informs
https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1523-4614, EISSN 1526-5498

Submitted to *Manufacturing & Service Operations Management*

Reinforcement Learning in MDPs with Information-Ordered Policies

(Authors' names are not included for peer review)

Abstract. Problem definition: We introduce an epoch-based reinforcement learning algorithm for infinite-horizon average-cost MDPs that exploits a partial order over a policy class. In this structure, $\pi_{\theta'} \preceq \pi_\theta$ if data collected under π_θ can be used to estimate the performance of $\pi_{\theta'}$, enabling counterfactual inference without additional environment interaction. **Methodology/results:** Leveraging this partial order, we show that our algorithm achieves a (nearly matching) regret bound of $\tilde{O}(\sqrt{w \log(|\Theta|) T})$, where w is the width of the partial order. Notably, the bound is independent of the state and action space sizes. **Managerial insights:** We illustrate the applicability of these partial orders in many domains in operations research, including inventory control and queuing systems. For each, we apply our framework to that problem, yielding new theoretical guarantees and strong empirical results without imposing extra assumptions such as convexity in the inventory model or specialized arrival-rate structure in the queuing model. Our results suggest that, in sequential decision-making within operations research, the width of the order in the corresponding operational problems, rather than the conventionally emphasized size of the state and action spaces, constitutes the primary source of problem complexity.

Key words: Reinforcement learning, Exogenous MDPs, Feedback Structure, Inventory control

1. Introduction

Recent breakthroughs in reinforcement learning (RL) have demonstrated superhuman-level performance in complex board and video games, such as Go (Silver et al. 2016, 2017), Atari (Mnih et al. 2015), and StarCraft (Vinyals et al. 2017, 2019). While these advances open exciting new avenues for controlling *complex* and *unknown* systems, their success relied on access to millions or even billions of data points (Vinyals et al. 2017, 2019), collected via interactions with high-fidelity simulators such as the StarCraft game engine. However, despite this progress, the deployment of RL in real-world operational systems remains nascent. While there have been promising developments in inventory management (Madeka et al. 2022, Alvo et al. 2023, Eisenach et al. 2024), queuing systems (Anselmi et al. 2022), supply chain optimization (Chong et al. 2022, Rolf et al. 2023), and algorithmic pricing (Bertsimas and Vayanos 2017, Zhang et al. 2022), wide-scale adoption has been hindered by persistent challenges related to the *required sample complexity*.

A recent approach to bridge the gap between designing algorithms for operational systems is to leverage *side observations* (i.e., rich feedback) to improve the learning efficiency of RL. For example, consider the tabular setting with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. When no side observations are available, it is well known that the regret scales polynomially with respect to $|\mathcal{S}|$ and $|\mathcal{A}|$ (Auer et al. 2008). On the other end, under full-feedback settings where the transition dynamics and rewards of all state-action pairs can be accessed simultaneously (Dann et al. 2020, Sinclair et al. 2023b, Wan et al. 2024), regret no longer depends on $|\mathcal{S}|$ and $|\mathcal{A}|$. Unfortunately, the full-feedback assumption is rarely satisfied in real-world contexts. However, many domains offer partial feedback that, while limited, can still be systematically leveraged to yield gains in learning efficiency.

Motivation from inventory control. As a concrete example, let us briefly consider the classical perishable newsvendor model from supply chain management (Goldberg et al. 2021). Consider operations in a retail store where the goal is to manage inventory (products in the store) by making ordering decisions subject to stochastic demand. The seller decides on an ordering quantity q, faces i.i.d. demand D, and incurs a (random) cost $C(q) = h(q-D)^+ + p(D-q)^+$, where $(\cdot)^+ = \max\{\cdot, 0\}$ and h and p are the holding and lost-sales cost coefficients respectively. If the demand D is fully observed, this reduces to the *full-feedback* setting (the algorithm can use data collected from any ordering level to estimate $C(q)\ \forall q$), and existing work leverages this structure for improved sample-complexity guarantees (Levi et al. 2007). However in practice, one only observes *sales* $N(q) = \min\{q, D\}$ instead of the true demand D (Besbes and Muharremoglu 2013, Hssaine and Sinclair 2024). Our insight is that the lost-sales feedback induces a natural *order* over ordering levels q, since observed sales $N(q)$ can be deterministically used to infer $N(q')$ for any $q' \leq q$. This structure induces an ordering $\pi_{q'} \preceq \pi_q$ over policies (characterized by their ordering level) with *width* one (Dushnik and Miller 1941). Although the policy ordering $\pi_{q'} \preceq \pi_q$ is trivial since the MDP has only a single state, we will show later that this measurability-based policy order extends to more complex MDPs, including inventory models with positive lead times, enabling the analysis of many real-world operations research problems.

This idea of leveraging partial feedback is intuitive and has been explored in simple models under "one-sided" feedback over actions (Zhao and Chen 2019, Gong and Simchi-Levi 2024). However, prior approaches have been ad-hoc and tailored to specific domains. In contrast, we provide a unified framework for modeling partial feedback directly over the policy class via a partial ordering. This shift not only generalizes prior work, but also allows us to systematically harness partial feedback in far more complex settings, such as inventory systems with positive lead times and multiple suppliers, as well as queuing systems. In particular, we seek to answer the following key research questions:

How can we formalize partial feedback over policies in general MDPs? Can we leverage them to improve sample efficiency? What real-world examples have natural policy classes exhibiting this structure?

1.1. Contributions

We answer these questions affirmatively by introducing the notion of an *information order* over policies grounded in σ-algebra measurability (Definitions 2 and 3). The primary contributions of this paper are as follows:

Policy Ordering Based on Information Structures. We propose the concept of policy ordering as a complexity measure for policy classes in general MDPs. We consider a policy class parameterized by $\theta \in \Theta$ and for two policies π_θ and $\pi_{\theta'}$, we write $\pi_{\theta'} \preceq \pi_\theta$ if an "estimate" for the long-run average cost under $\pi_{\theta'}$ is measurable with respect to a trajectory generated by π_θ. We consider three separate cases in which $\pi_{\theta'} \preceq \pi_\theta$. First, the *Sample-Path Policy Order* (Definition 2) holds when the *empirical average cost* (Definition 1) under $\pi_{\theta'}$ can be computed directly from a trajectory generated by π_θ, which typically occurs when the counterfactual trajectory under $\pi_{\theta'}$ is fully recoverable. Second, the *Distributional Policy Order* (Definition 3), which is a generalization of the Sample-Path Policy Order, applies when the empirical average cost under $\pi_{\theta'}$ is not directly obtainable, but there exists a random variable measurable with respect to the π_θ trajectory, whose distribution lies within a small total variation distance of the empirical average cost under $\pi_{\theta'}$. In Section 4.3 we further show a distributional policy order based on analytical estimates (e.g., via the stationary distribution) rather than the empirical average cost. Thus, policy ordering can be defined based on any type of estimator (exact or approximate, based on empirical average cost or not), and the corresponding regret bound can be obtained accordingly.

Regret Bounds Independent of State and Action Space Sizes. Given such a (partial) order of width w, our main result (Theorem 1) shows that the regret is bounded by $\tilde{O}(\sqrt{w \log(|\Theta|)T})$ (with extensions to continuous Θ). Importantly, this result is independent of $|\mathcal{S}|$ and $|\mathcal{A}|$ (which may be infinite when Θ is continuous), and measures the statistical complexity of learning based on the *width* of the partial order. We later emphasize how this result bridges the gap between the no-feedback ($w = |\Theta|$) and full-feedback ($w = 1$) regimes (Sinclair et al. 2023b). We also establish in Theorem 2 and Theorem C.1 that the bound in Theorem 1 is nearly tight, with the gap between upper and lower bounds matching that of classical rich-feedback bandit problems (Auer et al. 2008, Osband and Van Roy 2016). It is worth noting that our main result in Theorem 1 depends only on measurability-based policy ordering and mild assumptions on the MDP, without requiring strong structural properties on the long-run average cost function such as convexity.

The improved regret upper bound in Theorem 1 arises from a more structured exploration-exploitation tradeoff. Although Algorithm 1 follows a conventional UCB-style approach, exploration is no longer uniform over the policy class. Instead, the policy partial order actively guides exploration toward a small set of maximal policies (see Section 2.1.3) whose trajectories are sufficiently informative to counterfactually estimate the performance of all dominated policies. This structure effectively concentrates exploration where it is most valuable, enabling systematic exploitation of shared information and dramatically shrinking the

search space. Conceptually, this principle parallels Information-Directed Sampling (Russo and Van Roy 2018), which also tailors exploration to information gain via a "regret-per-bit" criterion, but differs in two key ways: it focuses on Bayesian regret and requires online estimation of information gains. In contrast, our policy-order framework delivers worst-case $\tilde{O}(\sqrt{T})$ regret directly, at the cost of requiring an explicit low-width ordering to be constructed for each problem.

Application in Real-World MDPs. In Section 4, we illustrate the versatility of policy orders and Theorem 1 by applying it to several canonical problems in operations research, including inventory control with positive lead times and lost-sales (Goldberg et al. 2021), lost-sales dual-sourcing systems governed by dual index policies (Whittemore and Saunders 1977), and queuing models with state-dependent service rates (Puterman 2014). In each case, we identify natural policy classes that admit an information order with low width, yielding new algorithms and regret bounds.

In particular, the dual index policy case study highlights that our framework captures structural properties of MDPs that go beyond those traditionally explored in the literature, such as convexity (Agrawal and Jia 2022, Gong and Simchi-Levi 2024, Chen and Shi 2025) and feedback graph connectivity (Dann et al. 2020). Notably, the lost-sales dual index policy setting lacks convexity (Veeraraghavan and Scheller-Wolf 2008), and applying existing learning algorithms yields regret of order $\tilde{O}(T^{2/3})$, whereas our approach achieves $\tilde{O}(\sqrt{T})$ regret. We also obtain an $\tilde{O}(\sqrt{T})$ regret bound for the queuing problem studied in Section 4.3, which is independent of the action space size without any additional structural assumptions on the system dynamics. In contrast, the $\tilde{O}(\sqrt{T})$ regret bound in Anselmi et al. (2022), although independent of the state space size, depends critically on assuming a state-dependent decay in the arrival rate to limit visits to high-occupancy states. Moreover, our approach extends naturally to more complex scenarios, such as state-dependent arrival rates in queuing systems.

We also perform extensive numerical simulations on all three case studies in Section 5, and show that our algorithm (Algorithm 1) consistently matches or exceeds the performance of specialized, problem-specific baselines. These results confirm the practical effectiveness and scalability of our policy-order-based approach.

1.2. Literature Review

Reinforcement Learning Algorithms for Tabular MDPs. Reinforcement learning algorithms for tabular (discrete) MDPs has been extensively studied in both model-based (Auer et al. 2008, Osband et al. 2016) and model-free settings (Strehl et al. 2006, Jin et al. 2018, Wei et al. 2020). In the sequential interaction setting (with finite horizon H) without access to a generative model, the best-known regret scales as $\tilde{O}(H^{3/2}\sqrt{|\mathcal{S}||\mathcal{A}|T})$ in both the model-based (Azar et al. 2017) and model-free (Jin et al. 2018) regimes, which becomes prohibitive as the state and action spaces grow. These challenges have motivated a growing body of research that exploits problem-specific structure to enhance learning efficiency, including linear

function approximation (Wang et al. 2019, Jin et al. 2020), Lipschitz continuity (Sinclair et al. 2023a), Eluder dimension (Osband and Van Roy 2014), and convexity (Agrawal and Jia 2022). In contrast, we explore a new type of structural assumption: information orders over the policy class.

MDPs with Rich Feedback. The ordering structure we investigate corresponds to the broader notion of rich feedback in both bandit and RL problems. In the bandit literature, rich feedback has been extensively analyzed (Mannor and Shamir 2011, Alon et al. 2015), with foundational results on feedback graphs and side-observations. In contrast, for RL and MDPs, the study of rich feedback remains relatively nascent. Dann et al. (2020) analyze the regret of a confidence set elimination algorithm under general feedback graphs, and show that in tabular MDPs, certain structural properties of the graph can be used to obtain regret bounds that do not depend on $|\mathcal{S}|$ and $|\mathcal{A}|$. While the setting studied in Dann et al. (2020) is similar to ours, their method yields a regret upper bound of $\tilde{O}(T^{2/3})$ in several important operations research domains due to the tabular assumption, including the lost-sales inventory control and the dual sourcing problem we study (see Section 4 for details). In contrast, our approach directly applies to continuous state-action MDPs and achieves a tighter bound of $\tilde{O}(\sqrt{T})$ in these settings. Furthermore, Wan et al. (2024) analyze two extreme feedback regimes in exogenous MDPs—no feedback and full feedback (their "no observation" and "full observation"), and point out a $\sqrt{d}$ gap in the regret of these two cases, where d is the size of the exogenous state space. Our work differs from Wan et al. (2024) in two key respects. First, we handle general MDPs rather than only exogenous ones, thus their analysis, which is based on linear mixture MDPs, would fail; Second, we introduce a partial-order framework that interpolates between these extremes, yielding greater generality than either no-feedback or full-feedback analyses.

MDPs with One-Sided Information. Our concept of policy order is motivated by (and indeed, builds upon) one-sided information, which represents a special case of rich feedback. Zhao and Chen (2019) and Yuan et al. (2021) investigate one-sided information in the context of bandit problems. Similarly, Gong and Simchi-Levi (2020) study a general Q-learning algorithm under both full-feedback and one-sided-feedback settings, establishing a learning framework that achieves $\tilde{O}(\sqrt{T})$ regret independent of the state and action spaces. They later instantiate this framework in a zero lead-time inventory problem with cyclic demand. The information orders in Gong and Simchi-Levi (2020, 2024) are defined at the action-space level, and do not apply to inventory control models with positive lead times. We extend this idea to define information order over *policies* instead of *actions*, and show how our results apply to a larger class of inventory models.

Algorithms Leveraging Problem-Specific Structure. From an application perspective, several prior works have investigated leveraging problem-specific structure for real-world MDPs. A closely related work is Agrawal and Jia (2022), which establishes a regret bound of $\tilde{O}(\sqrt{T})$ for the same lost-sales inventory problem we consider in Section 4.1 by exploiting convexity. However, this convexity-based approach cannot be extended to the dual-sourcing problem in Section 4.2, since dual index policies lack the requisite

convexity property (Veeraraghavan and Scheller-Wolf 2008). In contrast, our policy-ordering framework applies naturally to both problems. Chen and Shi (2025) derive an $\tilde{O}(\sqrt{T})$ bound for tailored base-surge policies in the dual-sourcing setting by relying on convexity over tailored base-surge policies, an assumption that fails for dual index policies and thus prevents a direct extension of their method. Tang et al. (2024) analyze the dual index policy under the backlog formulation, which is a simpler setting with full feedback compared to the lost-sales setting considered in this paper.

Note that for real-world MDPs with finite state and action spaces, standard RL algorithms apply directly (Strehl et al. 2006, Auer et al. 2008, Osband et al. 2016, Jin et al. 2018, Wei et al. 2020, Khodadadian et al. 2021, Chen et al. 2021). Anselmi et al. (2022) study the service-rate control queuing problem in Section 4.3, using a model-based estimator that exploits problem-specific structure to obtain an $\tilde{O}(\sqrt{T})$ regret bound that is independent of the state space size, assuming a state-dependent decay in the arrival rate. In contrast, our policy-ordering method achieves an $\tilde{O}(\sqrt{T})$ regret bound independent of the action space size, without explicit estimation of transition dynamics or rewards and without requiring any arrival-rate decay assumptions.

Lastly we emphasize numerous work leveraging domain-specific structure for improved online learning algorithms in particular operations research applications, including assortment optimization (Cao and Sun 2019), joint pricing and inventory control (Chen et al. 2022), dynamic pricing with product returns (Chen et al. 2025), reward programs in revenue management (Hssaine et al. 2025), and pricing for service systems with reusable resources (Jia et al. 2024). This growing body of work complements our focus on studying leveraging information structures in real-world MDPs for improved guarantees.

Empirical RL Applications. We note that many empirical algorithms have been applied to a variety of real-world systems without explicitly exploiting their MDP structure. For example, Feng et al. (2021) use deep learning methods to optimize ride-sharing dispatch, Dai and Gluzman (2022) apply policy-gradient methods to dynamic queuing networks, and Fang et al. (2019) explore empirical approaches for jitter-buffer management in streaming. While these methods have shown empirical success in many applications, they typically incur high sample complexity because they do not exploit the underlying problem structure (see comparison with PPO in Section 5). By incorporating policy ordering, our method offers the potential to improve sample efficiency and overall performance in these domains.

Paper Organization. The remainder of the paper is organized as follows. In Section 2, we introduce the necessary notation and formally define the MDP problem, along with the information ordering over the policy class that underpins our framework. Section 3 presents our main theoretical contributions: we derive upper and lower regret bounds under partial information orders over policies, and show that these bounds are nearly tight. In Section 4, we demonstrate the applicability of our framework through three case studies:

single-retailer inventory control with positive lead time in Section 4.1, the dual index policy for the lost-sales dual sourcing problem in Section 4.2, and an M/M/1/L queue with service-rate control in Section 4.3. In each case, we compare the theoretical performance of our algorithm against existing approaches (when available). Section 5 presents detailed numerical simulations that validate our theoretical guarantees and empirically examine the impact of model parameters in each case study. Finally, Section 6 concludes with a discussion of potential extensions, and main technical proofs and supplementary details are deferred to the appendix. Due to space constraints, the full version of the paper is available online.

2. Preliminary

Technical Notation. We use $[T] = \{1, \dots, T\}$. For random variables X and Y, we write $X \in Y$ to indicate that X is measurable with respect to the σ-algebra generated by Y. When $X \in Y$, we sometimes write $X(Y)$ to emphasize this measurability. The total variation distance between X and Y (in the sense of their pushforward measure) is denoted by $d_{\mathrm{TV}}(X, Y)$. Unless otherwise specified, all norms over the policy class Θ refer to the ℓ_∞ norm.

MDP and Policies. We consider an agent interacting with an underlying infinite-horizon average-cost Markov Decision Process (MDP) over T sequential periods. The underlying MDP is given by a five tuple $(\mathcal{S}, \mathcal{A}, P, C, s_1)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, P the state transition kernel, $C : \mathcal{S} \times \mathcal{A} \to [0, 1]$ the stochastic cost function, and s_1 the initial state (Puterman 2014). We assume that the agent has access to a class of stationary deterministic policies $\pi_\theta : \mathcal{S} \to \mathcal{A}$ parameterized by a value $\theta \in \Theta$. We use θ and π_θ interchangeably to denote both the parameter and the policy.

The system evolves as follows. The initial state is given by s_1. At each time step t, the agent observes the current state $S_t \in \mathcal{S}$, selects a policy parameter $\theta_t \in \Theta$, and applies the corresponding action $A_t = \pi_{\theta_t}(S_t)$. The agent then observes a realized cost $C(S_t, A_t)$, and the state transitions to S_{t+1} according to the transition kernel $P(\cdot \mid S_t, A_t)$.

Loss and Bias. For any fixed policy θ, the *long-run average cost* $g_\theta(s)$ is the cumulative average cost starting from state s. Similarly, the *bias* $v_\theta(s)$ is the total difference in the average cost from the asymptotic average cost starting from state s. More formally we have (Puterman 2014):

$$g_\theta(s) = \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T} \sum_{t \le T} C(S_t, A_t) \;\Big|\; S_1 = s\Big], \quad v_\theta(s) = \mathbb{E}\Big[\lim_{T\to\infty} \sum_{t \le T} (C(S_t, A_t) - g_\theta(S_t)) \;\Big|\; S_1 = s\Big].$$

Here $\theta^* = \arg\min_{\theta \in \Theta} g_\theta(s_1)$ denotes the optimal parameter, and $\pi^* = \pi_{\theta^*}$ denotes the optimal in-class policy.

Online Learning Structure. We assume that over a series of rounds $t \in [T]$ the agent needs to make sequential decisions using one of the policies π_{θ_t} for $\theta_t \in \Theta$. Their goal is to minimize the total cost, $\sum_t C(S_t, A_t)$ starting from $S_1 = s_1$ and following their decisions $A_t \sim \pi_{\theta_t}(S_t)$. We benchmark the agent on their *regret*: the additive loss over all periods the agent experiences using their policy instead of the optimal in-class one. In particular, the regret is defined as:

$$\mathsf{Regret}(T) = \sum_{t=1}^{T} C(S_t, A_t) - T g_{\theta^*}(s_1), \tag{1}$$

where $A_t = \pi_{\theta_t}(S_t)$ and $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. We consider in-class policy learning, since in many practical reinforcement learning and operations research settings, the policy class Θ is chosen to represent a family of structured or heuristic decision rules that are computationally and operationally tractable (Karlin 1958, Veeraraghavan and Scheller-Wolf 2008, Huh et al. 2009, Xin and Van Mieghem 2023) (including those examined in Section 4). The quality of our regret metric therefore depends on the expressiveness of Θ, i.e., how close the in-class optimal policy π_{θ^*} is to the true optimal.

2.1. Information Orders over Policies

In many problem domains, data collected from executing one policy can be used to infer the performance of others. In the newsvendor example above, parameterizing policies according to their ordering level q yields that samples collected under policy π_q can be used to estimate the performance of $\pi_{q'}$ for $q' \leq q$. We introduce a partial order over policies capturing this phenomenon represented via $\pi_{q'} \preceq \pi_q$. Before defining this we start with some necessary technical notation.

