be $v = H^2\sqrt{K}$, the Softmax parameter be $\eta = \sqrt{K}$, and the failure probability be $p = 1/K$. Let $\iota := \log\left(\frac{2MdTA}{p}\right)$. Let $\theta^\star := \sup_{(r,c)\in\bar{\Gamma}}\frac{r}{c}$, and let $(r^\star, c^\star) \in \bar{\Gamma}$ be an optimal performance pair such that $\theta^\star = r^\star/c^\star$. Then the performance ratio of our algorithm satisfies*

$$\frac{\sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi^k}(m_k, s_1^k)\right]}{\sum_{k=1}^{K}\mathbb{E}\left[W_1^{\pi^k}(m_k, s_1^k)\right]} \ge \theta^\star - \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Proof. For brevity, write

$$\hat{V}_1^k := \hat{V}_1^k(m_k, s_1^k), \qquad \check{W}_1^k := \check{W}_1^k(m_k, s_1^k),$$

and similarly

$$V_1^{\pi^k} := V_1^{\pi^k}(m_k, s_1^k), \qquad W_1^{\pi^k} := W_1^{\pi^k}(m_k, s_1^k).$$

By Lemma 17,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\hat{V}_1^k\Big] \geq \theta^\star \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big[\check{W}_1^k\big] - \mathcal{O}\bigg(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\bigg).$$

By Lemma 7,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\hat{V}_1^k] \leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[V_1^{\pi^k}] + \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right),$$

and

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\check{W}_1^k] \geq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}] - \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Combining the above inequalities yields

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[V_1^{\pi^k}] \geq \theta^\star \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}] - \mathcal{O}\bigg(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\bigg) - \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Since $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}] \geq Hc_{\min} > 0$, dividing both sides by $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}]$ gives

$$\frac{\sum_{k=1}^{K}\mathbb{E}[V_1^{\pi^k}]}{\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}]} \geq \theta^\star - \mathcal{O}\bigg(\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp\bigg) - \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Finally, setting $v = H^2\sqrt{K}$, $\eta = \sqrt{K}$, and $p = 1/K$ gives

$$\frac{H^4}{v} + \frac{v}{K} + \frac{H\log A}{v\eta} + Hp = O\bigg(\frac{H^2}{\sqrt{K}} + \frac{\log A}{HK} + \frac{H}{K}\bigg).$$

Recall that $\iota = \log\Big(\frac{2MdTA}{p}\Big)$, so $\log A \leq \iota$. Therefore the structural term is dominated by the statistical term, we have,

$$\frac{\sum_{k=1}^{K}\mathbb{E}[V_1^{\pi^k}]}{\sum_{k=1}^{K}\mathbb{E}[W_1^{\pi^k}]} \geq \theta^\star - \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right),$$

which proves the lemma. □

Proof of Theorem 1. Lemma 18 implies that

$$\mathrm{Gap}(K) \leq \mathcal{O}\left(\frac{\sqrt{Md^3H^3T\,\iota^2}}{K}\right).$$

Since $\mathcal{R}(K) = K\,\mathrm{Gap}(K)$, it follows that

$$\mathcal{R}(K) \leq \mathcal{O}\Big(\sqrt{Md^3H^3T\,\iota^2}\Big).$$

□

F Construction of the experimental environments

This appendix describes the construction of the two experimental environments used in our evaluation: a synthetic multi-task linear MDP and an empirical multi-task finite-horizon MDP built from KuaiRand interaction logs (Gao et al., 2022).

F.1 Synthetic linear MDP

We use a synthetic multi-task episodic linear MDP as a controlled testbed for the experiments in the main text. The environment is designed to match the structural assumptions of our model. In each episode, a task index $m \in \{1, \dots, M\}$ is drawn uniformly at random, the initial state is fixed to $s_1 = 0$, and the interaction lasts for $H = 5$ stages. We set $M = 4$, $|\mathcal{S}| = 20$, $A = 10$, and $d = 6$.