DEFINITION 1. For any policy $\theta \in \Theta$ and $T \in \mathbb{N}^+$, we denote $\mathcal{H}_\theta^T$ to refer to a sample path of the form $\{(S_t, A_t, C(S_t, A_t), S_{t+1})\}_{t\in[T]}$, collected under policy π_θ where the starting state $S_1 = s_1$. We define $G_\theta(\mathcal{H}_\theta^T)$ as the empirical average cost over a length T trajectory $\mathcal{H}_\theta^T$ starting from s_1:

$$G_\theta(\mathcal{H}_\theta^T) = \frac{1}{T}\sum_{t=1}^{T} C(S_t, A_t), \quad \text{where } S_1 = s_1. \tag{2}$$

The dependence of $G_\theta(\mathcal{H}_\theta^T)$ on s_1 is implicit and omitted for brevity. Note that the random variable $G_\theta(\mathcal{H}_\theta^T)$ is measurable with respect to $\mathcal{H}_\theta^T$, indicating that by sampling a trajectory from θ one can evaluate $G_\theta(\mathcal{H}_\theta^T)$ to estimate its long-run average cost $g_\theta(s_1)$.

2.1.1. Sample-Path Policy Ordering

The most straightforward way to perform counterfactual estimation is when $G_{\theta'}(\mathcal{H}_{\theta'}^T) \in \mathcal{H}_\theta^T$, i.e., the empirical average cost under θ' is estimatable using data collected from θ. This is exemplified in the newsvendor discussion above. While this condition may appear restrictive, we show that it holds in some cases, like the inventory control problem with positive lead time and lost-sales in Section 4.1. The corresponding policy order (referred to as the *sample-path* policy order) is defined as follows:

DEFINITION 2. For two policies $\theta', \theta \in \Theta$, we write $\pi_{\theta'} \preceq \pi_\theta$ if, for all $T \in \mathbb{N}^+$, it holds that

$$G_{\theta'}(\mathcal{H}_{\theta'}^T) \in \mathcal{H}_\theta^T. \tag{3}$$

The term *sample-path* arises from the observation that Definition 2 is satisfied whenever $\mathcal{H}_{\theta'}^T \in \mathcal{H}_\theta^T$, since $G_{\theta'}(\mathcal{H}_{\theta'}^T)$ is measurable with respect to $\mathcal{H}_{\theta'}^T$ trivially.

We revisit the newsvendor problem discussed above to illustrate the sample-path policy ordering. With zero lead time, the state space is trivial, and observed sales $N(q) = \min\{q, D\}$ (D refers to the stochastic demand) deterministically imply $N(q')$ for any $q' \leq q$, so $N(q') \in N(q)$. Hence, for all T, $\mathcal{H}_{q'}^T \in \mathcal{H}_q^T$, and the sample-path policy ordering $\pi_{q'} \preceq \pi_q$ follows immediately.

2.1.2. Distributional Policy Ordering While the sample-path policy order captures exact measurability of one policy's cost under trajectories collected under another, many other domains only permit *approximate* inference. In such cases, it suffices to construct a random variable $\tilde{G}_{\theta'|\theta}$ such that (i) $\tilde{G}_{\theta'|\theta} \in \mathcal{H}_\theta^T$, and (ii) $\tilde{G}_{\theta'|\theta}$ approximates the distribution of $G_{\theta'}(\mathcal{H}_{\theta'}^{T'})$ for some $T' \in \mathbb{N}^+$. Measurability implies that we can construct $\tilde{G}_{\theta'|\theta}$ from $\mathcal{H}_\theta^T$ to estimate the performance of θ' using samples collected from θ. The second condition ensures that the estimate is "sufficiently good". Moreover, T' may need to be strictly less than T if the information collected under θ is insufficient to support counterfactual inference over the full horizon T. However, as long as T' is a *constant fraction* (denoted with $\alpha \in (0, 1]$) of T, then using $\tilde{G}_{\theta'|\theta}$ in lieu of $G_{\theta'}(\mathcal{H}_{\theta'}^T)$ is sufficient. The formal definition of this *distributional* policy ordering is given by:

DEFINITION 3. Given a fixed $\alpha \in (0, 1]$, we write $\pi_{\theta'} \preceq \pi_\theta$ if there exists a random variable $\tilde{G}_{\theta'|\theta}$ such that, for any $\delta > 0$, there exists $T_h(\delta) \in \mathbb{N}^+$ satisfying

$$d_{\mathrm{TV}}(\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T), G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha T})) \leq \delta \quad \text{for all} \quad T \geq T_h(\delta). \tag{4}$$

Here $T_h(\delta)$ is a constant that depends on the system dynamics and the confidence level δ. Note that Definition 2 implies Definition 3 under $\alpha = 1$ and $T_h(\delta) = 1$, since we set $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) = G_{\theta'}(\mathcal{H}_{\theta'}^T)$. We note that our algorithms require knowledge of $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ in order to perform counterfactual inference. However, these orders allow us to capture settings where exact sample-path measurability fails, by permitting approximate estimation over a fraction of the trajectory and tolerating small distributional discrepancies. In Section 4, we show that the distributional policy order is powerful in some cases (Section 4.2 and Section 4.3), where the sample-path policy order does not hold.

2.1.3. Width of the Partial Order Our results rely on the *width* of the partial order, defined as the size of the largest set $\{\theta_1, \ldots, \theta_w\}$ of elements in Θ such that $\pi_{\theta_i} \not\preceq \pi_{\theta_j}$ for all $i \neq j$ (Dushnik and Miller 1941). Note that any finite policy class Θ admits a trivial partial order, where no two policies are related. In particular, one has $\pi_{\theta'} \not\preceq \pi_\theta$ for all $\theta' \neq \theta$, resulting in a partial order with *width* $w = |\Theta|$. While such orderings always exist, they offer no benefit in improving sample efficiency. In contrast, when Θ has a partial

order with small width, policies can share more information with one another, leading to faster learning and lower regret. Note that $w = 1$ implies the existence of a policy capable of counterfactual estimation for every policy in Θ. The full-feedback bandit setting is a special case of the $w = 1$ full ordering (Alon et al. 2015). More surprisingly, is that several classic problems in operations have natural policy classes with information orders of small width.

2.2. Assumptions on the MDP

Our algorithm and results rely on several mild assumptions on the underlying MDP and parameterized policy class Θ.

ASSUMPTION 1. *The parameterized policy class $\Theta \subset \mathbb{R}^d$ is bounded by U, i.e. $\sup_{\theta\in\Theta}\|\theta\| \le U$.*

The boundedness assumption on Θ is mild and commonly adopted in the literature on policy learning problems (Agrawal and Jia 2022, Gong and Simchi-Levi 2024, Chen and Shi 2025). The next assumption requires that there is a policy which can be used to return to the starting state s_1 in a bounded number of timesteps.

ASSUMPTION 2. *There exists a policy $\theta_R \in \Theta$ such that the expected time to reach state s_1 from any other state $s \in \mathcal{S}$ is bounded by $D_\Theta \in \mathbb{N}^+$ under policy π_{θ_R}.*

Note that the *MDP diameter* introduced by Auer et al. (2008) can serve the same purpose as Assumption 2 (see Algorithm 1 for details), in which case the initial state s_1 may be chosen arbitrarily. We adopt the uniform-policy formulation in Assumption 2 for clarity, as in many operations research settings there typically exists a single policy θ_R that can return the system to a designated reference state s_1 from any $s \in \mathcal{S}$ within a finite expected time. A typical example is the no-admission policy in a queuing system. Suppose the system state is defined as the number of customers in the queue, with the initial state corresponding to zero customers. Under the no-admission policy (no new customers are allowed to enter the system), the queue will quickly return to the empty state, regardless of the state at which the no-admission policy is applied. The last assumption requires that the loss function is uniform and the span of the bias is bounded.

ASSUMPTION 3. *For any policy $\theta \in \Theta$, the following properties are satisfied:*

1. *The average loss function $g_\theta(s)$ is uniform (i.e. $g_\theta(s) = g_\theta$ for all $s \in \mathcal{S}$).*
2. *The span of the bias is bounded by H (i.e. $\max_{s,s'\in\mathcal{S}^2} |v_\theta(s) - v_\theta(s')| \le H$).*
3. *The cost function g_θ is Lipschitz continuous with respect to θ, with Lipschitz constant L_Θ.*

Beyond assumptions on the loss and bias functions in Assumption 3, the literature has also utilized mixing time (Kearns and Singh 2002) and MDP diameter (Auer et al. 2008) as alternative regularity conditions for analytical purposes.

Algorithm 1 Information-Ordered Epoch Based Policy Elimination Algorithm (lOPEA)

Discretize Θ: $\Theta_1 = \mathcal{N}_r(\Theta)$ *// Discretization*
for epoch $k = 1, \ldots, K$ **do**
  Compute the maximal policies of Θ_k: $\Theta_k^{\max} = \{\pi_{\theta_{k1}}, \ldots, \pi_{\theta_{kw}}\}$ following order $\preceq$
  // Compute most informative policies
  for each $j \in [w]$ **do**
    play $\pi_{\theta_{kj}}$ for N_k timesteps to obtain $\mathcal{H}_{\theta_{kj}}^{N_k}$
    Play π_{θ_R} until returning to state s_1 *// Return to starting state*
  end for
  Use $\{\mathcal{H}_{\theta_{k1}}^{N_k}, \ldots, \mathcal{H}_{\theta_{kw}}^{N_k}\}$ to estimate $\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$ for all $\theta' \in \Theta_k$ *// Derive counterfactual estimates*
  Update $\Theta_{k+1} = \{\theta' \in \Theta_k \mid \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - \min_{\theta \in \Theta_k} \tilde{G}_{\theta|\tilde{\theta}_k(\theta)} \leq 2\beta_k\}$
end for

3. Main Results

In this section, we present our Information-Ordered Epoch-Based Policy Elimination Algorithm (lOPEA), as well as its regret guarantee. At a high level, our algorithm builds on the insight that policy learning can be accelerated by systematically leveraging the partial order over the policy class for counterfactual inference. Across a series of epochs we maintain a confidence set Θ_k of near-optimal policies. This set is refined over time throughout the course of learning as more data is collected. Rather than exploring each policy within Θ_k across each epoch, we sample the *most informative policies* (the "maximal" policies in the information order, those that enable counterfactual evaluation of all other policies in Θ_k), thereby improving sample efficiency and regret guarantees. We now describe each stage in detail (see Algorithm 1 for full pseudocode).

Discretizing the Policy Set Θ. If the policy class Θ is continuous, we begin by discretizing the policy space $\Theta \subset \mathbb{R}^d$ with an r-net $\Theta_1 = \mathcal{N}_r(\Theta)$ under the ℓ_∞-norm. Note that by Assumption 1 we know that Θ is bounded, and so we will later set $r = (1/T)^{1/2}$. This balances the loss due to learning and the loss due to discretization (handled since g_θ is Lipschitz under Assumption 3).

Epochs and Policy Sampling. The learning process proceeds over a series of epochs $k \in [K]$, where within each epoch we maintain a set $\Theta_k \subseteq \Theta_1$ of near-optimal policies. In order to collect samples to further estimate policy performance and refine Θ_k, we select the set of *maximal* policies of Θ_k with respect to the partial order $\preceq$:

$$\Theta_k^{\max} = \{\pi_{\theta_{k1}}, \ldots, \pi_{\theta_{kw}}\} \subset \Theta_k,$$

where w is the *width* of the partial order $\preceq$ over Θ (Dushnik and Miller 1941). For each policy in $\Theta_k^{\max}$, we sample a trajectory of length $N_k = 4^k T_h(\delta) \frac{\log(T)}{\alpha}$. Note that this scales exponentially with the epoch index, ensuring sufficient samples for reliable estimation. Lastly, after playing each of the w policies, we require a

"restart" step to ensure that we start collecting our trajectories from the fixed initial state s_1. However, the cumulative regret incurred by these restart steps is bounded due to Assumption 2.

Updating Confidence Set. Due to the policy order, for each $\theta' \in \Theta_k$, there exists $\tilde{\theta}_k(\theta') \in \Theta_k^{\max}$ such that $\pi_{\theta'} \preceq \pi_{\tilde{\theta}_k(\theta')}$. Using this ordering, we update the confidence set at the end of each epoch:

$$\Theta_{k+1} = \left\{ \theta' \in \Theta_k \mid \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - \min_{\theta \in \Theta_k} \tilde{G}_{\theta|\tilde{\theta}_k(\theta)} \leq 2\beta_k \right\}, \tag{5}$$

where β_k is a time-dependent confidence parameter that decreases with rate $\tilde{O}(\frac{1}{\sqrt{N_k}})$. This reflects the intuition that, as more data is accumulated, the true optimal policy lies in a progressively smaller region of the parameter space. Appropriate tuning of β_k ensures that the confidence set contains the optimal policy with high probability throughout the learning process. We are now ready to state our main theorem (see Section A.1 for the proof).

THEOREM 1 **(Regret Upper Bound)**. *Let w denote the width of the partial order $\preceq$ restricted to $\Theta_1 = N_r(\Theta)$ for $r = T^{-1/2}$. Then under Assumptions 1 to 3, for any $\delta > 0$, by setting $\beta_k = \frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4|\mathcal{N}_r(\Theta)|K/\delta)}{\alpha N_k}}$ and $N_k = \frac{4^k T_h(\delta)}{\alpha}\log(T)$, Algorithm 1 ensures that, for any $T \geq \frac{4wT_h(\delta)}{\alpha}$, with probability at least $1-3\delta$,*

$$\textit{Regret}(T) = O\left(\left((H+2)\sqrt{\frac{wd}{\alpha}\log\left(\frac{U\sqrt{T}\log T}{\delta}\right)}\log(T) + L_\Theta\right)\sqrt{T} + \left(\frac{wH}{\alpha} + wD_\Theta\right)\log(T)\right), \tag{6}$$

where $O(\cdot)$ hides absolute constants and logarithmic factors of α, $1/w$, and $1/T_h(\delta)$.

While we assume T is known, the standard doubling trick can be applied to obtain an anytime version of the algorithm without affecting the regret guarantees (Lattimore and Szepesvári 2020).

Since the sample-path policy order is a special case of the distributional policy order (with $\alpha = 1$ and $T_h(\delta) = 1$), a direct corollary of Theorem 1 yields a regret bound of $\tilde{O}(H\sqrt{wdT})$ in this case. Moreover, if the policy class Θ is discrete, then d is replaced by $\log(|\Theta|)$ in the regret bound.

Our analysis assumes knowledge of the policy order, including the parameters w, α, and $T_h(\delta)$, the Lipschitz constant L_Θ, as well as the bias upper bound H. These quantities can often be bounded using known structural properties of the problem studied (Fruit et al. 2018). We will illustrate in Section 4 how these quantities can be bounded for each case study we investigated individually. Furthermore, when $\alpha \to 0$ the regret bound deteriorates. However, this is a scenario where the distributional policy order is *uninformative*. If $|\Theta|$ is finite, one can obtain better regret guarantees with the uninformative order (no two policies are related) with width $w = |\Theta|$ and $\alpha = 1$. This further implies that whenever a low-width policy order can be established, the regret bound depends on w rather than $|\Theta|$, offering a sharper characterization of learning

complexity under our policy order framework. Moreover, the value of H does not necessarily grow with the size of the state space; see Section 4.1 and Section 4.2 for examples with uncountable state space.

The regret bound in Theorem 1 enables a unified comparison across different feedback regimes. When no side observations are available ($w = |\Theta|$), our bound reduces to the trivial $\tilde{O}(\sqrt{|\Theta|T})$, matching standard concentration-based results. At the other extreme of full feedback, we have $w = 1$, and our result recovers the $\tilde{O}(\sqrt{T})$ bound of Dann et al. (2020), which is independent of state and action space sizes. While Dann et al. (2020) is restricted to tabular MDPs, our framework applies to continuous state and action spaces whenever the policy order holds. Beyond the no-feedback and full-feedback settings above, our regret guarantee (i) interpolates between these two extremes, and (ii) allows for more complex information structures within the distributional policy order. In Section 4 we provide several case studies in operations research and specialize Theorem 1 to those case studies.

Lastly, we complement this regret upper bound with a nearly matching lower bound for the sample-path policy order. Indeed, we have (see Section A.2 for the proof and discussion):

THEOREM 2 **(Regret Lower Bound)**. *For any $w, |\Theta|$, and H, with $|\Theta| \geq H$, for any algorithm, there exists an MDP and finite policy class Θ satisfying the sample-path policy order of width w and Assumptions 1 to 3, such that*

$$\mathbb{E}[\textit{Regret}(T)] = \begin{cases} \Omega(\sqrt{H \log(|\Theta|)T}) & w = 1 \\ \Omega(\sqrt{HwT}) & w > 1. \end{cases}$$

This shows that the dependence on regret with respect to T and w is nearly-optimal up to logarithmic factors. Our results are not optimal up to $\log(|\Theta|)$, in the same way for bandit and full-feedback settings there is a gap between $\sqrt{K}$ and $\sqrt{\log(K)}$ in the lower bounds, where K is the number of arms. Our bound is also not optimal up to $\sqrt{H}$. Existing lower bounds on average-cost MDPs exhibit similar dependence on $\sqrt{H}$ instead of the upper bounds of H (see, for instance Auer et al. (2008)). As such, this $\sqrt{H}$ discrepancy in the upper and lower bounds is common in the literature. Theorem 2 also ignores the potential dependence on α. This arises from the fact that for any finite policy class Θ, the trivial order has width $w = |\Theta|$, so one always obtains $\tilde{O}(\sqrt{H|\Theta|T})$ regret. This precludes a general $\tilde{\Omega}\big(\sqrt{HwT/\alpha}\big)$ lower-bound, since as $\alpha \to 0$ the naive $\tilde{O}(\sqrt{H|\Theta|T})$ upper-bound outperforms it. However, our notion of distributional policy order, as well as IOPEA, does not assume that $C_t(S_t, A_t)$ is observed at every step. To capture the α dependence, we therefore analyze a family of MDPs where costs are revealed only on an α-fraction of time steps. We show by explicit construction that with access to exactly αT cost observations, any algorithm incurs $\tilde{\Omega}(\sqrt{HwT/\alpha})$ regret. Thus our results are tight in α as well. More details are provided in Theorem C.1.

Choice of Policy Order. When the policy class Θ admits multiple valid orderings, it is essential to compare and select among them. Among *sample-path* orders ($\alpha = 1$), Theorem 1 asserts that the ordering with minimal width w yields the tightest bound. For *distributional* orders, minimizing the ratio w/α directly

reduces the leading constant in Theorem 1; Nevertheless, this criterion does not always lead to the optimal policy ordering. The distributional policy order only yields regret bounds for horizons $T \geq T_h(\delta)$ at fixed confidence δ, where the threshold $T_h(\delta)$ itself depends on the chosen ordering, making order selection ambiguous in non-asymptotic T settings. Empirically, in our dual-sourcing experiments (Section 4.2) where the worst-case bound on α decays exponentially in the lead time L_r, we still observe rapid convergence (Figure 1), indicating that the *effective* mixing rate is typically much higher (often $O(1)$) than the pessimistic estimate. Thus, a pragmatic heuristic is to choose the order that minimizes w first, regardless of α. A systematic study of policy-order selection across diverse MDPs is left for future work.

Proof Sketch of Theorem 1. For each policy $\theta' \in \Theta_k$ we evaluate at epoch k, we first establish the concentration of $\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$ to its true long-run average cost $g_{\theta'}$ by combining a martingale concentration argument and the bounded total variation distance from the distributional policy order. Thus we establish that with high probability, $|\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - g_{\theta'}| \leq \beta_k$.

Next, we show by induction (Lemma A.6) that the discretized optimal policy $\theta_r^* = \arg\min_{\theta \in \mathcal{N}_r(\Theta)} g_\theta$ is never eliminated for each epoch k. Specifically, at each epoch k, any surviving policies $\theta, \theta' \in \Theta_k$ satisfy $\left|g_\theta - g_{\theta'}\right| \leq 6\beta_k$.

Finally, we decompose the total regret over T steps by summing over each epoch. In epoch k, at most w policies are implemented for N_k steps each, incurring regret at most $wN_k \cdot \tilde{O}(\beta_k)$. Returning to the initial state adds at most wD_Θ cost per epoch in expectation. Choosing $N_k = \tilde{O}\big(4^k \alpha^{-1} \log T\big)$ forces the number of epochs $K = \tilde{O}(\log T)$. Summing $wN_k \cdot \tilde{O}(\beta_k)$ over $k = 1, \ldots, K$ and adding the returning cost and discretization error $O(T\, r\, L_\Theta)$ yields

$$\mathsf{Regret}(T) \leq O\left(\frac{wH}{\alpha} + wD_\Theta\right) \log_4 \frac{\alpha T}{wT_h(\delta)} + O\left((H+2)\sqrt{\frac{w}{\alpha} \log(4|\mathcal{N}_r(\Theta)|K/\delta)T \log(T)}\right) + TL_\Theta r.$$

By taking $r = (1/T)^{1/2}$, we have $|\mathcal{N}_r(\Theta)| = O((U\sqrt{T})^d)$, which completes the proof. □

4. Case Studies

To demonstrate the practical relevance of our framework, we instantiate it in several canonical problems in operations research, including inventory control with positive lead time (Goldberg et al. 2021), dual sourcing (Whittemore and Saunders 1977), and queuing models with state-dependent service rates (Puterman 2014). In each case, we identify natural policy classes that admit a low-width information order, enabling new algorithms and novel regret guarantees. We further complement our theoretical findings with numerical simulations comparing our method to existing baselines in Section 5. Additional proofs and details are in Appendix D in the full version of the paper.

4.1. Single-Retailer Inventory Control with Positive Lead Time

Our first case study considers a single-retailer inventory control problem with positive lead time, a canonical supply chain management problem in the operations research literature (see Goldberg et al. (2021) for a survey of recent structural results on this problem).

4.1.1. Model A retailer is faced with making ordering decisions Q_t online over a period of rounds $t = 1, \dots, T$. At the beginning of each step t, the inventory manager observes the current inventory level I_t as well as the L previous unfulfilled orders in the pipeline, denoted as $Q_{t-L}, \dots, Q_{t-1}$. Here, the integer $L \geq 1$ is the so-called *lead time*, or delay in the number of steps between placing an order and receiving it. The system evolves according to the following dynamics. At the beginning of each stage t, the inventory manager picks an order Q_t to arrive at stage $t + L$. Then, the order Q_{t-L} that was made L time steps earlier arrives. Next, an unobserved demand $D_t \geq 0$ is generated independently from an unknown distribution $\mathcal{F}$, which we assume is supported on $[0, U]$ for simplicity; however, our results extend to more general demand distributions. We assume that $\mathbb{P}(D_t = 0) > 0$. The number of products sold is the minimum of on-hand inventory and demand, i.e. $\min\{I_t + Q_{t-L}, D_t\}$. Note that the decision maker only observes the sales, and not the actual demand D_t.

MDP Formulation. To formulate this as an MDP, the state space is $\mathcal{S} = [0, (L+1)U]^{L+1}$, where each state s consists of the current inventory level I_t along with the previous L unfulfilled orders $Q_{t-L}, \dots, Q_{t-1}$ in the pipeline. The action space is given by $\mathcal{A} = [0, U]$ where action $A_t = Q_t$ denotes the order placed at time t. The transition and reward dynamics are:

$$S_{t+1} = ((I_t + Q_{t-L} - D_t)^+, Q_{t-L+1}, \dots, Q_t) \tag{7}$$

$$C_t(S_t) = h(I_t + Q_{t-L} - D_t)^+ + p(D_t - (I_t + Q_{t-L}))^+, \tag{8}$$

where h and p are the holding and lost-sales cost coefficients. Finally, we assume a fixed starting state of $s_1 = (0, \dots, 0)$, corresponding to no on-hand inventory and no outstanding orders in the pipeline.

Modifications to Cost Function. As written, $C_t(S_t)$ is not observed, since it depends on the realized demand D_t whereas the algorithm only has access to *sales* data $N_t = \min\{D_t, I_t + Q_{t-L}\}$. However, one can transform the cost to a so-called *pseudo-cost* which maintains the same average cost up to a constant independent of the policy (Agrawal and Jia 2022):

$$\tilde{C}(S_t) = C(S_t) - pD_t = h(I_t + Q_{t-L} - \min\{I_t + Q_{t-L}, D_t\}) - p\min\{I_t + Q_{t-L}, D_t\}.$$

As is common in the literature, our results on the information order leverage this pseudo-cost, but when reporting the performance of the algorithms, we report the true cost $C_t(S_t)$. We emphasize that our results apply regardless of the choice of cost function, since unlike existing literature, we do not rely on convexity (Agrawal and Jia 2022). Indeed, there are several well-studied models in operations (including step-dependent holding and lost-sales costs) where the cost function is no longer convex (Chen and Yu 2021).

4.1.2. Policies and Information Order We consider the class of *base-stock* policies parameterized by their base stock level $\theta \in [0, U] = \Theta$ (Goldberg et al. 2021). Intuitively, these policies order a quantity that brings the sum of leftover inventory and outstanding orders to θ. Formally, fixing the base-stock level θ, the action at step t is given by:

$$\pi_\theta(I_t, Q_{t-L}, \ldots, Q_{t-1}) = Q_t = (\theta - I_t - \sum_{i \in [L]} Q_{t-i})^+. \tag{9}$$

Prior work such as Huh et al. (2009), Zipkin (2008) shows that base-stock policies are optimal when either $L = 0$ or the lost-sales cost $p \to \infty$. See Goldberg et al. (2021) for more discussion. Next up, we show that the base-stock policies satisfies a sample-path policy order, where $\pi_{\theta'} \preceq \pi_\theta$ whenever $\theta' \leq \theta$. We further note that computing the counterfactual estimates for $G_{\theta'}(\mathcal{H}_{\theta'}^T)$ from $\mathcal{H}_\theta^T$ can be done in $O(T)$ time (as exemplified in the proof of the result).

LEMMA 1. *The set of base stock policies satisfies the sample-path policy order of width one, where $\pi_{\theta'} \preceq \pi_\theta$ whenever $\theta' \leq \theta$.*

4.1.3. Performance Guarantee Combining Lemma 1 with the fact that the resulting problem satisfies Assumptions 1 to 3 establishes:

COROLLARY 1. *Applying Algorithm 1 to the inventory control with positive lead time L yields an algorithm achieving a regret guarantee $\tilde{O}\left(L \max\{h, p\} U \sqrt{T \log(U\sqrt{T}/\delta)}\right)$.*

This result differs from the regret bound in Agrawal and Jia (2022) by an additional factor $\sqrt{\log(U)}$, yet it holds without the assumption that the long-run average cost g_θ is convex, thereby it applies to non-convex cost structures such as the step-wise ordering costs in Chen and Yu (2021).

Proof. In order to apply Theorem 1, it suffices to show that the set of base stock policies satisfies Assumptions 1 to 3, and to provide values for U, D_Θ, H, L_Θ, α, w, and d. Using Lemma C.1 in Agrawal and Jia (2022) we know g_θ is Lipschitz with $L_\Theta = \max\{h, p\}$. Similarly, Assumptions 1 and 3 are satisfied under $U = U$ and $H = 36 \max\{h, p\} LU$ (Lemma 2.8 of Agrawal and Jia (2022)). Assumption 2 holds under base-stock level $\theta = 0$ with $D_\Theta = U\mathbb{E}[\tau]$ where $\tau = \inf\{t \geq 0 : \sum_{t'=1}^t D_{t'} \geq 1\}$. We have $\alpha = 1$ by the sample-path policy order established in Lemma 1. We have $d = 1$ since $\theta \in [0, U]$. Furthermore, the policy order has width $w = 1$, as it is fully determined by the natural order on $\mathbb{R}_{\geq 0}$. □

4.2. Dual Index Policy for the Dual Sourcing Problem

Dual sourcing extends the single-channel inventory control model in Section 4.1 to a setting with two replenishment channels: a regular channel with lower cost and longer lead time, and an expedited channel with higher cost and shorter lead time. The optimal control of the dual sourcing problem with lost-sales is notoriously challenging (Xin and Van Mieghem 2023). Consequently, numerous heuristic policies have been proposed in the literature, such as dual index policies and tailored base-surge policies (Xin and Van Mieghem

2023). In this section, we focus on the dual index policies of Veeraraghavan and Scheller-Wolf (2008) for two reasons. First, the superior performance of dual index policies in dual-sourcing is well-documented (Li and Yu 2014, Hua et al. 2015), making them a preferred approach in practice. Second, despite their popularity, providing regret guarantees for dual index policies in the lost-sales context remains an open problem in the literature.

4.2.1. Model The system dynamics mirror those in Section 4.1 with two key extensions: (i) the retailer places two orders (Q_t^r, Q_t^e) from regular and expedited channels with lead times $L_r > L_e$, and (ii) outstanding orders from both channels are tracked separately.

At time t, the retailer observes the current on-hand inventory I_t, along with the pending regular orders $(Q_{t-L_r}^r, \ldots, Q_{t-1}^r)$ and expedited orders $(Q_{t-L_e}^e, \ldots, Q_{t-1}^e)$. The retailer then selects actions (Q_t^r, Q_t^e) and fulfills demand D_t using available inventory. Let the sales be $N_t = \min\{I_t + Q_{t-L_r}^r + Q_{t-L_e}^e, D_t\}$. Only sales N_t are observed instead of the true demand D_t. The full state is defined as

$$S_t = (I_t, Q_{t-L_r}^r, \ldots, Q_{t-1}^r, Q_{t-L_e}^e, \ldots, Q_{t-1}^e) \in [0, (L_r + L_e + 1)U]^{L_r+L_e+1},$$

and the action space is $\mathcal{A} = [0, U]^2$. The system evolves according to the following transition and cost dynamics:

$$\begin{aligned} I_{t+1} &= \big(I_t + Q_{t-L_r}^r + Q_{t-L_e}^e - D_t\big)^+ \\ C_t(S_t) &= h\big(I_t + Q_{t-L_r}^r + Q_{t-L_e}^e - D_t\big)^+ + p(D_t - I_t - Q_{t-L_r}^r - Q_{t-L_e}^e)^+ + c_r Q_t^r + c_e Q_t^e, \end{aligned}$$

where h and p are the holding and lost-sales cost coefficients, and c_r and c_e are the unit costs for regular and expedited orders. The next state S_{t+1} is obtained by shifting the regular and expedited pipelines one step forward, appending the newly placed orders, and updating the inventory level:

$$S_{t+1} = \big(I_{t+1}, Q_{t-L_r+1}^r, \ldots, Q_t^r, Q_{t-L_e+1}^e, \ldots, Q_t^e\big).$$

We assume a fixed initial state $s_1 = (0, \ldots, 0)$.

Modifications to Cost Function. Similar to Section 4.1 as written, $C_t(S_t)$ is not observed since it depends on the realized demand D_t. However, one can again transform the cost to a pseudo-cost maintaining the same average cost via:

$$\tilde{C}(S_t) = C(S_t) - pD_t.$$

Our results in this section still apply regardless of the choice of cost function, as in Section 4.1. When indicating the performance of the algorithms we measure the true cost $C(S_t)$ instead of the pseudocost $\tilde{C}$. We further assume without loss of generality that $C_t(S_t) \in [0, 1]$, which can be achieved by appropriate normalization.

4.2.2. Policies and Information Order We consider a class of heuristic policies known as dual index policies (Veeraraghavan and Scheller-Wolf 2008), which are known to perform well in practice despite the complexity of the optimal policy (Janakiraman and Seshadri 2017). Each policy $\theta = (z_e^\theta, z_r^\theta) \in \Theta \subset \mathbb{R}^2$ is parameterized by two base-stock levels: z_e^θ for the expedited channel and z_r^θ for the total inventory position. At each time step t, the expedited and regular orders are computed sequentially as:

$$Q_t^e = \left(z_e^\theta - I_t - \sum_{i=1}^{L_e} Q_{t-i}^e - \sum_{i=L_r-L_e}^{L_r} Q_{t-i}^r \right)^+$$
$$Q_t^r = \left(z_r^\theta - I_t - \sum_{i=1}^{L_e} Q_{t-i}^e - \sum_{i=1}^{L_r} Q_{t-i}^r - Q_t^e \right)^+.$$

The policy first brings the *expedited inventory position* (on-hand inventory plus expedited pipeline and truncated regular pipeline) up to z_e^θ (if possible), and then tops off the total inventory position to z_r^θ using regular orders. Although dual index policies do not satisfy a sample-path policy order, a distributional partial order holds.

More concretely, we construct our counterfactual estimator $\tilde{G}_{\theta'|\theta}$ based on the following key insight: *regardless of which policy θ is executed*, it is always possible to obtain an unbiased estimate of $g_{\theta'}$ for any $\theta' \in \Theta$, provided that enough uncensored demand samples are available. This follows from the fact that demand realizations are independent of both the state and the policy, together with the fact that observed sales are capped by the on-hand inventory (after order arrivals) at each step (which is policy dependent). Thus any demand sample censored under a higher inventory level remains valid for estimating counterfactual sales under any lower inventory level.

As a result, for any pair of policies θ and θ' with $z_r^{\theta'} \le z_r^\theta$, the segments of the trajectory $\mathcal{H}_\theta^T$ where the on-hand inventory equals z_r^θ can be used to construct an unbiased estimate of $g_{\theta'}$, since under θ', the inventory level never exceeds $z_r^{\theta'} \le z_r^\theta$. Moreover, trajectories induced by θ revisit high-inventory states with constant probability. This ensures that, with high probability, a sufficient number of usable samples are collected under $\mathcal{H}_\theta^T$ to accurately estimate $g_{\theta'}$ for all θ' such that $z_r^{\theta'} \le z_r^\theta$. We formalize this idea in what follows, beginning with the definition of the estimator $\tilde{G}_{\theta'|\theta}$.

DEFINITION 4. Consider a fixed $\alpha \in (0, 1]$ and a trajectory $\mathcal{H}_\theta^T$ under policy θ. Define the hitting times

$$\tau_i = \inf\{t > \tau_{i-1} : I_t^\theta = z_r^\theta\}, \quad \tau_0 = 0,$$

and let $\mathcal{I}_\theta = \{\tau_1, \tau_2, \dots\}$. Denote $\tilde{D}_t$ as the sales observed at time t. For any $\theta' \in \Theta$, if $|\mathcal{I}_\theta| \ge \alpha T$, we define the counterfactual trajectory $\tilde{\mathcal{H}}_{\theta'}^{\alpha T}$ by simulating policy θ' from the initial state s_1 assuming the true demand sequence is $\{\tilde{D}_{\tau_i}\}_{i=1}^{\alpha T}$. Then the counterfactual estimate $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ is defined as:

$$\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) = \begin{cases} G_{\theta'}(\tilde{\mathcal{H}}_{\theta'}^{\alpha T}), & \text{if } |\mathcal{I}_\theta| \ge \alpha T, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Note that the dual-sourcing problem constitutes an exogenous MDP (Exo-MDP): demand realizations are independent of the MDP's actions and states (Sinclair et al. 2023b, Wan et al. 2024). This exogeneity directly enables counterfactual estimation, allowing us to infer the performance of one policy using data collected under another. Finally, we have:

LEMMA 2. *Set $\gamma = \mathbb{P}(D_t = 0)$ and let $\alpha = \frac{1-\gamma}{2}\gamma^{L_r}$. Fix any policy $\theta \in \Theta$, confidence level $\delta > 0$, and time horizon $T \geq T_h(\delta)$, where $T_h(\delta) = \Omega(\frac{\log(1/\delta)}{(1-\gamma)^4\gamma^{2L_r}})$. Let $\tilde{G}_{\theta'|\theta}$ be as specified in Definition 4. Then, for any policy $\theta' \in \Theta$ such that $z_r^{\theta'} \leq z_r^{\theta}$, it follows that $\pi_{\theta'} \preceq \pi_\theta$.*

4.2.3. Performance Guarantee We now state our regret bound for dual index policies in the lost-sales dual-sourcing problem. Notably, (i) the lost-sales dual index policy setting lacks convexity (Veeraraghavan and Scheller-Wolf 2008), and (ii) we achieve a $\tilde{O}(\sqrt{T})$ regret bound, improving on the $\tilde{O}(T^{2/3})$ bound of Dann et al. (2020). To our knowledge, this is the first $\sqrt{T}$-regret result for dual index policies under lost-sales. Related work includes Tang et al. (2024), who obtain $\tilde{O}(\sqrt{T})$ regret for dual index policies in a backlog model, and Chen and Shi (2025), who achieve $\tilde{O}(\sqrt{T})$ regret for tailored base-surge policies which rely on convexity.

The proof follows the same structure as that in Section 4.1. We first verify Assumptions 1 to 3:

LEMMA 3. *For the dual index policies in Section 4.2, the gain is uniform, and the span of the bias is upper bounded by $\frac{1}{(1-\gamma)\gamma^{L_r}}$.*

The exponential dependence on L_r for the span H matches the existing reinforcement learning literature (Agrawal and Jia 2022), but it remains open whether one can obtain $H = O(L_r)$ in the dual-sourcing setting. While Agrawal and Jia (2022) shows $H = O(L_r)$ for the single-channel case, no such linear bound is known here. Although our policy-order proof incurs $H = O((\gamma^{L_r})^{-1})$, empirical evidence suggests this exponential factor may be avoidable with a sharper argument. We leave this refinement to future work.

In order to apply Theorem 1 it suffices to show that g_θ is Lipschitz continuous in the base-stock levels θ. However, this property has not been studied in the literature. Here we present a proof for the case of "slow-moving" items, i.e. that the demand in each period is zero with high probability (Hahn and Leucht 2015). This assumption is realistic, since slow-moving products account for the bulk of modern inventories (Snyder et al. 2012, Bi et al. 2023).

LEMMA 4. *For any dual sourcing system, there exists $c_\gamma \in (0, 1)$ such that if $\gamma > c_\gamma$, the cost function g_θ is Lipschitz continuous with respect to θ.*

The constant c_γ can be computed explicitly, see Remark EC.3 for further details and an empirical evaluation. We observe that dual index policies remain Lipschitz empirically even without this "slow-moving" assumption. However, we leave a thorough treatment of the Lipschitz property for future work.

Combining the results above yields the desired regret bound from Theorem 1:

COROLLARY 2. *For any dual sourcing system with fixed (L_r, L_e), if g_θ is Lipschitz continuous with respect to θ, our algorithm IOPEA achieves $\tilde{O}(\gamma^{-\frac{3L_r}{2}}(1-\gamma)^{-\frac{3}{2}}\sqrt{T})$ regret.*

Proof. First, note the bound on the bias is given in Lemma 3. The final regret bound follows from Lemma 2 by taking $d=2, w=1, \alpha=\frac{1-\gamma}{2}\gamma^{L_r}$. □

Note that the $\sqrt{T}$-regret bound of Corollary 2 holds for *lost-sales* dual index policies, which are strictly more challenging than the *backlog* dual index policies studied in Tang et al. (2024). In the backlog case, demand is fully observed at every time step, enabling counterfactual estimation under any policy, which is the basis of the algorithm introduced in Tang et al. (2024).

We also remark that in certain dual-sourcing settings, like when $L_r - L_e = 1$, one can obtain a two-dimensional sample-path policy order over the dual index policy class Θ: for two policies $\theta = (z_e^\theta, z_r^\theta), \theta' = (z_e^{\theta'}, z_r^{\theta'}) \in \Theta$, if $z_e^{\theta'} \le z_e^\theta$ and $z_r^{\theta'} \le z_r^\theta$, then in sample-path policy order sense, $\pi_{\theta'} \preceq \pi_\theta$. However, under this sample-path policy order, the width w scales with $1/r$ (thus Theorem 1 fails as w is no longer a constant), leading to a $\tilde{O}(T^{2/3})$ regret—worse than the $\tilde{O}(\sqrt{T})$ bound in Corollary 2. This shows that even when both sample-path and distributional orders are available, the distributional order can yield superior T-dependence ($\sqrt{T}$ vs. $T^{2/3}$), despite having constants exponential in L_r.

4.3. M/M/1/L Queuing Model with Service Rate Control

In our third case study, we consider an M/M/1/L queueing system with impatient customers and controllable service rates (Walton and Xu 2021). Despite the simplicity of the model (featuring only two unknown parameters: arrival and deadline rates), existing learning algorithms suffer from sample complexity that grows with the size of the action space. However, while the queue length is always inherently bounded, the action space may be large when we allow the service rate to take many possible values. We show that, by constructing a suitable distributional policy order, our framework yields a regret bound independent of the size of the action space in this setting. Additionally, this problem has no *partial feedback*, as required in Dann et al. (2020).

4.3.1. Model We consider an M/M/1/L queuing system in which the decision-maker dynamically selects a service rate from a finite set of options. Note that these models have also been proposed to represent a dynamic voltage and frequency scaling processor control (Anselmi et al. 2021). Jobs arrive to a finite buffer of size L according to a Poisson process of (unknown) rate λ. Upon each arrival of a job, they draw an unobserved deadline from an exponential distribution with (unknown) rate μ. Each job requires exactly one unit of processing work.

When a job arrives to the system and the queue is full (i.e. the total number of jobs in the system is L), the job is lost. Otherwise, it enters the queue and waits for service. However, if a job's deadline elapses before the service completes, it departs immediately and the algorithm incurs a fixed penalty C. We assume that the queue starts at time $t=1$ and is initially empty.

To control the queue, the algorithm may choose at each state (total number of customers in the system), a processing speed $a \in \{0, 1, \ldots, A_{\max}\}$. Operating at speed a processes a work-units per time unit, and incurs power cost $w(a)$, where $w(\cdot)$ is an arbitrary convex cost function. While the assumption that $w(\cdot)$ is convex aligns with Anselmi et al. (2022), our results apply to arbitrary bounded cost functions. The goal of the controller is to trade off running faster (higher $w(a)$) against letting jobs miss their deadlines (each incurring a cost C).

MDP Formulation. In order to analyze this problem using a discrete time formulation, we apply the uniformization trick with a constant $U = \lambda_{\max} + L\mu_{\max} + A_{\max}$, where $\lambda_{\max}$ and $\mu_{\max}$ are known upper bounds on the arrival and deadline rates (Anselmi et al. 2022). We construct an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C, s_1)$ as follows.

The state space for the MDP is $\mathcal{S} = \{0, \ldots, L\}$ corresponding to the total number of jobs in the system. The action space is $\mathcal{A} = \{0, 1, \ldots, A_{\max}\}$. Taking action a in state s, the system transitions to a new state via:

$$P(s' \mid s, a) = \begin{cases} \frac{\lambda_i}{U} & s' = s+1, s < L, \\ \frac{s\mu + a}{U} & s' = s-1, s > 0, \\ 1 - \frac{\lambda_i + s\mu + a}{U} & s' = s, \end{cases}$$

where $\lambda_i = \lambda\left(1 - \frac{i}{L}\right)$ is the state-dependent decaying arrival rate (Anselmi et al. 2022). We adopt this decaying rate, in which λ_i represents the true arrival rate at state i from Anselmi et al. (2022) to enable a fair comparison, but our results do not rely on this strong assumption (see Remark 4). Further discussion regarding such decaying rate is in Remark 4.