The environment is generated from a shared feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ together with task- and stage-specific linear parameters. Each feature vector $\phi(s, a)$ is sampled once at initialization from $\mathrm{Dirichlet}(\mathbf{1}_d)$. For each task m and stage h, we sample a transition measure $\mu_h(m, \cdot) \in \mathbb{R}^{d \times |\mathcal{S}|}$, with each row drawn as a probability vector, and define $P_h(s' \mid m, s, a) = \phi(s, a)^\top \mu_h(m, s')$. Since $\phi(s, a)$ is also a probability vector, this defines a valid transition distribution. The one-step reward and cost are given by $r_h(m, s, a) = \phi(s, a)^\top \theta_{r,h}(m)$ and $c_h(m, s, a) = \phi(s, a)^\top \theta_{c,h}(m)$, where each coordinate of $\theta_{r,h}(m)$ is sampled independently from $\mathrm{Uniform}(0, 1)$, and each coordinate of $\theta_{c,h}(m)$ is sampled independently from $\mathrm{Uniform}(0.01, 0.5)$.

This construction yields a controlled multi-task linear environment with a shared state-action representation and task-dependent reward, cost, and transition models. In the experiments, all compared methods are evaluated on the same generated environment for $K = 15000$ episodes.

Implementation details and hyperparameters. The 20 independent trials were run with random seeds $42, 43, \dots, 61$. Hyperparameters were selected by a small pilot tuning procedure on separate random seeds. For COLD, we tuned $v \in \{1, 5, 10, 20\}$, $\eta \in \{10, 100, 1000\}$, $\lambda \in \{0.1, 1, 10\}$, and $\beta \in \{0.1, 0.5, 1.0\}$, and used $v = 10$, $\eta = 1000$, $\lambda = 1$, and $\beta = 0.5$ in the reported experiments. The Lyapunov variables were initialized with $J_1 = 0$, $\gamma_{\min} = 1/(Hc_{\max}) = 0.4$, and $\gamma_{\max} = 1/(Hc_{\min}) = 20.0$, corresponding to the initial cost target $1/\gamma_{\min} = 2.5$. For RM^{2}VI, we tuned the Robbins–Monro stepsize scale $\alpha_\eta \in \{0.1, 0.5, 1.0\}$, where $\eta = \alpha_\eta/(c_{\min}\sqrt{K})$, and the bonus scale $\alpha_b \in \{0.001, 0.005, 0.01, 0.05\}$, where $b_n = \alpha_b H\sqrt{\log(SAHK)/n}$. The final configuration was $\alpha_\eta = 1.0$ and $\alpha_b = 0.01$. For Ratio-LSVI-UCB, we used the same λ and β as COLD for a fair comparison.

F.2 Empirical KuaiRand MDP

This section describes the empirical multi-task finite-horizon MDP used in our KuaiRand experiments. Starting from the KuaiRand interaction logs (Gao et al., 2022), specifically the KuaiRand-1K data released under the CC BY-SA 4.0 license, we first define task, state, action, reward, and cost variables at the interaction level, then extract fixed-horizon trajectories, and finally estimate an empirical multi-task finite-horizon MDP $(\nu^{\mathrm{emp}}, P^{\mathrm{emp}}, R^{\mathrm{emp}}, C^{\mathrm{emp}})$, where ν^{emp} denotes the empirical initial-state distribution, and $P_h^{\mathrm{emp}}, R_h^{\mathrm{emp}}, C_h^{\mathrm{emp}}$ are the stage-dependent empirical transition, reward, and cost functions. We also specify the shared feature map $\phi(s, a)$ used by the linear methods on this empirical environment.

Task, state, and action spaces. We define the task space using user activity groups. After merging the interaction log with user activity metadata, each retained interaction is assigned to one of the following seven task types:

$$\mathcal{M} = \{\texttt{2_14_day_new}, \texttt{30day_retention}, \texttt{single_low_active}, \texttt{low_active}, \texttt{middle_active}, \texttt{high_active}, \texttt{full_active}\}.$$

Thus, each trajectory is associated with a task index $m \in \{0, \dots, M - 1\}$.

To define the state space, we map each interaction to a discrete engagement state. We first construct a coarse feedback state with five levels:

$$\begin{aligned} &0 : \text{weak or no feedback},\\ &1 : \text{click},\\ &2 : \text{long view},\\ &3 : \text{like or comment},\\ &4 : \text{follow or profile enter}. \end{aligned}$$

The assignment is hierarchical: if multiple feedback conditions are satisfied, the interaction is mapped to the highest-priority one. We then compute the play ratio

$$\rho = \frac{\texttt{play_time_ms}}{\texttt{duration_ms}},$$

clip it to $[0, 1]$, and quantize it into three buckets. The final state is defined by combining the coarse feedback level with the play-ratio bucket, yielding a total of 15 discrete states:

$$\mathcal{S} = \{0, 1, \ldots, 14\}.$$

Equivalently, the state index can be written as

$$s = 3\, s_{\text{coarse}} + \kappa(\rho),$$

where $\kappa(\rho) \in \{0, 1, 2\}$ denotes the play-ratio bucket.