The cost $C(s, a)$ in state s having taken action a is a random variable defined as:

$$C(s, a) = \begin{cases} \frac{w(a)}{U} + C, & \text{with probability } \frac{s\mu}{U}, \\ \frac{w(a)}{U}, & \text{otherwise.} \end{cases}$$

This corresponds to paying $w(a)$ for servicing jobs at a rate a, with a potential cost of C if the current job's deadline is before the service is complete.

4.3.2. Policies and Information Order

Since the state and action space are finite, we consider the class of all *deterministic* policies, i.e. $\Theta = A_{\max}^{L+1}$. Each policy $\theta \in \Theta$ is of the form

$$\pi_\theta(s) = \theta_s, \tag{11}$$

corresponding to the service speed $\theta_s \in \{0, \ldots, A_{\max}\}$ employed when the queue has s jobs in the system. See Anselmi et al. (2022) for more discussion on properties of the optimal policy.

Note that under mild stability conditions, the controlled Markov chain admits a unique stationary distribution $m_\theta \in \Delta(\mathcal{S})$ for any policy $\theta \in \Theta$. Moreover, the long-run average cost of π_θ can be written as:

$$g_\theta = \sum_{s=0}^{L} m_\theta(s)\mathbb{E}[C(s, \pi_\theta(s))]. \tag{12}$$

Construction of the Information Order. Note that the MDP has only two unknown parameters: λ and μ. Hence, no matter the implemented policy θ, if we can estimate λ and μ accurately, we can counterfactually estimate the performance of any other policy $\theta' \in \Theta$.

Indeed, we start by showing how to use $\mathcal{H}_\theta^T$ to estimate λ and μ directly, regardless of the policy θ. With this, we construct the stationary distribution $\hat{m}_{\theta'}$ for an arbitrary policy θ' as its stationary distribution when the arrival and deadline rates are the estimates, $\hat{\lambda}$ and $\hat{\mu}$, accordingly. We then estimate its average cost via:

$$\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) = \sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}\Big[\hat{C}(s, \pi_{\theta'}(s)\Big],$$

where $\hat{C}$ differs from C in that μ is replaced by $\hat{\mu}$. Hence, in order to formally define $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ we just need to define how we estimate λ and μ, and in turn calculate $\hat{m}_{\theta'}$. Note that the stationary measures of the CTMC and the uniformized discrete chain are the same, thus we don't distinguish them below.

Estimating λ and μ. Given a trajectory $\mathcal{H}_\theta^T$ collected under policy π_θ, we denote by $\tau_{\theta,i}$ as the time of the i-th "jump" of the process. Hence, the amount of time spent in state $S_{\tau_{\theta,i}}$ is $\tau_{\theta,i+1} - \tau_{\theta,i}$. Note that although the chain evolves in discrete time, $\tau_{\theta,i}$ and i may not coincide, as the chain can remain in the same state for multiple consecutive time steps. We also denote $\tau_{\theta,0} = 1$ and $N_\theta(T) = \max\{n \mid \tau_{\theta,n} \le T\}$ to refer to the total number of jumps by time T.

We first consider estimating λ. Note that whenever the system is in state $s = 0$, it leaves only by an arrival of a new job (which occurs with rate λ). Hence we can estimate λ via

$$\hat{\lambda} = \frac{\sum_{i=0}^{N_\theta(T)-1} \mathbb{I}(s_{\tau_{\theta,i}} = 0)}{\sum_{i=0}^{N_\theta(T)-1} (\tau_{\theta,i+1} - \tau_{\theta,i})\mathbb{I}(s_{\tau_{\theta,i}} = 0)} \tag{13}$$

as the inverse of the average time it takes to depart state $s = 0$. Similarly, when the system is in state $s = L$, it leaves only by a job departure at rate $L\mu + \theta_L$. Thus we estimate μ via

$$\hat{\mu} = \frac{1}{L}\left(\frac{\sum_{i=0}^{N_\theta(T)-1} \mathbb{I}(s_{\tau_{\theta,i}} = L)}{\sum_{i=0}^{N_\theta(T)-1} (\tau_{\theta,i+1} - \tau_{\theta,i})\mathbb{I}(s_{\tau_{\theta,i}} = L)} - \theta_L\right).$$

REMARK 1. There are various other estimators for λ and μ. For example, treating arrivals and departures as Bernoulli events, or using data from states beyond just $s = L$ in order to estimate μ, may be helpful. It can be shown that this approach (at least) yields an improved $O(L^2\sqrt{T})$ regret. However, since our primary aim is to demonstrate that our algorithm achieves regret that does not scale polynomially with $A_{\max}$, we employ the simple estimators above for expository purposes.

Calculating $\hat{m}_{\theta'}(s)$. Since the state space is finite, the Markov chain is positive recurrent. Therefore, under mild regularity conditions, for any values of $\hat{\lambda}$ and $\hat{\mu}$, one can calculate the stationary distribution $\hat{m}_{\theta'}(s)$. This can be done, for instance, by calculating the generator matrix Q for the underlying process and solving the stationary equations (see Puterman (2014)).

REMARK 2. With uniformization, the system becomes a standard discrete time Markov chain, on which our algorithm operates. The key estimates, $\hat{\lambda}$ and $\hat{\mu}$, are computed directly from running the algorithm over the discrete chain. References to the continuous time process in this section and the appendix are included only to simplify certain derivations and proofs.

Based on the previous construction, we are ready to formally state the distributional policy order.

LEMMA 5. *For any T sufficiently large and any two policies θ and θ', there exists $\tilde{G}_{\theta'|\theta}$ such that with probability at least $1-\delta$,*

$$|\tilde{G}_{\theta'|\theta} - g_{\theta'}| \leq O\left(L^3\sqrt{\frac{\log(1/\delta)}{T}}\right).$$

As a result, $\pi_{\theta'} \preceq \pi_\theta$.

The policy order $\pi_{\theta'} \preceq \pi_\theta$ follows Definition 3 but is slightly modified since we estimate $g_{\theta'}$ directly rather than via empirical average cost. Details are deferred to Section D.3.1 in the full version of the paper, as this change relies on the proof of Theorem 1 in the Appendix.

4.3.3. Performance Guarantee Combining Lemma 5 with the fact that the resulting problem satisfies Assumptions 1 to 3 establishes:

COROLLARY 3. *IOPEA achieves a regret bound of $\tilde{O}\big(L^{7/2}\sqrt{\log(A_{\max})T}\big)$.*

Proof. First note that a uniform bound on the bias of $H = O(L\log L)$ is given by Lemma 3.2 of Anselmi et al. (2022). Since the state and action spaces are finite, Assumption 1 and Assumption 2 hold trivially. Lastly, since $|\Theta| = A_{\max}^{L+1}$ we have that $\log(|\Theta|) = (L+1)\log(A_{\max})$. Combining this with Theorem 1 gives that, with probability at least $1-\delta$, $\mathsf{Regret}(T) = \tilde{O}\Big(L^{7/2}\sqrt{\log(A_{\max})T}\Big)$. □

Our regret bound only scales logarithmically with $A_{\max}$ (actually, it is independent of $A_{\max}$, see Remark 3), achieved by taking into account the information order over policies. In contrast, Anselmi et al. (2022) derive a regret bound of $\tilde{O}(\sqrt{A_{\max}T})$, meaning our result improves upon theirs by a factor of $\sqrt{A_{\max}}$, at the expense of an additional factor of $L^{7/2}$. This highlights that, for the queuing problem considered, our approach excels when the state space is small but the action space is relatively large. Moreover, our results extend seamlessly to more complex settings (for example, with state-dependent arrival rates).

REMARK 3. By modifying the proof of Theorem 1 one can avoid a union bound over $\Theta = A_{\max}^{L+1}$, since we are estimating $\hat{\mu}$ and $\hat{\lambda}$ first before constructing counterfactual estimators $\tilde{G}_{\theta'|\theta}$. This modification would yield a regret guarantee of $O(L^3\sqrt{T})$, independent of $\log(A_{\max})$. We keep the current version here for expository purposes, since it aligns with the regret proofs for the two other case studies.

REMARK 4. Although we use the decaying arrival rate λ_i, this assumption is not necessary for our $O(L^3\sqrt{T})$ regret bound (Remark 3); our analysis holds for any fixed arrival rate. In contrast, Anselmi et al. (2022) rely on λ_i to remove state-space dependence (Lemma C.2 of Anselmi et al. (2022)). Without the decaying-rate assumption, our algorithm still attains $O(L^3\sqrt{T})$ regret, whereas the regret bound in (Anselmi et al. 2022) would scale with the sizes of both the state and action space.

5. Numerical Simulations

Finally, we evaluate IOPEA on all three case studies above, showing that it achieves improved performance over state-of-the-art algorithms tailored to each of the case studies, despite being a more general solver.

5.1. Baselines and Simulation Results

We compare the performance of IOPEA against several baselines:

- Model-Based RL with Feedback Graphs (Dann et al. 2020) (hereafter referred to as Feedback Graph): This is a generic model-based algorithm for finite state-action MDPs. Unlike our policy ordering framework which relies on counterfactual estimation between policies, the Feedback Graph method of Dann et al. (2020) uses observations from the implemented state–action pair (s, a) to infer counterfactual outcomes for other state-action pairs at each time step. This algorithm is restricted to tabular MDPs. Therefore, when applied to problems with continuous state and action spaces, as in Sections 4.1 and 4.2, we discretize the state and action spaces.
- Proximal Policy Optimization (PPO) (Schulman et al. 2017): A popular on-policy gradient method that updates the policy via the clipped surrogate objective.
- Stochastic Gradient Descent (SGD): An SGD method based on finite-difference gradient estimates. While prior works in operations research have proposed problem-specific gradient or subgradient constructions for certain models (Huh and Rusmevichientong 2009), we employ finite-difference estimators for broader applicability.
- Random Policy: A baseline that selects actions uniformly at random.
- Problem-Specific: Algorithms tailored to each case study.
  - Inventory Control: The algorithm of Agrawal and Jia (2022) proposes a convexity-based policy-elimination method for the inventory problem (Section 4.1), maintaining a one-dimensional confidence interval that shrinks at each epoch (whereas the confidence sets of IOPEA need not be intervals).
  - Dual Sourcing: The BASA method of Chen and Shi (2025), which solves the lost-sales dual-sourcing problem (Section 4.2) over tailored base-surge policies using stochastic gradient descent and convexity. We note that tailored base-surge policies and dual index policies are two different policy classes.
  - Queuing Model: UCRL2 (Anselmi et al. 2022), which addresses the service-rate control problem (Section 4.3) by building confidence sets over MDPs and using extended value iteration to select the most optimistic model in each epoch.
- Optimal In-Class Policy: The best policy within the given policy class, computed with full knowledge of the system.

- ERM-Based: An algorithm adapted from Sinclair et al. (2023b), which leverages full-feedback information. The ERM-Based algorithm assumes access to unobserved randomness thus is not implementable in practice.

Although it lacks theoretical guarantees, we include PPO (Schulman et al. 2017) as a representative empirical RL method.

Table 1 Average cost of the learned policy at the timestep g_{θ_T} for $T = 10^5$ (small-scale) and $T = 3 \times 10^5$ (large-scale) achieved by each algorithm, under three demand distributions. The queuing case study is estimated at $T = 3 \times 10^5$. $\star$ indicates a significant improvement and $\circ$ a significant decrease over Problem-Specific by Welch's t-test with $p < 0.05$. In parentheses we report relative performance $(g_{\theta_T} - g_{\theta^*})/g_{\theta^*}$. Feedback Graph is omitted in large-scale due to compute constraints. The performance of SGD is omitted in the queuing models as it is essentially equivalent to PPO. See Table EC.2 for scenario hyperparameters.

Distribution	Algorithm	Inventory Control		Dual Sourcing		Queuing
		Small	Large	Small	Large	Large
Exponential	Optimal (g_{θ^*})	2.5	39.1	1.9	20.3	9.5
	ERM (Sinclair et al. 2023b)	2.6(2%)	39.2(0%)	1.9(1%)	20.3(0%)	9.6(1%)
	IOPEA	$\star$2.6(3%)	40.1(2%)	$\star$2.0(5%)	$\star$26.0(28%)	$\star$10.5(11%)
	PPO (Schulman et al. 2017)	3.8(49%)	$\circ$132.4(238%)	2.9(53%)	$\circ$112.0(451%)	$\star$9.8(3%)
	SGD	$\star$2.6(3%)	$\circ$43.0(10%)	$\star$2.1(10%)	$\circ$55.2(172%)	–
	Feedback Graph (Dann et al. 2020)	$\circ$5.2(103%)	–	$\star$2.4(26%)	–	11.5(21%)
	Problem-Specific	4.1(60%)	40.3(3%)	3.4(78%)	30.1(48%)	11.3(19%)
	Random	$\star$3.6(43%)	$\circ$182.4(366%)	3.7(95%)	$\circ$121.5(498%)	11.1(17%)
Normal	Optimal (g_{θ^*})	2.2	39.9	1.3	17.2	–
	ERM (Sinclair et al. 2023b)	2.2(1%)	40.3(1%)	1.3(3%)	17.5(2%)	–
	IOPEA	$\star$2.3(1%)	$\star$40.2(1%)	$\star$1.3(3%)	$\star$17.7(3%)	–
	PPO (Schulman et al. 2017)	$\circ$3.4(52%)	$\circ$133.6(235%)	$\circ$2.3(76%)	$\circ$148.8(765%)	–
	SGD	$\star$2.3(4%)	$\star$41.9(5%)	1.8(31%)	$\circ$45.1(162%)	–
	Feedback Graph (Dann et al. 2020)	$\circ$5.8(158%)	–	1.7(30%)	–	–
	Problem-Specific	2.9(29%)	44.9(13%)	1.6(22%)	18.8(9%)	–
	Random	$\circ$3.4(54%)	$\circ$96.5(142%)	$\circ$3.0(127%)	$\circ$107.4(524%)	–
Uniform	Optimal (g_{θ^*})	2.9	49.0	1.7	22.9	–
	ERM (Sinclair et al. 2023b)	3.0(1%)	49.5(1%)	1.7(1%)	23.2(1%)	–
	IOPEA	$\star$3.0(2%)	49.5(1%)	$\star$1.7(1%)	$\star$23.5(3%)	–
	PPO (Schulman et al. 2017)	$\circ$4.1(41%)	$\circ$157.3(221%)	$\circ$3.7(116%)	$\circ$147.7(545%)	–
	SGD	3.1(7%)	$\circ$51.5(5%)	$\circ$2.7(59%)	$\circ$37.5(64%)	–
	Feedback Graph (Dann et al. 2020)	$\circ$5.8(99%)	–	2.6(53%)	–	–
	Problem-Specific	3.2(11%)	49.6(1%)	2.3(37%)	30.0(30%)	–
	Random	$\circ$3.9(34%)	$\circ$136.3(178%)	$\circ$3.1(81%)	$\circ$108.8(375%)	–

Simulation Results. In Table 1 we include a comparison of the algorithm performance at $T = 10^5$ (small scale) and $T = 3 \times 10^5$ (large-scale). We use Welch's t-test, at a significance level of 0.05, to assess whether each algorithm differs significantly from Problem-Specific in each setting. In parentheses we report relative performance $(g_{\theta_T} - g_{\theta^*})/g_{\theta^*}$. Details on the hyperparameters and problem specifications are in Section B.1. We also provide the numerical experiment results for the queuing case study under the fixed arrival rate in Section B.2.

First, we observe that IOPEA nearly matches the performance of ERM-Based algorithm (Sinclair et al. 2023b) despite only having access to partial feedback. For example, in the small-scale inventory

control problem, IOPEA achieves within 3% of the optimal cost, while PPO and the Problem-Specific algorithm (Agrawal and Jia 2022) lag behind at 49% and 60%, respectively. Also, IOPEA delivers stable and competitive performance, consistently matching or outperforming Problem-Specific algorithms. In contrast, the performances of certain baselines are unstable: The convergence behavior of Feedback Graph (Dann et al. 2020) depends strongly on the size of the state space. It excels in dual-sourcing but struggles in inventory control, where the state space is roughly 10 times larger, and performs even better in the queuing problem with the smallest state space.

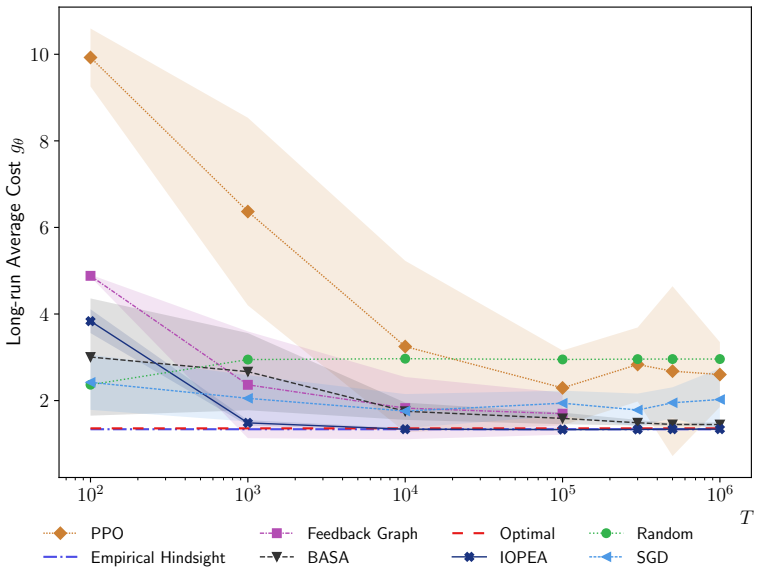

Figure 1 **Comparison of the long-run average cost g_θ over time T for different algorithms in the small-scale dual-sourcing problem up to $T = 10^6$. We assume the demand follows the normal distribution as in Table EC.2. For Feedback Graph (Dann et al. 2020), data are shown only to $T = 10^5$. IOPEA attains near-optimal gain by $T = 10^4$; BASA (Chen and Shi 2025) remains suboptimal and continues to decline even at $T = 10^6$; Feedback Graph matches BASA's long-run cost at $T = 10^5$ but with much larger variance; PPO converges markedly slower than all other methods. SGD fails to converge effectively, as the problem is non-convex. Note that in some trials PPO and Feedback Graph may exceed the "Optimal" curve, because "Optimal" denotes the best dual index policy rather than the true problem-wide optimum.**

Second, IOPEA converges substantially faster, indicating superior sample efficiency: by $T = 3 \times 10^5$ in the large-scale regime, its average cost has stabilized in nearly all simulation settings, whereas all other algorithms require $T > 10^6$ to achieve similar performance. Further evidence of IOPEA's rapid convergence is presented in Figure 1. In the small-scale dual sourcing problem, IOPEA achieves near-optimal performance by $T = 10^4$ and remains stable thereafter, converging more quickly than all other algorithms we evaluated.

An intuitive explanation for the fast convergence of IOPEA lies in the policy order: the rich feedback structure it leverages improves sample efficiency, as observations collected under one policy can be reused to estimate the performance of others, thereby accelerating convergence. A promising direction for future work is to compare the sample efficiency of low-width policy orders with stochastic gradient descent methods (Huh and Rusmevichientong 2009, Agrawal and Jia 2022) in convex models; Interestingly, in our single-channel inventory experiments (which is convex), IOPEA even outperforms gradient-based methods in such settings, despite not leveraging convexity.

Moreover, IOPEA is highly scalable. The training of IOPEA completes in under three minutes in all simulation experiments, whereas PPO and Feedback Graph are significantly slower or infeasible. This scalability stems from the use of the information order, which enables focused exploration. While computing the policy order and associated counterfactual estimates is lightweight during training, it does require domain-specific insight into the structure of the policy class to obtain the policy order.

Suboptimality of BASA (Chen and Shi 2025). While BASA (Chen and Shi 2025) also converges quickly in the dual-sourcing problem (see Fig. 1), it learns over the class of tailored base-surge policies. These are known to underperform dual index policies in most non-asymptotic settings (i.e. when $L_r < \infty$ (Xin and Van Mieghem 2023)). This limitation is reflected in the observed performance gap between IOPEA and BASA in our simulations, since BASA converges to a significantly suboptimal policy. This reinforces the practical advantage of learning within the stronger-performing policy classes, such as dual index policies.

Suboptimality of SGD. We observe three main limitations of SGD. First, it requires substantially more samples to converge. In the small-scale inventory control problem, SGD performs comparably to IOPEA, but its performance deteriorates significantly in the large-scale setting. Since this model is convex and SGD is theoretically guaranteed to converge to the optimum, this gap highlights its slower convergence rate. Second, the performance of SGD is highly sensitive to hyperparameter choices, such as the initial point and learning rate (see Section B.1 for details). Finally, in the dual sourcing problem, SGD performs markedly worse, as expected, since the model is no longer convex. In contrast, our method IOPEA consistently maintains strong performance in the dual sourcing setting.

Computational Limitations of Feedback Graph (Dann et al. 2020). Note that for the large-scale inventory control and dual sourcing problem, we do not include the performance of Feedback Graph, since this model-based approach requires storing a $O(|S|^2|A|)$ length vector for updating transition dynamics during training. In our large-scale inventory control problem with $L = 6$ and demand mean $40/3$, a reasonable discretization of the state space with radius as 0.1 leads to a state space of size $\Theta(10^{14})$ after discretization, which is not tractable. Hence, we omit these comparisons.

6. Conclusion

In this paper, we introduced the novel concept of an *information order* over a class of policies in infinite-horizon average-cost MDPs, and showed how to leverage this partial ordering to improve sample efficiency in learning. Our main theoretical contribution is a regret bound that scales as $\tilde{O}\big(\sqrt{\frac{wdT}{\alpha}}\big)$, where w is the width of the information-order, which interpolates smoothly between the no-feedback ($w = |\Theta|$) and full-feedback ($w = 1$) regimes.