The action space is defined using video content categories. After merging category metadata into the interaction log, we retain the top-8 most frequent first-level video categories and map them to action indices. This gives

$$\mathcal{A} = \{0, 1, \ldots, 7\}.$$

Hence, each action corresponds to recommending a video from one of the selected top-level categories.

Reward and cost definitions. Let $\bar{d}$ denote the median positive video duration in the merged log. We use $\bar{d}$ as a common normalization scale for both reward and cost. The base watchtime reward is defined as

$$r_i^{\text{wt}} = \max\left(\frac{\texttt{play_time_ms}_i}{\bar{d}},\, 0\right),$$

and the base duration cost is

$$c_i^{\text{dur}} = \max\left(\frac{\texttt{duration_ms}_i}{\bar{d}},\, \varepsilon_c\right), \qquad \varepsilon_c = 10^{-3}.$$

To reduce the effect of heavy tails, both quantities are optionally truncated at high empirical quantiles.

We consider three reward variants in the main paper.

Watchtime-only reward. The first variant uses only normalized watchtime with a 95th-percentile cap:

$$r_i = \min\left(r_i^{\text{wt}},\, q_{0.95}(r^{\text{wt}})\right).$$

Uniform engagement bonus. The second variant uses the same capped watchtime base and the same overall bonus budget, but distributes the bonus uniformly across active engagement events:

$$r_i = \min\left(r_i^{\text{wt}},\, q_{0.95}(r^{\text{wt}})\right) + \sum_{e \in \mathcal{E}} \beta_e^{\text{uni}} \mathbf{1}\{e \text{ occurs on interaction } i\}.$$

Rarity-aware engagement bonus. The third variant adds an engagement bonus to the capped watchtime reward. Let

$$\mathcal{E} = \{\texttt{click}, \texttt{like}, \texttt{comment}, \texttt{follow}\}$$

denote the set of engagement events, and let π_e be the empirical occurrence rate of event $e \in \mathcal{E}$. The total bonus budget is set proportional to the scale of the watchtime reward, and rarer events are assigned larger coefficients. The resulting reward takes the form

$$r_i = \min\left(r_i^{\text{wt}},\, q_{0.95}(r^{\text{wt}})\right) + \sum_{e \in \mathcal{E}} \beta_e \mathbf{1}\{e \text{ occurs on interaction } i\},$$

where β_e increases as the event becomes rarer.

Trajectory extraction. After assigning task, state, action, reward, and cost labels to each interaction, we group interactions by user and extract fixed-horizon trajectories by a sliding-window procedure with horizon $H = 5$. For a user sequence $g = (x_0, x_1, \ldots, x_{T_u - 1})$, each valid window defines one episode. If x_t is the current interaction, then the state at stage h is taken from the current interaction, while the action, reward, cost, and next state are taken from the next interaction in the window. Therefore, each stage contributes a transition of the form

$$(h, m, s, a, r, c, s').$$

This construction yields both a collection of initial states and a set of stage-labeled transitions from which the empirical MDP is estimated.

Empirical MDP estimation. From the extracted trajectories, we estimate a stage-dependent empirical MDP

$$\mathcal{M}_{\mathrm{emp}} = (\mathcal{M}, \mathcal{S}, \mathcal{A}, H, \nu^{\mathrm{emp}}, P^{\mathrm{emp}}, R^{\mathrm{emp}}, C^{\mathrm{emp}}).$$

The initial distribution $\nu^{\mathrm{emp}}(m)$ is obtained by normalizing the empirical frequencies of initial states for each task m. For each stage h, task m, state s, and action a, we estimate the transition kernel $P_h(\cdot \mid m, s, a)$, reward $R_h(m, s, a)$, and cost $C_h(m, s, a)$ from the corresponding empirical transition counts and sample averages. When some tuples are sparsely observed, we use smoothed empirical estimates based on broader task-level or global statistics so that the resulting tensors remain dense.

The final empirical MDP is represented by

$$\nu^{\mathrm{emp}} \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{S}|}, \qquad P^{\mathrm{emp}} \in \mathbb{R}^{H\times|\mathcal{M}|\times|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{S}|}, R^{\mathrm{emp}}, C^{\mathrm{emp}} \in \mathbb{R}^{H\times|\mathcal{M}|\times|\mathcal{S}|\times|\mathcal{A}|}.$$

These objects are then used as the common evaluation environment for all compared methods.