We instantiated our Information-Ordered Epoch-Based Policy Elimination Algorithm (IOPEA) on three canonical operations research problems: inventory control with positive lead time, dual sourcing, and an M/M/1/L queue with controllable service rates, obtaining $\tilde{O}(\sqrt{T})$ regret guarantees in each setting with a low-width policy order. Notably, as far as we know, this is the first result obtaining $\sqrt{T}$-regret for dual index policies in lost-sales dual sourcing problem. Numerical simulations confirm that IOPEA attains performance close to the theoretical optimum and significantly outperforms baseline methods.

Beyond these case studies, our policy-order framework offers a unified approach to leveraging partial feedback in a wide range of decision-making problems and opens several avenues for future research. One could investigate policy ordering and counterfactual estimation in nonstationary environments, such as inventory systems with time-varying demand distributions. Another future direction is to address settings where observations from a single policy θ only partially inform about another policy θ'. In such cases, producing an accurate counterfactual estimate for $\pi_{\theta'}$ requires implementing multiple policies and collecting data under each one. This challenge appears in many operations research problems, such as online assortment optimization, where different policies have overlapping but nonidentical informational content.

References

Agrawal S, Jia R (2022) Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Operations Research* 70(3):1646–1664.

Alon N, Cesa-Bianchi N, Dekel O, Koren T (2015) Online learning with feedback graphs: Beyond bandits. *Conference on Learning Theory*, 23–35 (PMLR).

Alvo M, Russo D, Kanoria Y (2023) Neural inventory control in networks via hindsight differentiable policy optimization. *arXiv preprint arXiv:2306.11246* .

Anselmi J, Gaujal B, Rebuffi LS (2021) Optimal speed profile of a dvfs processor under soft deadlines. *Performance Evaluation* 152:102245.

Anselmi J, Gaujal B, Rebuffi LS (2022) Reinforcement learning in a birth and death process: breaking the dependence on the state space. *Advances in Neural Information Processing Systems* 35:14464–14474.

Auer P, Jaksch T, Ortner R (2008) Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems* 21.

Azar MG, Osband I, Munos R (2017) Minimax regret bounds for reinforcement learning. *International Conference on Machine Learning*, 263–272 (PMLR).

Bartlett PL, Tewari A (2012) Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661* .

Bertsimas D, Vayanos P (2017) Data-driven learning in dynamic pricing using adaptive optimization. *Optimization Online* 20.

Besbes O, Muharremoglu A (2013) On implications of demand censoring in the newsvendor problem. *Management Science* 59(6):1407–1424.

Bi S, He L, Teo CP (2023) Taming the long tail: The gambler's fallacy in intermittent demand management. *Manufacturing & Service Operations Management* 25(5):1692–1710.

Cao J, Sun W (2019) Dynamic learning with frequent new product launches: A sequential multinomial logit bandit problem. *International Conference on Machine Learning*, 912–920 (PMLR).

Chen B, Jasin S, Luo Q, Zhang M (2025) Managing lost-sale inventory systems under unknown demand and return distributions. *Available at SSRN 5624311* .

Chen B, Li M, Simchi-Levi D (2022) Dynamic pricing with infrequent inventory replenishments. *Available at SSRN 4240137* .

Chen B, Shi C (2025) Tailored base-surge policies in dual-sourcing inventory systems with demand learning. *Operations Research* 73(4):1723–1743.

Chen H, He Y, Zhang C (2024) On interpolating experts and multi-armed bandits. *International Conference on Machine Learning*, 6776–6802 (PMLR).

Chen T, Yu S (2021) An eoq model with stepwise ordering cost and the finite planning horizon under carbon cap-and-trade regulations. *Decis. Sci. Lett* 10:337–350.

Chen Z, Maguluri ST, Shakkottai S, Shanmugam K (2021) A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567* .

Chong JW, Kim W, Hong J (2022) Optimization of apparel supply chain using deep reinforcement learning. *IEEE Access* 10:100367–100375.

Dai JG, Gluzman M (2022) Queueing network controls via deep reinforcement learning. *Stochastic Systems* 12(1):30–67.

Dann C, Mansour Y, Mohri M, Sekhari A, Sridharan K (2020) Reinforcement learning with feedback graphs. *Advances in Neural Information Processing Systems* 33:16868–16878.

Drekic S, Spivey MZ (2021) On the number of trials needed to obtain k consecutive successes. *Statistics & Probability Letters* 176:109132.

Dushnik B, Miller EW (1941) Partially ordered sets. *American journal of mathematics* 63(3):600–610.

Eisenach C, Ghai U, Madeka D, Torkkola K, Foster D, Kakade S (2024) Neural coordination and capacity control for inventory management. *arXiv preprint arXiv:2410.02817* .

Eldowa K, Esposito E, Cesari T, Cesa-Bianchi N (2023) On the minimax regret for online learning with feedback graphs. *Advances in Neural Information Processing Systems* 36:46122–46133.

Fang J, Ellis M, Li B, Liu S, Hosseinkashi Y, Revow M, Sadovnikov A, Liu Z, Cheng P, Ashok S, Zhao D, Cutler R, Lu Y, Gehrke J (2019) Reinforcement learning for bandwidth estimation and congestion control in real-time communications. *arXiv preprint arXiv:1912.02222* .

Feng J, Gluzman M, Dai JG (2021) Scalable deep reinforcement learning for ride-hailing. *2021 American Control Conference (ACC)*, 3743–3748 (IEEE).

Fruit R, Pirotta M, Lazaric A, Ortner R (2018) Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *International Conference on Machine Learning*, 1578–1586 (PMLR).

Goldberg DA, Reiman MI, Wang Q (2021) A survey of recent progress in the asymptotic analysis of inventory systems. *Production and Operations Management* 30(6):1718–1750.

Gong XY, Simchi-Levi D (2020) Provably more efficient q-learning in the one-sided-feedback/full-feedback settings. *arXiv preprint arXiv:2007.00080* .

Gong XY, Simchi-Levi D (2024) Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Management Science* 70(9):6139–6157.

Hahn G, Leucht A (2015) Managing inventory systems of slow-moving items. *International Journal of Production Economics* 170(PB):543–550.

Hssaine C, Hu Y, Pike-Burke C (2025) Learning fair and effective points-based rewards programs. *arXiv preprint arXiv:2506.03911* .

Hssaine C, Sinclair SR (2024) The data-driven censored newsvendor problem. *arXiv preprint arXiv:2412.01763* .

Hua Z, Yu Y, Zhang W, Xu X (2015) Structural properties of the optimal policy for dual-sourcing systems with general lead times. *IIE Transactions* 47(8):841–850.

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009) Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Science* 55(3):404–420.

Huh WT, Rusmevichientong P (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research* 34(1):103–123.

Janakiraman G, Seshadri S (2017) Dual sourcing inventory systems: On optimal policies and the value of costless returns. *Production and Operations Management* 26(2):203–210.

Jia H, Shi C, Shen S (2024) Online learning and pricing for service systems with reusable resources. *Operations Research* 72(3):1203–1241.

Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is q-learning provably efficient? *Advances in neural information processing systems* 31.

Jin C, Yang Z, Wang Z, Jordan MI (2020) Provably efficient reinforcement learning with linear function approximation. *Conference on Learning Theory*, 2137–2143 (PMLR).

Karlin S (1958) Inventory models of the arrow-harris-marschak type with time lag. *Studies in the mathematical theory of inventory and production* .

Kearns M, Singh S (2002) Near-optimal reinforcement learning in polynomial time. *Machine learning* 49:209–232.

Khodadadian S, Chen Z, Maguluri ST (2021) Finite-sample analysis of off-policy natural actor-critic algorithm. *International Conference on Machine Learning*, 5420–5431 (PMLR).

Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press).

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research* 32(4):821–839.

Li Q, Yu P (2014) Multimodularity and its applications in three stochastic dynamic inventory problems. *Manufacturing & Service Operations Management* 16(3):455–463.

Madeka D, Torkkola K, Eisenach C, Luo A, Foster DP, Kakade SM (2022) Deep inventory management. *arXiv preprint arXiv:2210.03137* .

Mannor S, Shamir O (2011) From bandits to experts: On the value of side-observations. *Advances in neural information processing systems* 24.

Meyn SP, Tweedie RL (2012) *Markov chains and stochastic stability* (Springer Science & Business Media).

Mitrophanov AY (2003) Stability and exponential convergence of continuous-time markov chains. *Journal of applied probability* 40(4):970–979.

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. (2015) Human-level control through deep reinforcement learning. *nature* 518(7540):529–533.

Osband I, Van Roy B (2014) Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems* 27.

Osband I, Van Roy B (2016) On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732* .

Osband I, Van Roy B, Wen Z (2016) Generalization and exploration via randomized value functions. *International Conference on Machine Learning*, 2377–2386 (PMLR).

Puterman ML (2014) *Markov decision processes: Discrete stochastic dynamic programming* (John Wiley & Sons).

Rolf B, Jackson I, Müller M, Lang S, Reggelin T, Ivanov D (2023) A review on reinforcement learning algorithms and applications in supply chain management. *International Journal of Production Research* 61(20):7151–7179.

Ross SM (2014) *Introduction to probability models* (Academic press).

Russo D, Van Roy B (2018) Learning to optimize via information-directed sampling. *Operations Research* 66(1):230–252.

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. (2016) Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484.

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, et al. (2017) Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* .

Sinclair SR, Banerjee S, Yu CL (2023a) Adaptive discretization in online reinforcement learning. *Operations Research* 71(5):1636–1652.

Sinclair SR, Frujeri FV, Cheng CA, Marshall L, Barbalho HDO, Li J, Neville J, Menache I, Swaminathan A (2023b) Hindsight learning for mdps with exogenous inputs. *International Conference on Machine Learning*, 31877–31914 (PMLR).

Snyder RD, Ord JK, Beaumont A (2012) Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting* 28(2):485–496.

Strehl AL, Li L, Wiewiora E, Langford J, Littman ML (2006) Pac model-free reinforcement learning. *Proceedings of the 23rd international conference on Machine learning*, 881–888.

Tang J, Chen B, Shi C (2024) Online learning for dual-index policies in dual-sourcing systems. *Manufacturing & Service Operations Management* 26(2):758–774.

Veeraraghavan S, Scheller-Wolf A (2008) Now or later: A simple policy for effective dual sourcing in capacitated systems. *Operations Research* 56(4):850–864.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Vinyals O, Babuschkin I, Chung J, Mathieu M, Jaderberg M, Czarnecki WM, Dudzik A, Huang A, Georgiev P, Powell R, Ewalds T, Horgan D, Kroiss M, Danihelka I, Agapiou J, Oh J, Dalibard V, Choi D, Sifre L, Sulsky Y, Vezhnevets S, Molloy J, Cai T, Budden D, Paine T, Gulcehre C, Wang Z, Pfaff T, Pohlen T, Wu Y, Yogatama D, Cohen J, McKinney K, Smith O, Schaul T, Lillicrap T, Apps C, Kavukcuoglu K, Hassabis D, Silver D (2019) AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

Vinyals O, Ewalds T, Bartunov S, Georgiev P, Vezhnevets AS, Yeo M, Makhzani A, Küttler H, Agapiou J, Schrittwieser J, et al. (2017) Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* .

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).

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

Wan J, Sinclair SR, Shah D, Wainwright MJ (2024) Exploiting exogenous structure for sample-efficient reinforcement learning. *arXiv preprint arXiv:2409.14557* .

Wang Y, Wang R, Du SS, Krishnamurthy A (2019) Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136* .

Wei CY, Jahromi MJ, Luo H, Sharma H, Jain R (2020) Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. *International conference on machine learning*, 10170–10180 (PMLR).

Whittemore AS, Saunders S (1977) Optimal inventory under stochastic demand with two supply options. *SIAM Journal on Applied Mathematics* 32(2):293–305.

Xin L, Van Mieghem JA (2023) Dual-sourcing, dual-mode dynamic stochastic inventory models. *Research Handbook on Inventory Management*, 165–190 (Edward Elgar Publishing).

Yuan H, Luo Q, Shi C (2021) Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science* 67(10):6089–6115.

Zhang M, Ahn HS, Uichanco J (2022) Data-driven pricing for a new product. *Operations Research* 70(2):847–866.

Zhao H, Chen W (2019) Stochastic one-sided full-information bandit. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 150–166 (Springer).

Zipkin P (2008) Old and new methods for lost-sales inventory systems. *Operations research* 56(5):1256–1263.

Table EC.1 Common notation

Symbol	Definition
	Problem setting specifications
T	Total time horizon
$X, Y, X \in Y$	Two random variables, and shorthand for X measurable with respect to Y
$\mathcal{S}, \mathcal{A}, P, C, s_1$	State and action space, transition distribution, cost function, starting state
Θ, θ	Policy class, and arbitrary policy in the policy class
π_θ	$\pi_\theta : \mathcal{S} \to \mathcal{A}$, policy parameterized by $\theta \in \Theta$
S_t, θ_t, A_t	State, selected policy parameter, and action at time t where $A_t = \pi_{\theta_t}(S_t)$
$g_\theta(s), v_\theta(s)$	Long-run average cost and bias for policy π_θ
θ^*	Optimal policy in Θ
$\mathsf{Regret}(T)$	$\sum_{t=1}^{T} C(S_t, A_t) - T g_{\theta^*}(s_1)$
$\mathcal{H}_\theta^T$	Sample path trajectory of length T collected via π_θ
$G_\theta(\mathcal{H}_\theta^T)$	Empirical average cost of policy θ
$\preceq, w, \alpha$	Policy order relation, width, and "informativeness" factor
$\tilde{G}_{\theta' \mid \theta}(\mathcal{H}_\theta^T)$	Counterfactual estimate of θ' from θ
$T_h(\delta)$	A constant (horizon threshold) that depends on the confidence level δ
U, H	Upper bound on Θ, and bound on the bias
L_Θ, θ_R	Lipschitz constant of g_θ, and the policy with finite expected return time to s_1
	Algorithm specification
k, Θ_k	Epoch and confidence set in epoch k
Θ_k^{max}	Set of maximal policies according to $\preceq$ to Θ_k
r	Discretization parameter
N_k, β_k	Number of trajectories sampled in epoch k, and confidence term
	Inventory case study
I_t, Q_t	Inventory level and ordering level at time t
D_t	Demand at time t
L, L_r, L_e	Lead time parameters
h, p	Holding and lost-sales cost
c_r, c_e	The long-lead purchasing cost and short-lead purchasing cost coefficient in dual-sourcing
Q_t^e, Q_t^r	Order of expedited and regular channels
γ	$\mathbb{P}(D_t = 0)$
	Queuing case study
L, λ, μ	Buffer size, arrival rate, deadline rate
$C, w(a), A_{\max}$	Cost for missed deadline, power cost, and maximum service rate
$U, \lambda_{\max}, \mu_{\max}$	Uniformization parameter, with known upper bounds on λ and μ
$m_\theta(s)$	Stationary distribution of policy θ
V	The confidence interval in the policy order

Appendix A: Proof of Theorems in Section 3

A.1. Regret Upper Bound (Theorem 1)

We start off this section by providing a complete proof of the regret upper bound of our algorithm IOPEA (see Algorithm 1).

THEOREM 1 **(Regret Upper Bound)**. *Let w denote the width of the partial order $\preceq$ restricted to $\Theta_1 = N_r(\Theta)$ for $r = T^{-1/2}$. Then under Assumptions 1 to 3, for any $\delta > 0$, by setting $\beta_k = \frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4|\mathcal{N}_r(\Theta)|K/\delta)}{\alpha N_k}}$ and $N_k = \frac{4^k T_h(\delta)}{\alpha}\log(T)$, Algorithm 1 ensures that, for any $T \geq \frac{4wT_h(\delta)}{\alpha}$, with probability at least $1-3\delta$,*

$$\textit{Regret}(T) = O\left(\left((H+2)\sqrt{\frac{wd}{\alpha}\log\left(\frac{U\sqrt{T}\log T}{\delta}\right)}\log(T) + L_\Theta\right)\sqrt{T} + \left(\frac{wH}{\alpha} + wD_\Theta\right)\log(T)\right), \tag{6}$$

where $O(\cdot)$ hides absolute constants and logarithmic factors of α, $1/w$, and $1/T_h(\delta)$.

Note that the regret bound for IOPEA explicitly relies on the choice of policy order. While all policy classes have a trivial information order, the resulting regret guarantee will be exponential with respect to d. Hence, our results require a policy class that is sufficiently "informative" (i.e. of low width). See Section 3 (Choice of Policy Order) for more discussion on the choice of policy order and its implications on regret.

We emphasize here that the total number of epochs K is trivially upper bounded by T. Our result requires a series of lemmas. The first one establishes concentration on the estimates $\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$ versus $g_{\theta'}$ in terms of the number of "useful" samples in the epoch αN_k.

LEMMA A.1. *Recall that with fixed $\alpha \in (0,1]$, $\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$ is the counterfactual average cost estimate of policy θ' from observed sample trajectories under the maximal policy $\tilde{\theta}_k(\theta') \in \Theta_k^{max}$. For any fixed $\delta > 0$, define the following event:*

$$\mathcal{E}_0 = \left\{\forall k, \theta' \in \Theta_k \mid |\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - g_{\theta'}| \leq \frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4|\mathcal{N}_r(\Theta)|K/\delta)}{\alpha N_k}}\right\}.$$

Then we have that $\mathcal{E}_0$ occurs with probability at least $1-2\delta$.

We establish Lemma A.1 by showing that the counterfactual estimate $\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$ concentrates around the true cost $g_{\theta'}$ uniformly over all $\theta' \in \mathcal{N}_r(\Theta)$ and all epochs $k \in [K]$ with high probability. The main idea is to apply the concentration bound of a Markov process (Lemma 3 of Agrawal and Jia (2022)), restated in Lemma A.2 for completeness, together with the fact that $\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$ approximates the empirical estimate of $G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha T})$ (as guaranteed by the policy order in Definition 3). The final result follows via a union bound over both K and the covering size $|\mathcal{N}_r(\Theta)|$.

LEMMA A.2 **(Lemma 3 of Agrawal and Jia (2022)).** *Suppose that for any given policy θ', the gain function $g_{\theta'}(s)$ is constant across all states and the span of the bias is at most H. Then, given a sample trajectory $\mathcal{H}_{\theta'}^T$ of length T, we have that, for any $\delta > 0$, with probability at least $1-\delta$,*

$$|G_{\theta'}(\mathcal{H}_{\theta'}^T) - g_{\theta'}(s_1)| \le \frac{H}{T} + (H+2)\sqrt{\frac{2\log(4/\delta)}{T}}.$$

Proof. We recall the Lemma 8.2.6 from Puterman (2014) first:

LEMMA A.3 **(Restate the Theorem 8.2.6 of Puterman (2014)).** *For any fixed policy θ', for any t, and for any state $S_t \in \mathcal{S}$, the gain and bias satisfy*

$$g_\theta(S_t) = \mathbb{E}_{\theta'}[C(S_t, \pi_{\theta'}(S_t))] + \mathbb{E}_{S'\sim P_{\theta'}(S_t)}[v_{\theta'}(S')] - v_{\theta'}(S_t).$$

We bound $\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\theta'}[C(S_t,\pi_{\theta'}(S_t))] - g_{\theta'}\right|$ first. Denote $S_1,\dots,S_T$ as the observed states within the sample trajectory $\mathcal{H}_{\theta'}^T$,

$$\begin{aligned}
\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\theta'}\left[C(S_t,\pi_{\theta'}(S_t))\right] - g_{\theta'}\right| &= \frac{1}{T}\left|\sum_{t=1}^T \left(\mathbb{E}_{\theta'}[C(S_t,\pi_{\theta'}(S_t))] - g_{\theta'}\right)\right| \\
&= \frac{1}{T}\Big|\sum_{t=1}^T \Big(\mathbb{E}_{\theta'}[C(S_t,\pi_{\theta'}(S_t))] \\
&\quad - \left[\mathbb{E}_{\theta'}[C(S_t,\pi_{\theta'}(S_t))] + \mathbb{E}_{S'\sim P_{\theta'}(S_t)}[v_{\theta'}(S')] - v_{\theta'}(S_t)\right]\Big)\Big| \\
&= \frac{1}{T}\left|\sum_{t=1}^T \left(v_{\theta'}(S_t) - \mathbb{E}_{S'\sim P_{\theta'}(S_t)}[v_{\theta'}(S')]\right)\right| \\
&\le \frac{1}{T}\left|v_{\theta'}(S_1) - \mathbb{E}_{S'\sim P_{\theta'}(S_T)}[v_{\theta'}(S')]\right| \\
&\quad + \frac{1}{T}\left|\sum_{t=1}^{T-1}\left(v_{\theta'}(S_{t+1}) - \mathbb{E}_{S'\sim P_{\theta'}(S_t)}[v_{\theta'}(S')]\right)\right| \\
&\le \frac{H}{T} + \frac{1}{T}\left|\sum_{t=1}^{T-1}\Delta_{t+1}\right|,
\end{aligned}$$

where the second line is due to Lemma A.3 and Assumption 3, and we set

$$\Delta_{t+1} := v_{\theta'}(S_{t+1}) - \mathbb{E}_{S'\sim \mathbb{P}_{\theta'}(S_t)}[v_{\theta'}(S')].$$

Noting $\mathbb{E}[\Delta_{t+1}|S_t] = 0$ and $|\Delta_{t+1}| \le H$, Azuma–Hoeffding (Wainwright 2019) yields for any $\epsilon > 0$,

$$\mathbb{P}\Big(\Big|\sum_{t=2}^T \Delta_t\Big| \ge \epsilon\Big) \le 2\exp\Big(-\frac{\epsilon^2}{2(T-1)H^2}\Big).$$

By setting $\epsilon = H\sqrt{2(T-1)\log(2/\delta)}$, we have that with probability at least $1-\delta$,

$$\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\theta'}\left[C(S_t,\pi_{\theta'}(S_t))\right] - g_{\theta'}\right| \le \frac{H}{T} + H\sqrt{\frac{2\log(2/\delta)}{T}}. \tag{EC.1}$$

Now we bound $|G_{\theta'}(\mathcal{H}_{\theta'}^T) - g_{\theta'}(s_1)|$ correspondingly. Define $X_t = C(S_t, \pi_{\theta'}(S_t)) - \mathbb{E}_{\theta'}[C(S_t, \pi_{\theta'}(S_t))]$, then

$$|X_t| \leq |C(S_t, \pi_{\theta'}(S_t))| + |\mathbb{E}_{\theta'}[C(S_t, \pi_{\theta'}(S_t))]| \leq 2.$$

Hence we have that $\{X_t\}_{t=1}^T$ is also a bounded martingale difference sequence. With Azuma–Hoeffding (Wainwright 2019), with probability at least $1 - \frac{\delta}{2}$,

$$\frac{1}{T}\Big|\sum_{t=1}^{T} C(S_t, \pi_{\theta'}(S_t)) - \mathbb{E}_{\theta'}[C(S_t, \pi_{\theta'}(S_t))]\Big| \leq 2\sqrt{\frac{2\log(4/\delta)}{T}}. \tag{EC.2}$$

The proof of Lemma A.2 is complete by combining Equation (EC.1) and (EC.2) above. □

With the previous lemma in hand, we are now ready to show Lemma A.1.

Proof of Lemma A.1. Observe that each epoch k begins from the same fixed initial state s_1, regardless of the policy being implemented. As a counterfactual, consider executing policy θ' for $\alpha N_k = 4^k T_h(\delta)\log(T)$ steps during epoch k, in place of the maximal policies in $\Theta_k^{\max}$. This would yield an empirical estimate $G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha N_k})$ of $g_{\theta'}$, which concentrates around $g_{\theta'}$ by Lemma A.2:

$$|G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha N_k}) - g_{\theta'}| \leq \frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4/\delta)}{\alpha N_k}}.$$

However, the policy θ' may not be actually implemented in epoch k, as θ' may not belong to $\Theta_k^{\max}$. Due to the policy order, we know that there exists $\tilde{\theta}_k(\theta') \in \Theta_k^{\max}$ such that $\pi_{\theta'} \preceq \pi_{\tilde{\theta}_k(\theta')}$. Under Assumption 3, note that the average cost $g_{\theta'}$ has no dependence on the state s, so $[g_{\theta'} - (\frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4/\delta)}{\alpha N_k}}), g_{\theta'} + (\frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4/\delta)}{\alpha N_k}})]$ is Borel measurable. Then due to $N_k = \frac{4^k T_h(\delta)}{\alpha} \geq T_h(\delta)$, we have

$$d_{\mathrm{TV}}(\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}, G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha N_k})) \leq \delta.$$

By the definition of the total variation distance we get

$$\mathbb{P}\left(|\tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - g_{\theta'}| \leq \frac{H}{\alpha N_k} + (H+2)\sqrt{\frac{2\log(4/\delta)}{\alpha N_k}}\right) \geq 1 - \delta - \delta = 1 - 2\delta.$$

The final result follows via a union bound over both K and $|\mathcal{N}_r(\Theta)|$. □

Similarly, we have concentration for the empirical costs corresponding to the actual policies implemented from $\Theta_k^{\max}$ in each epoch k. Note that we analyze the costs for these policies separately since we can directly construct $G_\theta(\mathcal{H}_\theta^{N_k})$ as these policies were employed, instead of using the distributional policy order.

LEMMA A.4. *Define the following event*

$$\mathcal{E}_1 = \left\{\forall k, \theta \in \Theta_k^{\max} \,|\, |G_\theta(\mathcal{H}_\theta^{N_k}) - g_\theta| \leq \frac{H}{N_k} + (H+2)\sqrt{\frac{2\log(4wK/\delta)}{N_k}}\right\}.$$

Then we have that $\mathcal{E}_1$ occurs with probability at least $1-\delta$.

This lemma follows immediately from Lemma A.2 and a union bound.

Without loss of generality, we assume that $w \leq |\mathcal{N}_r(\Theta)|$ and note that $\alpha \leq 1$. Applying a union bound over the events $\mathcal{E}_0$ and $\mathcal{E}_1$ then yields the following:

LEMMA A.5. *Define the event*

$$\mathcal{E}_2 = \left\{ \forall k,\, \theta' \in \Theta_k,\, \theta \in \Theta_k^{\mathsf{max}} \;\middle|\; \left| \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - g_{\theta'} \right| \leq \beta_k \text{ and } \left| G_\theta(\mathcal{H}_\theta^{N_k}) - g_\theta \right| \leq \beta_k \right\}.$$

Then $\mathcal{E}_2$ holds with probability at least $1 - 3\delta$.

Next we establish that under the event $\mathcal{E}_2$, IOPEA maintains that the optimal parameter $\theta_r^* \in \Theta_k$ for each k with high probability, where

$$\theta_r^* = \operatorname*{arg\,min}_{\theta \in \mathcal{N}_r(\Theta)} g_\theta.$$

We similarly establish that the average cost of *all* policies contained in Θ_k is bounded by $O(\beta_k)$.

LEMMA A.6. *Under the event $\mathcal{E}_2$ we have that $\theta_r^* \in \Theta_k$ for all k. Moreover, for any two $\theta, \theta' \in \Theta_k^2$ we have that:*

$$|g_\theta - g_{\theta'}| \leq 6\beta_k.$$

Proof. We first start by showing that $\theta_r^* \in \Theta_k$ for all k. We show this via induction over the epochs k. Clearly for the base case when $k = 1$ we have that $\theta_r^* \in \mathcal{N}_r(\Theta) = \Theta_1$. For the step case $k \to k+1$ suppose that $\theta_r^* \in \Theta_k$. Then we have that if $\hat{\theta}_k = \arg\min_{\theta' \in \Theta_k} \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}$:

$$\begin{aligned}
\tilde{G}_{\theta_r^*|\tilde{\theta}_k(\theta_r^*)} - \tilde{G}_{\hat{\theta}_k|\tilde{\theta}_k(\hat{\theta}_k)} &= \tilde{G}_{\theta_r^*|\tilde{\theta}_k(\theta_r^*)} - g_{\theta_r^*} + g_{\theta_r^*} - g_{\hat{\theta}_k} + g_{\hat{\theta}_k} - \tilde{G}_{\hat{\theta}_k|\tilde{\theta}_k(\hat{\theta}_k)} \\
&\leq \tilde{G}_{\theta_r^*|\tilde{\theta}_k(\theta_r^*)} - g_{\theta_r^*} + g_{\hat{\theta}_k} - \tilde{G}_{\hat{\theta}_k|\tilde{\theta}_k(\hat{\theta}_k)} \\
&\leq 2\beta_k.
\end{aligned}$$

Note that in the first inequality we used that $g_{\theta_r^*} \leq g_{\hat{\theta}_k}$ since θ_r^* is the optimizer over $\mathcal{N}_r(\Theta)$, and the second inequality is from the definition of the event $\mathcal{E}_2$.

To show the second property we note that for any $\theta, \theta' \in \Theta_k$ that $|\tilde{G}_{\theta|\tilde{\theta}_k(\theta)} - \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')}| \leq 4\beta_k$. Hence we have that

$$|g_\theta - g_{\theta'}| = |g_\theta - \tilde{G}_{\theta|\tilde{\theta}_k(\theta)} + \tilde{G}_{\theta|\tilde{\theta}_k(\theta)} - \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} + \tilde{G}_{\theta'|\tilde{\theta}_k(\theta')} - g_{\theta'}| \leq 2\beta_k + 4\beta_k = 6\beta_k,$$

where we again used the event $\mathcal{E}_2$ and the fact that both $\theta, \theta' \in \Theta_k$. □

With the previous two lemmas in hand we are finally ready to prove Theorem 1.

Proof of Theorem 1. We condition the remainder of the proof on the event $\mathcal{E}_2$, which holds with probability at least $1 - 3\delta$. Each epoch k consists of at most $w(N_k + D_\Theta)$ timesteps in expectation: the first wN_k steps correspond to executing the maximal policy set Θ_k^{max} to estimate the performance of all policies in Θ_k,

and the remaining wD_Θ steps account for returning to the fixed initial state s_1, as ensured by Assumption 3. With epochs $k \in [K]$, we use this structure to derive a regret decomposition for the algorithm:

$$\textsf{Regret}(T) = \sum_{k=1}^{K} \sum_{\theta \in \Theta_k^{\max}} N_k (G_\theta(\mathcal{H}_\theta^{N_k}) - g_{\theta_r^*}) + \tilde{O}(KwD_\Theta) + TL_\Theta r.$$

The first term reflects the cost incurred while executing each $\theta \in \Theta_k^{\max}$ within an epoch. The second term accounts for the D_Θ steps required to return to s_1 at the end of each policy's execution, and follows from the fact that the per-step cost is bounded in $[0,1]$. The third term captures the discretization error $|g_{\theta_r^*} - g_{\theta^*}|$, which is bounded by the Lipschitz continuity of the cost function.

Given that $N_k = \frac{2^{2k} T_h(\delta)}{\alpha} \log(T)$ we can bound the number of epochs K by using the fact that:

$$\begin{aligned} T \geq \sum_{k=1}^{K} wN_k &= \sum_{k=1}^{K} w \left(\frac{2^{2k} T_h(\delta)}{\alpha} \log(T) \right) \\ &= \frac{4wT_h(\delta)\log(T)}{3\alpha}(4^K - 1). \end{aligned}$$

Solving this equation for K gives that $K = \tilde{O}(\log_4 \frac{\alpha T}{wT_h(\delta)}) = \tilde{O}(\log(T))$. Note that this also requires $T \geq 4\frac{wT_h(\delta)}{\alpha}$, since K must be greater than 1.

For the first term we use the definition of the event $\mathcal{E}_2$ to have

$$\begin{aligned} \sum_{k=1}^{K} \sum_{\theta \in \Theta_k^{\max}} (N_k (G_\theta(\mathcal{H}_\theta^{N_k}) - g_{\theta_r^*})) &\leq \sum_{k=1}^{K} (wN_k \beta_k + wN_k (g_{\theta_k}(s_1) - g_{\theta^*}(s_1))) \\ &\leq \sum_{k=1}^{K} 7wN_k \beta_k. \end{aligned}$$

The first inequality follows from the event $\mathcal{E}_2$, while the second follows from Lemma A.6, since both θ^* and θ_k belong to Θ_k. With the definition of N_k and β_k we have:

$$\begin{aligned} 7\sum_{k=1}^{K} wN_k \beta_k &= 7w \sum_{k=1}^{K} N_K \left(\frac{H}{\alpha N_k} + (H+2) \sqrt{\frac{2\log(4|\mathcal{N}_r(\Theta)|K/\delta)}{\alpha N_k}} \right) \\ &= \frac{7wKH}{\alpha} + 7w(H+2)\sqrt{\frac{2}{\alpha}\log(4|\mathcal{N}_r(\Theta)|K/\delta)} \sum_{k=1}^{K} \sqrt{N_k} \\ &= \tilde{O}\left(\frac{wH}{\alpha} \log_4 \frac{\alpha T}{wT_h(\delta)} \right) + 7w(H+2)\sqrt{\frac{2}{\alpha}\log(4|\mathcal{N}_r(\Theta)|K/\delta)} \sum_{k=1}^{K} \sqrt{N_k}. \end{aligned}$$

Note that $\sum_{k=1}^{K} \sqrt{N_k} = O\left(\sqrt{\frac{T_h(\delta)}{\alpha}} \log(T) (\frac{\alpha T}{wT_h(\delta)})^{1/2} \right) = O\left(\sqrt{\frac{T}{w}} \log(T) \right)$. Combining all the terms together gives that:

$$\textsf{Regret}(T) \leq O\left(\frac{wH}{\alpha} + wD_\Theta \right) \log_4 \frac{\alpha T}{wT_h(\delta)} + O\left((H+2)\sqrt{\frac{w}{\alpha}\log(4|\mathcal{N}_r(\Theta)|K/\delta)T}\log(T) \right) + TL_\Theta r.$$

By taking $r = (1/T)^{1/2}$, we have $|\mathcal{N}_r(\Theta)| = O((U\sqrt{T})^d)$. □

A.2. Regret Lower Bound (Theorem 2)

We begin this section by proving the lower bound stated in Section 3. We then extend this result to a more challenging setting in which the cost is *not observed* by the algorithm, showing that the lower bound increases by a factor of $1/\alpha$. Notably, our algorithm and regret guarantees apply even under such partial observability. As a result, the regret bound in Theorem 1 continues to be minimax-optimal in α, despite the presence of unobserved costs.

THEOREM 2 **(Regret Lower Bound).** *For any $w, |\Theta|$, and H, with $|\Theta| \geq H$, for any algorithm, there exists an MDP and finite policy class Θ satisfying the sample-path policy order of width w and Assumptions 1 to 3, such that*

$$\mathbb{E}[\mathit{Regret}(T)] = \begin{cases} \Omega(\sqrt{H \log(|\Theta|)T}) & w = 1 \\ \Omega(\sqrt{HwT}) & w > 1. \end{cases}$$

Proof. We organize the proof in several parts. First, we show that for any algorithm, there exists an MDP and a policy class of size $|\Theta|$ with width w, such that its regret is lower-bounded by

$$\mathbb{E}[\mathsf{Regret}(T)] = \begin{cases} \Omega\Big(\sqrt{\log(|\Theta|)T}\Big), & w = 1, \\ \Omega\Big(\sqrt{wT}\Big), & w > 1. \end{cases}$$

Second, we explain how to augment these constructions to include an arbitrary H, yielding the final bound as in Theorem 2. Our lower bound construction is based on reducing the problem of learning in an average-cost MDP to that of learning in a bandit instance. Before presenting the full construction, we start with a discussion on the bandit lower bound instances. While for consistency with the literature, these are written in terms of reward maximization, they can equivalently be done for cost minimization, as the focus of our paper, by multiplying the rewards by negative one.

Part 1: $\Omega(\sqrt{\log(|\Theta|)T})$ Full-Feedback Lower Bound. We begin by considering the full-feedback Θ armed bandit, an instance of our model with no states and $w = 1$. At each time step t, the algorithm selects $\pi_\theta = A_t \in \{1, \ldots, |\Theta|\}$ and observes the entire rewards vector $Y_t = (Y_t(1), \ldots, Y_t(|\Theta|))$, with each $Y_t(a)$ i.i.d. and following $\mathrm{Bernoulli}(\mu_a) \in [0, 1]$. To construct the hard instance, we pick a uniform hidden index $J \in \{1, \ldots, |\Theta|\}$, and under environment P_j where $j \in \{1, \ldots, |\Theta|\}$,

$$Y_t(a) \sim \begin{cases} \mathrm{Bernoulli}\big(\frac{1}{2} + \Delta\big), & a = j, \\ \mathrm{Bernoulli}\big(\frac{1}{2}\big), & a \neq j. \end{cases}$$

Let the hidden index $J \sim \mathsf{Unif}\{1, \ldots, |\Theta|\}$, also let $N_T(a) = \sum_{t=1}^{T} \mathbb{I}(A_t = a)$ and $\widehat{J} = \arg\max_a N_T(a)$. Recall the definition of regret for a bandit instance:

$$R_T = \sum_{t=1}^{T} \mu^* - \mu(A_t),$$

where $\mu^* = \arg\max_a \mu_a$. Then we have that under instance P_j, the regret R_T satisfies:

$$\mathbb{E}[R_T|P_j] = \mathbb{E}[\Delta\big(T - N_T(j)\big)|P_j] \geq \frac{\Delta T}{2}\,\mathbb{P}_j(\widehat{J} \neq J). \tag{EC.3}$$

Taking expectation over J and the learner's randomness gives

$$\mathbb{E}[R_T] \geq \frac{\Delta T}{2}\,\mathbb{P}(\widehat{J} \neq J). \tag{EC.4}$$

Thus with Fano's inequality,

$$\mathbb{P}(\widehat{J} \neq J) \geq 1 - \frac{I(X^T; J) + \log(2)}{\log(|\Theta|)}, \tag{EC.5}$$

where $X^T = (A_1, Y_1, \ldots, A_T, Y_T)$ is the vector of observed arm reward pairs. In any single time step, due to full-feedback, for any $j \neq j'$, the information of environment j and j' are always observable, thus:

$$\begin{aligned}\mathrm{D_{KL}}\big(P_j(Y_t) \parallel P_{j'}(Y_t)\big) &= \mathrm{D_{KL}}\big(\mathrm{Bernoulli}(\tfrac{1}{2}+\Delta) \| \,\mathrm{Bernoulli}(\tfrac{1}{2})\big) + \mathrm{D_{KL}}\big(\mathrm{Bernoulli}(\tfrac{1}{2}) \| \,\mathrm{Bernoulli}(\tfrac{1}{2}+\Delta)\big) \\ &\leq 4\Delta^2.\end{aligned}$$

By independence over t, $\mathrm{D_{KL}}\big(P_j(X^T) \parallel P_{j'}(X^T)\big) \leq 4\Delta^2 T$. Then we have

$$I(X^T; J) \leq 4\Delta^2\, T. \tag{EC.6}$$

Setting $\Delta = \sqrt{\frac{\log(|\Theta|)}{16T}}$ gives in (EC.5) that $\mathbb{P}(\widehat{J} \neq J) \geq \frac{1}{2}$ for sufficiently large $|\Theta|$. Plugging into (EC.4) yields

$$\mathbb{E}[R_T] = \Omega\big(\sqrt{\log(|\Theta|T}\big)$$

as required.

Part 2: $\Omega(\sqrt{wT})$ Bandit Lower Bound. Next we consider constructing a modified bandit instance with partial feedback according to a policy order of width w. Fix any width $w \geq 2$ and any $|\Theta|$, and consider a standard $|\Theta|$-arm stochastic bandit with unknown arm means in $[0, 1]$. For simplicity, we assume $\frac{|\Theta|}{w}$ is an integer. We divide $|\Theta|$ into w disjoint blocks, and assume that all arms in the same block have the same reward distribution.

This problem clearly admits a policy class (parametrized by the set of arms Θ) of width w (since counterfactual estimates are directly obtained for all other arms in the same block). Furthermore, this problem degenerates into a w-arm bandit problem with no side information. Hence, by the classical minimax lower bound for stochastic K-armed bandits (see Chapter 15 of Lattimore and Szepesvári (2020)), we obtain an instance for which, for any policy,

$$\mathbb{E}[R_T] = \Omega\big(\sqrt{wT}\big).$$

Part 3: Extension to General H. Finally, we construct our "hard instance" of an MDP with an arbitrary bound H on the bias, which reduces to learning in either of the bandit instances described above. For simplicity we assume that T/H is an integer.

MDP Definition. We describe an MDP with finite episode length H, i.e. after H timesteps the state transitions back to a fixed starting state s_1 (Puterman 2014). Note that if the rewards are bounded in $[0,1]$, this yields an MDP where the span of the bias is upper bounded by H.

Before formally stating the MDP, we consider a bandit problem described above in either Part 1 or Part 2, referring to a set of the form $\{\mu_a, \forall a \in \Theta\}$. The action space for our MDP is fixed as $\mathcal{A} = \Theta$. The state space for the MDP is:

$$\mathcal{S} = \{s_1\} \cup \{(\Theta, b, h)\}_{b \in \{0,1\}, h \in [H]}.$$

The first index corresponds to a fixed initial state. The second component corresponds to an action, observed reward, as well as the timestep in the episode. The dynamics for the MDP are as follows:

The initial state is a fixed special state s_1. In state s_1, selecting action $a \in \Theta$ yields a cost C from a Bernoulli random variable with mean $\mu(a)$ (as in Part 2), or (ii) a sample from $\mu(a)$ for all a (as in Part 1). The state then transitions to $(a, C, 2)$. From this point forward, the only valid action is a, yielding cost of C and transitioning to state $(a, C, h+1)$. This repeats H steps until the system returns to state s_1.

Part 4: Verifying Assumptions. First note the MDP has costs bounded in $[0,1]$ and periods of length H, so the span of the MDP is upper bounded by H. Moreover, the resulting MDP has a finite state and action space, so both Assumptions 1 and 3 hold trivially. Returning to s_1 every H periods ensures Assumption 2.

Our policy class can be parameterized as Θ corresponding to the fixed arm $a \in \Theta$ that is played in the bandit instance. If the bandit instances are constructed as in Part 1, it is clear that this policy class admits a sample path policy order of width $w = 1$, since there are full observations on the resulting costs. If the bandit instances are constructed as in Part 2, we see the policy class has a sample path policy order of width w corresponding to the size of the blocks, as discussed above.

Part 5: Establishing Lower Bound. With the MDP defined above, we are ready to put together the pieces to establish the lower bound. Note that the optimal policy θ^* satisfies $g_{\theta^*} = \mu^*$, the mean of the optimal action for the underlying bandit instance. Thus we can rewrite regret via:

$$\begin{aligned}
\mathbb{E}[\mathsf{Regret}(T)] &= \sum_t \mathbb{E}[C_t(S_t, A_t)] - \mu^* \\
&= H \sum_{c=1}^{T/H} \mathbb{E}[C_{Hc}(S_{Hc}, A_{Hc})] - \mu^* \\
&= H\mathbb{E}\left[R_{T/H}\right] \\
&\geq H\sqrt{BT/H} = \sqrt{HBT},
\end{aligned}$$

where B is either $\log(|\Theta|)$ or w, depending on whether the construction follows Part 1 or Part 2. □

REMARK EC.1. Theorem 2 shows that the dependence on regret with respect to T and w is minimax optimal. However, the results are not minimax optimal up to $\log(|\Theta|)$, in the same way for bandit and full-feedback settings there is a gap between $\sqrt{K}$ and $\log(K)$ in the lower bounds, where K is the number of

Table EC.2 Parameter specifications for the numerical simulations.