Feature map for linear methods. In the KuaiRand experiments, both COLD and Ratio-LSVI-UCB use the same structured feature map $\phi(s, a)$. For a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}$, we define

$$\phi(s, a) = \big[1,\ \mathrm{onehot}(s),\ \mathrm{onehot}(a),\ \mathrm{onehot}(s) \otimes \mathrm{onehot}(a)\big] \in \mathbb{R}^{\,1+|\mathcal{S}|+|\mathcal{A}|+|\mathcal{S}||\mathcal{A}|},$$

where $\mathrm{onehot}(s) \in \mathbb{R}^{|\mathcal{S}|}$ and $\mathrm{onehot}(a) \in \mathbb{R}^{|\mathcal{A}|}$ denote the one-hot encodings of the state and action, respectively, and $\otimes$ is the Kronecker product. Hence, the feature dimension is

$$d = 1 + |\mathcal{S}| + |\mathcal{A}| + |\mathcal{S}||\mathcal{A}|.$$

This feature map contains a global bias term, separate state and action effects, and explicit state-action interaction terms. As a result, it provides a structured but tabular-style linear representation, making the empirical setting relatively favorable to tabular methods while still allowing the linear algorithms to operate in a shared feature space.

Implementation details and hyperparameters. The empirical KuaiRand MDP was estimated from 20,682 user episodes and 103,410 stage-labeled transitions. Results were averaged over 20 independent runs with seeds $42, 43, \ldots, 61$. Hyperparameters were selected by a small pilot tuning procedure on separate random seeds using the watch time only setting, and then fixed for all three reward variants: watch time only, watch time plus uniform engagement bonus, and watch time plus weighted engagement bonus. In the watch-time-only setting, the 95th-percentile-capped reward and cost ranges were $r_{\min} = 0$, $r_{\max} = 0.3297$, $c_{\min} = 0.001$, and $c_{\max} = 3.6794$. For COLD, we tuned $v \in \{10, 30, 50, 80, 100\}$, $\eta \in \{10, 50, 100, 200\}$, $\lambda \in \{0.001, 0.01, 0.03, 0.1, 1\}$, and $\beta \in \{0.001, 0.005, 0.01, 0.02, 0.05\}$, and used $v = 80$, $\eta = 100$, $\lambda = 0.01$, and $\beta = 0.02$. We initialized $J_1 = 0$, $\gamma_{\min} = 1/(Hc_{\max}) \approx 0.0544$, and $\gamma_{\max} = 1/(Hc_{\min}) = 200$. For RM^{2}VI, we tuned $\alpha_\eta \in \{0.001, 0.005, 0.01, 0.05\}$ in $\eta = \alpha_\eta/(c_{\min}\sqrt{K})$ and $\alpha_b \in \{0.0005, 0.001, 0.005, 0.01\}$ in $b_n = \alpha_b H\sqrt{\log(SAHK)/n}$, and used $\alpha_\eta = 0.01$ and $\alpha_b = 0.001$. For Ratio-LSVI-UCB, we used the same structured feature map as COLD and set $\lambda = 0.01$ and $\beta = 0.02$ for a fair comparison.

Compute setting. All experiments were run on CPU using an Intel Xeon Gold 5318Y processor.

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: Claims in the abstract and introduction accurately reflect our paper's contributions and scope.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We discuss the limitations of our results after Theorem 1, including the current restriction to finite action spaces, stationary task distributions, and non-adversarial task arrivals.

   Guidelines:

   - The answer [N/A] 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: We provide a proof sketch of our theorem in the main paper and complete the full proof in the appendix.

Guidelines:

- The answer [N/A] 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: We disclose all the information needed to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] 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: We provide anonymized code and reproduction instructions in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the experimental settings in the supplemental material.

   Guidelines:

   - The answer [N/A] 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: We report error bars correctly in the paper.

   Guidelines:

   - The answer [N/A] 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: We provide the sufficient information on the computer resources.

   Guidelines:

   - The answer [N/A] 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: The research conducted in this paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] 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: We cite the original paper and abide by the license.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: We provide anonymized code for the proposed algorithm, baselines, synthetic environment generation, and KuaiRand-based empirical MDP construction in the supplementary material.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not conduct new crowdsourcing experiments or new human-subject studies.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve new human-subject data collection or intervention.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as part of the proposed method, theoretical analysis, or experimental pipeline.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.