	Small-scale	**Large-scale**
Inventory Control		
Lead Time (L)	2	6
Holding Cost (h)	1	1
Lost-sales Penalty (p)	10	10
$\mathbb{P}(D_t = 0)$ (γ)	0.3	0.3
Demand Support (d)	$[0, 3]$	$[0, 40]$
Distributions	Exp($\lambda = 1$) Normal($1.5, 0.5$) Uniform($[0, 3]$)	Exp($\lambda = \frac{40}{3}$) Normal($20, \frac{20}{3}$) Uniform($[0, 40]$)
Dual Sourcing		
Lead Time (L_r, L_e)	(1,0)	(5,0)
Holding Cost (h)	1	1
Shortage Penalty (p)	10	10
Purchasing Cost (c_r, c_e)	(0, 0.5)	(0, 0.5)
$\mathbb{P}(D_t = 0)$ (γ)	0.3	0.3
Demand Support (d)	$[0, 3]$	$[0, 40]$
Distributions	Exp($\lambda = 1$) Normal($1.5, 0.5$) Uniform($[0, 3]$)	Exp($\lambda = \frac{40}{3}$) Normal($20, \frac{20}{3}$) Uniform($[0, 40]$)
Queuing Model		
Buffer Size (L)	–	2
Maximum Service Rate ($A_{\max}$)	–	3
Arrival Rate ($\lambda, \lambda_{\max}$)	–	$(6, 10)$
Service Rate ($\mu, \mu_{\max}$)	–	$(3, 10)$
Power Cost Function ($w(a)$)	–	$w(a) = a^2$
Deadline Missing Penalty (C)	–	100

arms. Despite recent progress that closes the $\log(K)$ gap in the adversarial bandit setting (Eldowa et al. 2023, Chen et al. 2024), as far as we know, closing this gap in the stochastic setting remains open.

Our bound is also not minimax optimal up to $\sqrt{H}$. Existing lower bounds on average-cost MDPs exhibit similar dependence on $\sqrt{H}$ instead of the upper bounds of H (see, for instance Auer et al. (2008)). While Bartlett and Tewari (2012) claims to achieve a lower bound which is linear in H, Osband and Van Roy (2016) suggests that there are some mistakes in their construction. As such, this $\sqrt{H}$ discrepancy in the upper and lower bounds is common in the literature.

For a detailed discussion of α in the lower bound, please refer to the full version of the paper.

Appendix B: Numerical Simulations

B.1. Simulation Information

Computing Infrastructure. The experiments were conducted on a server with an Intel i7-14700K 20 Core Processor and 32GB of RAM. No GPUs were needed for the experiments. Each simulation (evaluating all algorithms included in the figures) took approximately 60 hours.

Experiment Setup. Each experiment was repeated 20 times, and all plots and metrics report averages over these runs. Each policy was evaluated on trajectories of length 10^5. Hyperparameters for all algorithms were set as follows: IOPEA uses only the constants specified in the main paper (e.g. H, α, r) with no additional tuning; PPO's constant learning rate (chosen from $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$), batch size ($[16, 32, 64, 128]$), and discount factor ($[0.9, 0.99, 0.999, 0.9999]$) were selected via grid search. SGD's constant learning rate (ranging between 10^{-2} and 4×10^4), sample size (chosen from $[10^0, 10^1, 10^2, 10^3, 10^4]$), and the gap width (ranging between 10^{-2} and 10^0) used in the finite-difference estimator were also tuned via grid search. Feedback Graph's (Dann et al. 2020) exploration bonus scale was selected via grid search over $[10^{-4}, 10^4]$. The convexity-based algorithm of Agrawal and Jia (2022) and BASA (Chen et al. 2024) involve no tuning. UCRL2's (Anselmi et al. 2022) exploration bonuses (fixed constants 7 and 14) are taken directly from the original paper, while its confidence parameter δ was tuned by grid search over $[0.01, 0.05, 0.1]$.

To identify the optimal policy θ^* in each setting, we discretized the policy class with a radius $r = 0.1$ (for the algorithms that need discretization, expect IOPEA, which is determined by $T^{-1/2}$).

In the inventory and dual-sourcing simulations, we train the algorithms on pseudocost $\tilde{C}$ but report performance on C (as discussed in Section 4.1 and 4.2). Additionally, for the dual-sourcing simulations, we assume $c_r = 0$ and $L_e = 0$, i.e., the purchasing cost in the regular channel is omitted, and we assume the lead time of the expedited channel is 0. This is without loss of generality since any dual sourcing problem can be transformed to this type (Chen and Shi 2025).

In Table 1 we report the performance of the M/M/1/L queueing model under exponential distributions, and leave the other two blocks empty. All other parameter specifications for the numerical simulations are in Table EC.2.

Table EC.3 Queuing Performance under Decaying vs. Fixed Arrival Rates

Algorithm	Queuing	
	Decaying Arrival Rate	Fixed Arrival Rate
Optimal (g_{θ^*})	9.5	11.6
ERM (Sinclair et al. 2023b)	9.6(1%)	11.6(0%)
IOPEA	*10.5(11%)	*13.1(13%)
PPO (Schulman et al. 2017)	*9.8(3%)	*11.8(2%)
Feedback Graph (Dann et al. 2020)	11.5(21%)	13.2(13%)
Problem-Specific	11.3(19%)	13.9(20%)
Random	11.1(17%)	13.5(16%)

B.2. Numerical Simulations for Queuing Model with Fixed Arrival Rate

In the queuing case study of Sections 4.3 and 5, we adopted the decaying arrival-rate setting from Anselmi et al. (2022) for both theoretical analysis and numerical experiments. In this appendix, we instead examine the classic M/M/1/L model with a fixed (constant) arrival rate, λ, and present the corresponding simulation results for various algorithms, including IOPEA. These results are summarized in Table EC.3. Aside from replacing the time-varying rates $\{\lambda_i\}$ of Section 4.3 with the constant rate λ (see Table EC.2), the experimental setup is identical to that of Section 5. We observe that nearly all methods exhibit performance comparable to the decaying-rate case, and thus the main conclusions of Section 5 remain unchanged under the fixed-rate setting.

We also observe that in both the decaying-rate and fixed-rate scenarios, the Problem-Specific algorithm (UCRL2) under performs the Random baseline. This is because the problem is relatively small (so Random performs reasonably) and the UCRL2 algorithm after $T = 3 \times 10^5$ is not long enough for the algorithm to approximate the optimal policy. We limit our experiments to this scale because UCRL2 requires comparing multiple MDP models at each epoch, which becomes prohibitively expensive in larger scale problems.

Due to space constraints, the remaining details are provided in the full version of the paper, available online.

Appendix C: Omitted Discussions in Section A.2

The lower-bound discussion in Section A.2 ignores dependence on α: for any finite policy class Θ, the trivial order has width $w = |\Theta|$, so one always obtains $\tilde{O}(\sqrt{|\Theta|T})$ regret. This precludes a general $\tilde{\Omega}\big(\sqrt{wT/\alpha}\big)$ lower-bound, since as $\alpha \to 0$ the naive $\tilde{O}(\sqrt{|\Theta|T})$ upper-bound outperforms it. Our notion of distributional policy order, as well as IOPEA, does not assume that $C_t(S_t, A_t)$ is observed at every step. To capture the α dependence, we therefore analyze a family of MDPs where costs are revealed only on an α-fraction of time steps. We show by explicit construction that with access to exactly αT cost observations, any algorithm incurs $\tilde{\Omega}(\sqrt{wT/\alpha})$ regret. Thus our results are tight in α as well.

THEOREM C.1. *For any $w, |\Theta|, \alpha$, and H, with $|\Theta| \geq H$, for any algorithm, there exists an MDP and finite policy class Θ satisfying the distributional policy order of width w and Assumptions 1 to 3, such that*

$$\mathbb{E}[\textit{Regret}(T)] = \begin{cases} \Omega(\sqrt{\frac{H \log(|\Theta|)T}{\alpha}}) & w = 1 \\ \Omega(\sqrt{\frac{HwT}{\alpha}}) & w > 1. \end{cases} \tag{EC.7}$$

Proof. Similar to the proof of Theorem 2, we organize the proof in several parts. For simplicity, we assume that $1/\alpha$ is an integer.

Part 1: Extension to General α via Reward Censoring. We begin by modifying the MDP from Part 3 of the proof of Theorem 2 as follows. First, whenever the chain visits s_1 and selects an action a, the algorithm observes the cost C which is incurred for the next H timesteps (as in Part 3). However, for the next $H(\frac{1}{\alpha} - 1)$ timesteps, the algorithm incurs additional costs according to their selected action a that are *unobserved*.

MDP Definition. We again consider a bandit problem referring to a set of the form $\{\mu_a, \forall a \in \Theta\}$. The action space for our MDP is fixed as $\mathcal{A} = \Theta$. The state space for the MDP is:

$$\mathcal{S} = \{s_1\} \cup \{(\Theta, b, h, m)\}_{b \in \{0,1\}, h \in [H], m \in [1/\alpha]}.$$

The first index corresponds to a fixed initial state. The second component corresponds to an action, observed reward, as well as a step in the episode h and block m pair. We use h to repeat the rewards H times, and m to denote blocks. The first block ($m = 1$) the cost is observed, but for the remaining $1/\alpha - 1$ blocks, the costs are unobserved.

The dynamics for the MDP are as follows: The initial state is a fixed initial state s_1. In state s_1, selecting action $a \in \Theta$ yields a cost C from a Bernoulli random variable with mean $\mu(a)$ (or a sample from $\mu(a)$ for all $a \in \Theta$). The state then transitions to $(a, C, 2, 1)$, corresponding to playing action a, observing cost C, step $h = 2$ within the episode, and in block $m = 1$. From this point forward, the only valid action is a. In state (a, C, h, m), if the current step $h < H$, the MDP transitions to $(a, C, h + 1, m)$, incurring the same cost C. However, if $h = H$ then the MDP transitions to $(a, C, 1, m + 1)$ where C now denotes a *new* but *unobserved* sample from $\mu(a)$.

Part 2: Verification of Policy Order. To distinguish $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ and $G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha T})$ carefully, we argue that $\mathcal{H}_\theta^T$ in $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ and $\mathcal{H}_\theta^T$ in $G_{\theta'}$ have different meanings. In $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$, the trajectory $\mathcal{H}_\theta^T$ refers to the measurability over observations, but in $G_{\theta'}$, it is used for the empirical estimation of the true gain function g_θ. Thus in our bandit cases above, although we collect samples from a length T trajectory, to empirically estimate the gain function g_θ, the length of "usable" trajectory is only of length αT. Thus, our constructed bandit examples above satisfy the distributional policy order with factor α, due to reward censoring, by letting $\tilde{G}_{\theta'|\theta}$ be the empirical estimate of g_θ only over the first block of the round at each episode.

Part 3: Verifying Assumptions. The MDP has a finite state and action space, and so Assumption 1 and the Lipschitz bound hold trivially. Returning to s_1 every H/α timesteps ensures Assumption 2. Since the costs are bounded within $[0,1]$, the span of the bias is also H/α. However, the only place we use the bound on the bias in the regret guarantee is in Lemma A.1. It is clear by the construction of our instance that one can still obtain $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ satisfying Lemma A.1 on the order of H/α by using the observed samples in the first episode of each block directly.

Part 4: Establishing Lower Bound. With the MDP defined above, we are ready to put together the pieces to establish the lower bound. Note again that the optimal policy θ^* satisfies $g_{\theta^*} = \mu^*$, the mean of the optimal action for the underlying bandit instance. Thus, we can rewrite regret via:

$$\begin{aligned}
\mathbb{E}[\mathsf{Regret}(T)] &= \sum_{t=1}^{T} \mathbb{E}[C_t(S_t, A_t)] - \mu^* \\
&= \frac{H}{\alpha} \sum_{c=1}^{\frac{T\alpha}{H}} \mathbb{E}\left[C_{\frac{Hc}{\alpha}}(S_{\frac{Hc}{\alpha}}, A_{\frac{Hc}{\alpha}})\right] - \mu^* \\
&= \frac{H}{\alpha} \mathbb{E}\left[R_{\frac{\alpha T}{H}}\right] \\
&\geq \frac{H}{\alpha} \sqrt{B \frac{\alpha T}{H}} = \sqrt{\frac{HBT}{\alpha}}.
\end{aligned}$$

Here, we implicitly use the fact that an algorithm only gets information at the start of each episode within a block. Hence, by rewriting the regret as the sum over those timesteps, we see that the regret for the algorithm can be represented as the regret of the *bandit* algorithm over the $\alpha T/H$ timesteps corresponding to the start of each block for $m = 1$, $h = 1$. While we omit the full proof of this lower bound (it follows similarly to Theorem 2, it leverages the fact that the mutual information does not increase for the *censored* samples. Again we use B to denote either $\log(|\Theta|)$ or w, depending on whether the construction follows Part 1 or Part 2. Using the appropriate full-feedback or partial-feedback bandit instance yields the result. □

REMARK EC.2. Theorem C.1 establishes that our regret in Theorem 1 is minimax optimal with respect to α. This is because IOPEA (and our result in Theorem 1) allows for reward censoring. Indeed, our algorithm only relies on counterfactual estimates up to a factor of α for every policy in Θ_k for epoch k, including the policy that is actually implemented. As a result, the algorithm remains valid even if rewards for all policies are partially censored.

Appendix D: Proofs in Section 4

D.1. Single-Retailer Inventory Control with Positive Lead Time (Section 4.1)

LEMMA 1. *The set of base stock policies satisfies the sample-path policy order of width one, where* $\pi_{\theta'} \preceq \pi_\theta$ *whenever* $\theta' \leq \theta$.

Proof of Lemma 1. We show the proof by establishing that $\mathcal{H}^T_{\theta'} \in \mathcal{H}^T_\theta$. To do so, we start off with the following lemma.

LEMMA D.1. *Suppose that* $\{S_t\}_{t=1}^T$ *is a state trajectory collected under a base stock policy* π_θ *and* $\{S'_t\}_{t=1}^T$ *collected under a base stock policy* $\pi_{\theta'}$ *over a fixed sequence of demands* $(D_1, \ldots, D_T)$ *with* $\theta \geq \theta'$. *Then if*

- $S_0 \geq S'_0$
- $\theta - \|S_0\|_1 \geq \theta' - \|S'_0\|_1 \geq 0$,

we have that for all $t \leq T$*:*

- $S_t \geq S'_t$
- $\theta - \|S_t\|_1 \geq \theta' - \|S'_t\|_1 \geq 0$.

Proof. We show the result by induction over the timestep t. For the base case all of the statements are true by assumption, so we only have to focus on the step case for t implying $t+1$. We abbreviate IH as the induction hypothesis.

Recall that for any $t \leq T$, the state $S_t = (I_t, Q_{t-L}, \ldots, Q_{t-1}) \in \mathbb{R}^{L+1}$, refers to on-hand inventory and L unfinished orders in the pipeline. For clarity, we denote the components as $(S_{t,0}, \ldots, S_{t,L}) \triangleq (I_t, Q_{t-L}, \ldots, Q_{t-1})$, and similariy, $S'_t = (S'_{t,0}, \ldots, S'_{t,L}) = (I'_t, Q_{t'-L}, \ldots, Q_{t'-1})$. Then, under the system dynamics given in Equation (7), and assuming the demands $D_1, \ldots, D_T$ are fixed, we have for all $t < T$:

$$S_{t+1} = ((S_{t,0} + S_{t,1} - D_t)^+, S_{t,2}, \ldots, S_{t,L}, (\theta - \|S_t\|_1)^+)$$
$$S'_{t+1} = ((S'_{t,0} + S'_{t,1} - D_t)^+, S'_{t,2}, \ldots, S'_{t,L}, (\theta' - \|S'_t\|_1)^+).$$

First we show that $S_{t+1} \geq S'_{t+1}$. For any index $j = 1, \ldots, L-1$ we have that

$$S_{t+1,j} = S_{t,j+1} \geq S'_{t,j+1} = S'_{t+1,j},$$

where the inequality in the middle uses the IH. For the index $j = L$ we have that:

$$S_{t+1,L} = (\theta - \|S^t\|_1) \geq (\theta' - \|S'_t\|_1) = S'_{t+1,L},$$

where again the last line uses the IH. For the first index $j = 0$ we have:

$$S_{t+1,0} = (S_{t,0} + S_{t,1} - D_t)^+ \geq (S'_{t,0} + S'_{t,1} - D_t)^+ = S'_{t+1,0}.$$

Lastly, we show that

$$\theta - \|S_{t+1}\|_1 \geq \theta' - \|S'_{t+1}\|_1 \geq 0.$$

The fact that $\|S'_{t+1}\|_1 \leq \theta'$ is clear by the induction hypothesis and the definition of the dynamics. We further note that

$$\theta - \|S_{t+1}\| = (S_{t,0} + S_{t,1} - (S_{t,0} + S_{t,1} - D_t)^+), \quad \theta' - \|S'_{t+1}\| = (S'_{t,0} + S'_{t,1} - (S'_{t,0} + S'_{t,1} - D_t)^+).$$

Now we compare the expressions under different realizations of D_t:

1. If $D_t \leq S'_{t,0} + S'_{t,1}$, then since $S'_{t,0} + S'_{t,1} \leq S_{t,0} + S_{t,1}$, we also have $D_t \leq S_{t,0} + S_{t,1}$. Thus:

$$(S_{t,0} + S_{t,1} - (S_{t,0} + S_{t,1} - D_t)^+) = D_t = (S'_{t,0} + S'_{t,1} - (S'_{t,0} + S'_{t,1} - D_t)^+).$$

2. If $D_t \geq S_{t,0} + S_{t,1}$, then since $S'_{t,0} + S'_{t,1} \leq S_{t,0} + S_{t,1}$, we have $D_t \geq S'_{t,0} + S'_{t,1}$. Thus:

$$(S_{t,0} + S_{t,1} - (S_{t,0} + S_{t,1} - D_t)^+) = S_{t,0} + S_{t,1} \geq S'_{t,0} + S'_{t,1} = (S'_{t,0} + S'_{t,1} - (S'_{t,0} + S'_{t,1} - D_t)^+).$$

3. If $S'_{t,0} + S'_{t,1} < D_t < S_{t,0} + S_{t,1}$, then:

$$(S_{t,0} + S_{t,1} - (S_{t,0} + S_{t,1} - D_t)^+) = D_t > S'_{t,0} + S'_{t,1} = (S'_{t,0} + S'_{t,1} - (S'_{t,0} + S'_{t,1} - D_t)^+).$$

Combining all of the different cases we see that $\theta - \|S_{t+1}\|_1 \geq \theta' - \|S'_{t+1}\|_1$ as needed. □

Now, in order to show that the base stock policies satisfy the information order, we recall that the starting state $s_1 = (0, \ldots, 0)$. Hence, by Lemma D.1 for any two base stock levels $\theta' \leq \theta$ we see that the on-hand inventory under base stock policy π_θ $(S_{t,0})$ is always larger than that under $\pi_{\theta'}$ $(S'_{t,0})$. Thus, we just need to show that $\tilde{C}(S'_t)$ is measurable with respect to $\mathcal{H}_\theta^T$. We show by induction on t that: $S'_t \in \mathcal{H}_\theta^T$ and $\tilde{C}(S'_t) \in \mathcal{H}_\theta^T$. Indeed, this is true for $t = 0$. For $t > 0$:

$$S'_{t+1} = ((S'_{t,0} + S'_{t,1} - D_t)^+, S'_{t,2}, \ldots, (\theta' - \|S'_t\|)^+).$$

However, by the induction step, we know that S'_t is measurable with respect to $\mathcal{H}_\theta^T$. Moreover, since $S'_t \leq S_t$ via Lemma D.1 we know if $S_{t+1,0} = (S_{t,0} + S_{t,1} - D_t)^+ = S_{t,0} + S_{t,1}$, then $S'_{t+1,0}$ is trivially in $\mathcal{H}_\theta^T$. Otherwise, if $S_{t+1,0} = 0$ then $D_t \geq S_{t,0} + S_{t,1} \geq S'_{t,0} + S'_{t,1}$ and so $S'_{t+1,0} = 0$ as well (implying $S'_{t+1,0} \in \mathcal{H}_\theta^T$). Overall, we see that $S'_t \in \mathcal{H}_\theta^T$ as required.

Similarly for the costs,

$$\tilde{C}(S'_t) = h(S'_{t,0} + S'_{t,1} - \min\{S'_{t,0} + S'_{t,1}, D_t\}) - p \min\{S'_{t,0} + S'_{t,1}, D_t\}).$$

However, $\min\{S'_{t,0} + S'_{t,1}, D_t\}) \in \mathcal{H}_\theta^T$ since $S'_t \leq S_t$ via Lemma D.1. Thus we have that $\tilde{C}(S'_t) \in \mathcal{H}_\theta^T$. All together, this implies that $\pi_{\theta'} \preceq \pi_\theta$ as needed. □

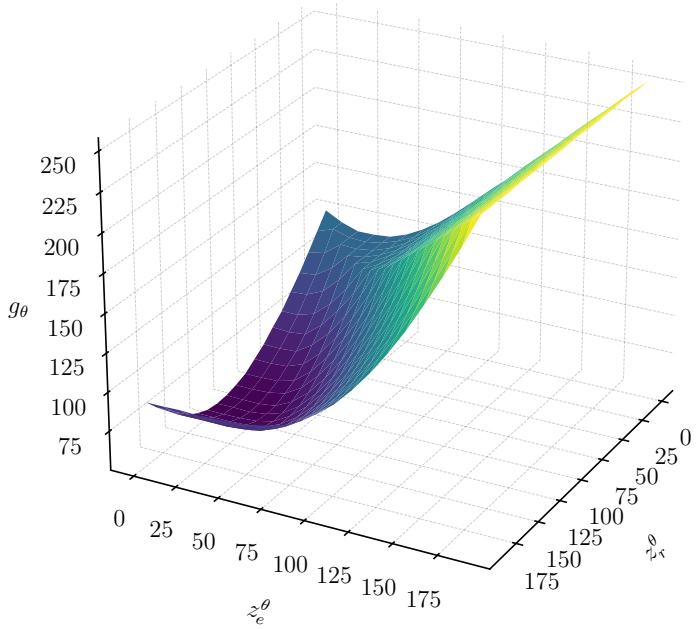

Figure EC.1 **An empirical illustration of non-convexity in the long-run average cost with respect to the base-stock levels (z_e^θ, z_r^θ) under the lost-sales dual index policy. Parameters are $L_r = 8$, $L_e = 2$, $c_r = 1$, $c_e = 8$, $h = 1$, $p = 10$, and demand follows an exponential distribution with rate $3/40$. Costs are approximated by sample averages over $T = 10^5$ steps, evaluated on a grid $(z_e^\theta, z_r^\theta) \in \{0, 5, 10, \ldots, 200\}^2$.**

D.2. Dual Index Policy for the Dual Sourcing Problem (Section 4.2)

D.2.1. Non-Convexity of the Problem Unlike the convex cost curve of the lost-sales single-channel base-stock policy in Agrawal and Jia (2022), and the empirically observed non-convexity of the backlog dual index policy in Veeraraghavan and Scheller-Wolf (2008), the convexity of the long-run average cost under the lost-sales dual index policy (our setting in Section 4.2) has received little attention in the literature. In Fig. EC.1, we follow Veeraraghavan and Scheller-Wolf (2008) to show empirically that this cost is non-convex in the base-stock parameters. Notably, when the expedited base-stock level $z_e^\theta = 0$, the dual index policy reduces to the single-channel base-stock policy and recovers the convex cost curve of Agrawal and Jia (2022) (see the $z_e^\theta = 0$ trace in Fig. EC.1). Figure EC.1 further suggests Lipschitz continuity for general cases beyond the "slow-moving" assumption in Lemma 4 (note that the exponential demand example in Fig. EC.1 does not satisfy that assumption).

D.2.2. Verification of Assumptions Before formally describing the information order over policies, we start by showing two required results to apply Theorem 1. First, we provide an upper bound on the bias in terms of $\gamma = \mathbb{P}(D_t = 0)$ and the long lead time L_r. Second, we provide conditions for which the cost function g_θ is Lipschitz.

Following Proposition 8.1.1 of Puterman (2014) and Remark 3 of Agrawal and Jia (2022), we will ignore discussion around the existence of any limits. Indeed, Agrawal and Jia (2022) establishes that employing an appropriate discretization of state space and action space allows the existence of all limits within our focus. Before presenting these results we start with a technical lemma.

LEMMA D.2. *Define Y_1 as the first time having L_r consecutive 0 demands, i.e.,*

$$Y_1 = \inf\{t \geq L_r \,|\, \forall k \in \{0, 1, \ldots, L_r - 1\}, D_{t-k} = 0\},$$

and $\gamma = \mathbb{P}(D_t = 0)$. Then

$$\mathbb{E}[Y_1] = \frac{1-\gamma^{L_r}}{(1-\gamma)\gamma^{L_r}} \leq \frac{1}{(1-\gamma)\gamma^{L_r}}.$$

One version of proof of Lemma D.2 based on induction is given in Ross (2014).

LEMMA 3. *For the dual index policies in Section 4.2, the gain is uniform, and the span of the bias is upper bounded by $\frac{1}{(1-\gamma)\gamma^{L_r}}$.*

Proof. Fix any policy $\theta \in \Theta$, and let s_1, s_2 be two arbitrary initial states. For any time horizon T, let $G_\theta(\mathcal{H}_\theta^T)(s)$ denote the empirical average cost from initial state s as defined in Definition 1. Then:

$$\left|TG_\theta(\mathcal{H}_\theta^T)(s_1) - TG_\theta(\mathcal{H}_\theta^T)(s_2)\right| \leq 2\min\{T, Y_1\} \leq 2Y_1,$$

where Y_1 is the first time L_r consecutive 0-demand events occur. Since the costs are bounded by 1, the upper bound of T follows immediately. However, if $Y_1 \leq T$, we note that for all $t' \geq Y_1$ the two systems will be in the same state. Hence, the different in costs is similarly upper bounded by Y_1. Taking expectations and applying Lemma D.2, we get:

$$\left|\mathbb{E}\left[TG_\theta(\mathcal{H}_\theta^T)(s_1) - TG_\theta(\mathcal{H}_\theta^T)(s_2)\right]\right| \leq 2\mathbb{E}[Y_1] \leq \frac{2}{(1-\gamma)\gamma^{L_r}}.$$

It follows that the long-run average cost is uniform:

$$|g_\theta(s_1) - g_\theta(s_2)| = \left|\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[TG_\theta(\mathcal{H}_\theta^T)(s_1) - TG_\theta(\mathcal{H}_\theta^T)(s_2)\right]\right| \leq \lim_{T\to\infty} \frac{2}{(1-\gamma)\gamma^{L_r}T} = 0.$$

Additionally, the span of the bias is bounded:

$$|v_\theta(s_1) - v_\theta(s_2)| = \left|\lim_{T\to\infty}\left[TG_\theta(\mathcal{H}_\theta^T)(s_1) - Tg_\theta - TG_\theta(\mathcal{H}_\theta^T)(s_2) + Tg_\theta\right]\right| \tag{EC.8}$$

$$\leq \frac{2}{(1-\gamma)\gamma^{L_r}}. \tag{EC.9}$$

All limits exist, and the interchange of limits and expectations is justified by the uniform integrability. For simplicity, we defer the proof of uniform integrability in Lemma 4. □

The exponential dependence on L_r for the span H matches some existing reinforcement learning literature (Anselmi et al. 2022), but it remains open whether one can obtain $H = O(L_r)$ in the dual-sourcing setting. While Agrawal and Jia (2022) shows $H = O(L_r)$ for the single-channel case, no such linear bound is known here. Although our policy-order proof incurs $H = O((\gamma^{L_r})^{-1})$, empirical evidence suggests this exponential factor may be avoidable with a sharper argument. We also leave this refinement to future work.

LEMMA 4. *For any dual sourcing system, there exists $c_\gamma \in (0,1)$ such that if $\gamma > c_\gamma$, the cost function g_θ is Lipschitz continuous with respect to θ.*

Proof. Recall that the policy class is characterized by the tuple (z_e^θ, z_r^θ) where z_e^θ and z_r^θ refer to the base-stock level of the short-lead channel and the base-stock level of the long-lead channel, respectively. Hence, the policy class is isomorphic to $\Theta \subset \mathbb{R}^2$.

Now we show that g_θ is Lipschitz continuous. Specifically, for any $\theta_1 = (z_e^{\theta_1}, z_r^{\theta_1})$ and $\theta_2 = (z_e^{\theta_2}, z_r^{\theta_2})$ such that $\|\theta_1 - \theta_2\|_\infty = \delta$, we intend to show that $|g_{\theta_1} - g_{\theta_2}| \le L_\Theta \delta$ for some L_Θ.

Recall that we denote the state at time t as

$$S_t = [I_t, (Q^r_{t-L_r}, Q^r_{t-L_r+1}, \dots, Q^r_{t-1}), (Q^e_{t-L_e}, Q^e_{t-L_e+1}, \dots, Q^e_{t-1})],$$

where I_t refers to the on-hand inventory at time step t, Q^e_t refers to the inventory in the short-lead pipeline that was ordered at time t, and Q^r_t refers to the inventory in the long-lead pipeline that was ordered at time t. Since g_θ is uniform for any θ (Lemma 3), we assume that for policy θ_1, the initial state is $[z_r^{\theta_1}, (0,0,\dots,0), (0,0,\dots,0)]$, and the initial state for policy θ_2 is $[z_r^{\theta_2}, (0,0,\dots,0), (0,0,\dots,0)]$.

Now denote C_t^θ as the observed cost at time t under policy θ. We introduce a sequence of random variables $(Y_i)_{i\in\mathbb{N}_+}$, where Y_i refers to the interarrival time between $(i-1)$-st occurrence of L_r consecutive periods where demands are all 0 and the i-th occurrence of L_r consecutive periods where demands are all 0. Formally, the random variables are defined recursively as

$$Y_i = \inf\{t \ge Y_{i-1} + L_r \mid \forall k \in \{0,1,\dots,L_r-1\}, D_{t-k} = 0\}\} - Y_{i-1},$$

where we set $Y_0 = 1$.

Note that following our definition, $\{Y_i, i \in \mathbb{N}^+\}$ is a sequence of non-negative and i.i.d. random variables. Following convention, if denoting

$$N(T) = \sup\left\{ n \mid \sum_{i=1}^{n} Y_i \le T \right\}, \tag{EC.10}$$

then $N(t)$ is a renewal process. By our construction of Y_i, we have $\mathbb{E}[Y_i] < \infty$, and $\mathbb{E}[\sum_{t\in Y_1} \left| C_t^{\theta_1} - C_t^{\theta_2} \right|] \le 2\mathbb{E}[Y_1] < \infty$. Here we slightly abuse notation by interpreting $\{t \in \mathbb{Z} \mid t \in Y_i, i \le N(T)\}$ as $\{t \in \mathbb{Z} \mid Y_{i-1} \le t \le Y_i, 1 \le i \le N(T)\}$, where we assume $Y_0 = 1$.

Furthermore, $\mathbb{E}[(\sum_{t=1}^{T}(C_t^{\theta_1}-C_t^{\theta_2}))^2]\leq 4T^2$. Using Markov's inequality we get for all $x>0$:

$$\mathbb{P}\left(\frac{\sum_{t=1}^{T}\left|C_t^{\theta_1}-C_t^{\theta_2}\right|}{T}>x\right)\leq\frac{4}{x^2}.$$

Using this uniform integrability, we are free to switch the limit and the expectation. Thus:

$$|g_{\theta_1}-g_{\theta_2}|=\left|\mathbb{E}[\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\sum_{t\in Y_i}(C_t^{\theta_1}-C_t^{\theta_2})]+\mathbb{E}[\lim_{T\to\infty}\frac{1}{T}\sum_{t=N(T)+1}^{T}(C_t^{\theta_1}-C_t^{\theta_2})]\right| \tag{EC.11}$$

$$\leq\mathbb{E}[\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\sum_{t\in Y_i}\left|C_t^{\theta_1}-C_t^{\theta_2}\right|] \tag{EC.12}$$

$$=\frac{\mathbb{E}[\sum_{t\in Y_1}\left|C_t^{\theta_1}-C_t^{\theta_2}\right|]}{\mathbb{E}[Y_1]}. \tag{EC.13}$$

Thus, to provide a bound on $|g_{\theta_1}-g_{\theta_2}|$ it suffices to bound $\mathbb{E}[\sum_{t\in Y_i}|\tilde{C}_t^{\theta_1}-\tilde{C}_t^{\theta_2}|]$, which we do by induction.

Before providing a bound, we give several iterations below to illustrate our regenerative cycle construction. We denote $I_t(\theta_i)$ as the on-hand inventory at t for policy π_i, (and $Q_t^e(\theta_i)$ and $Q_t^r(\theta_i)$) accordingly), and $S_t(\theta_i)$ as the state at time step t for policy θ_i. We further denote D_t as the realized demand at time t. Then we have:

$$S_1(\theta_1)=[z_r^{\theta_1},(0,0,\ldots,0),(0,0,\ldots,0)]$$
$$S_1(\theta_2)=[z_r^{\theta_2},(0,0,\ldots,0),(0,0,\ldots,0)],$$

by our construction described earlier. For $t=2$:

$$S_2(\theta_1)=[[z_r^{\theta_1}-D_1]^+,(0,0,\ldots,0),(0,0,\ldots,0)]$$
$$S_2(\theta_2)=[[z_r^{\theta_2}-D_1]^+,(0,0,\ldots,0),(0,0,\ldots,0)].$$

Next at $t=3$ we have:

$$S_3(\theta_1)=[I_3(\theta_1),(0,0,\ldots,Q_2^e(\theta_1)),(0,0,\ldots,Q_2^r(\theta_1))]$$
$$S_3(\theta_2)=[I_3(\theta_2),(0,0,\ldots,Q_2^e(\theta_2)),(0,0,\ldots,Q_2^r(\theta_2))].$$

In order to finish the proof, we establish the following:

LEMMA D.3. *For any $t\leq Y_1$, we have that*

$$|I_t(\theta_1)-I_t(\theta_2)|+\sum_{i=1}^{L_e}|Q_{t-i}^e(\theta_1)-Q_{t-i}^e(\theta_2)|+\sum_{i=1}^{L_r}|Q_{t-i}^r(\theta_1)-Q_{t-i}^r(\theta_2)|\leq W(t)\delta,$$

where $W(1)=1$ and $W(t)=4W(t-1)+3$. Correspondingly,

- $|I_t(\theta_1) - I_t(\theta_2)| \leq W(t)\delta$
- $|C_t^{\theta_1} - C_t^{\theta_2}| \leq (h+p+c_r+c_e)W(t)\delta$.

Proof. We show the proof by induction on t. For the base case when $t=1$ we have that $I_1(\theta_1) = z_r^{\theta_1}$ and $I_1(\theta_2) = z_r^{\theta_2}$. Thus, due to the fact that $|z_r^{\theta_1} - z_r^{\theta_2}| \leq \delta$, $|I_1(\theta_1) - I_2(\theta_2)| \leq \delta$. Next we note that $C_1^{\theta_1} \leq (h+p)I_1(\theta_1)$, so the bound on $|C_1^{\theta_1} - C_1^{\theta_2}|$ follows immediately.

Next we show that $t \to t+1$. By the induction hypothesis we assume that

$$|I_t(\theta_1) - I_t(\theta_2)| + \sum_{i=1}^{L_e} |Q_{t-i}^e(\theta_1) - Q_{t-i}^e(\theta_2)| + \sum_{i=1}^{L_r} |Q_{t-i}^r(\theta_1) - Q_{t-i}^r(\theta_2)| \leq W(t)\delta. \tag{EC.14}$$

Denote by

$$IP_t^e(\theta_i) = I_t(\theta_i) + \sum_{i=1}^{L_e} (Q_{t-i}^e(\theta_i) + Q_{t-L_r+L_e-i}^r(\theta_i)) + Q_{t-L_r+L_e}^r(\theta_i))$$
$$IP_t^r(\theta_i) = I_t(\theta_i) + \sum_{i=1}^{L_e} Q_{t-i}^e(\theta_i) + \sum_{i=1}^{L_r} Q_{t-i}^r(\theta_i) + Q_t^e(\theta_i)$$

as the total inventory position for the short-lead and long-lead channel respectively at timestep t. Note that by definition of the dual index base stock policies, $Q_t^e(\theta_i) = (z_e^{\theta_i} - IP_t^e(\theta_i))^+$ (and similarly with $Q_t^r(\theta_i)$). Then we have

$$\begin{aligned}
\left|IP_t^e(\theta_1) - IP_t^e(\theta_2)\right| &\leq \left|I_t(\theta_1) - I_t(\theta_2)\right| + \sum_{i=1}^{L_e} \left|Q_{t-i}^e(\theta_1) - Q_{t-i}^e(\theta_2)\right| \\
&\quad + \sum_{i=1}^{L_e} \left|Q_{t-L_r+L_e-i}^r(\theta_1) - Q_{t-L_r+L_e-i}^r(\theta_2)\right| \\
&\quad + \left|Q_{t-L_r+L_e}^r(\theta_1) - Q_{t-L_r+L_e}^r(\theta_2)\right| \\
&\leq W(t)\,\delta.
\end{aligned}$$

Thus due to $|z_e^{\theta_1} - z_e^{\theta_2}| \leq \delta$, we have that

$$|Q_t^e(\theta_1) - Q_t^e(\theta_2)| \leq |(z_e^{\theta_1} - IP_t^e(\theta_1)) - (z_e^{\theta_2} - IP_t^e(\theta_2))| \leq (W(t)+1)\delta, \tag{EC.15}$$

and

$$|Q_t^r(\theta_1) - Q_t^r(\theta_2)| \leq |(z_r^{\theta_1} - IP_t^r(\theta_1)) - (z_r^{\theta_2} - IP_t^r(\theta_2))| \tag{EC.16}$$
$$\leq W(t)\delta + |Q_t^e(\theta_1) - Q_t^e(\theta_2)| + |z_r^{\theta_1} - z_r^{\theta_2}| \tag{EC.17}$$
$$\leq (2W(t)+2)\delta. \tag{EC.18}$$

Note that

$$\begin{aligned}
|I_{t+1}(\theta_1) - I_{t+1}(\theta_2)| &= |[I_t(\theta_1) + Q_{t-L_e}^e(\theta_1) + Q_{t-L_r}^r(\theta_1) - D_t]^+ - [I_t(\theta_2) + Q_{t-L_e}^e(\theta_2) + Q_{t-L_r}^r(\theta_2) - D_t]^+| \\
&\leq |I_t(\theta_1) - I_t(\theta_2)| + |Q_{t-L_e}^e(\theta_1) - Q_{t-L_e}^e(\theta_2)| + |Q_{t-L_r}^r(\theta_1) - Q_{t-L_r}^r(\theta_2)|.
\end{aligned}$$

Correspondingly,

$$\begin{aligned}|I_{t+1}(\theta_1)-I_{t+1}(\theta_2)|+\sum_{i=1}^{L_e}|Q^e_{t-i+1}(\theta_1)-Q^e_{t-i+1}(\theta_2)|+\sum_{i=1}^{L_r}|Q^r_{t-i+1}(\theta_1)-Q^r_{t-i+1}(\theta_2)|\\ \le W(t)\delta+(W(t)+1)\delta+(2W(t)+2)\delta\\ \le (4W(t)+3)\delta=W(t+1)\delta,\end{aligned}$$

where the first part is from (EC.14), the second part is from (EC.15), and the third part is from (EC.18). This completes the induction part of Lemma D.3. Furthermore, we directly obtain that $|I_t(\theta_1)-I_t(\theta_2)|\le W(t)\delta$. Also,

$$|C^{\theta_1}_t-C^{\theta_2}_t|\le(h+p+c_r+c_e)[|I_t(\theta_1)-I_t(\theta_2)|+|Q^e_t(\theta_1)-Q^e_t(\theta_2)|+|Q^r_t(\theta_1)-Q^r_t(\theta_2)|] \tag{EC.19}$$

$$\le(h+p+c_r+c_e)W(t)\delta, \tag{EC.20}$$

as over-stocking and shortage penalty only relate to $|I_{t+1}(\theta_1)-I_{t+1}(\theta_2)|$, and purchasing cost only relates to $|Q^e_t(\theta_1)-Q^e_t(\theta_2)|$ and $|Q^r_t(\theta_1)-Q^r_t(\theta_2)|$. Here h, p, c_r, c_e refer to the holding cost, the shortage penalty, the long-lead purchasing cost, and the short-lead purchasing cost coefficient, respectively. □

Finally we combine Eq. (EC.13) with Lemma D.3 to establish the Lipschitz continuity. Note that

$$\begin{aligned}|g_{\theta_1}-g_{\theta_2}|&\le\frac{\mathbb{E}\left[\sum_{t\le Y_1}|C^{\theta_1}_t-C^{\theta_2}_t|\right]}{\mathbb{E}[Y_1]}\\ &\le\frac{\mathbb{E}\left[\sum_{t\le Y_1}(h+p+c_r+c_e)W(t)\delta\right]}{\mathbb{E}[Y_1]}\\ &\le(h+p+c_r+c_e)\delta\frac{\sum_{t=1}^{\infty}\mathbb{P}(Y_1=t)W(t)}{\mathbb{E}[Y_1]}.\end{aligned}$$

Also note that the distribution of Y_1 is given in Drekic and Spivey (2021), and known to be sub-exponential, and $\mathbb{E}[Y_1]$ is known to be finite. Also, $\sum_{t=1}^{\infty}\mathbb{P}(Y_1=t)W(t)$ is fully characterized by L_r and γ by our definition of Y_1. Assuming $\sum_{t=1}^{\infty}\mathbb{P}(Y_1=t)W(t)$ to be finite, the desired Lipschitz continuity follows with

$$L_\Theta=(h+p+c_r+c_e)\frac{\sum_{t=1}^{\infty}\mathbb{P}(Y_1=t)W(t)}{\mathbb{E}[Y_1]}.$$

Due to the distribution of Y_1 given in Drekic and Spivey (2021), $\sum_{t=1}^{\infty}\mathbb{P}(Y_1=t)W(t)$ is finite when $\gamma\to 1$. However, $\sum_{t=1}^{\infty}\mathbb{P}(Y_1=t)W(t)$ is not necessarily bounded for any $\gamma>0$. Correspondingly, with γ sufficiently large to have a bounded L_Θ, the L_Θ term in Theorem 1 is dominated by $H\sqrt{\frac{wd}{\alpha}\log\left(\frac{U\sqrt{T}\log T}{\delta}\right)}$ for sufficiently large T. For this reason, we omit the explicit L_Θ term in our regret bound. □

REMARK EC.3. While the exact distribution of Y_1 is given in Drekic and Spivey (2021), its closed-form is cumbersome to analyze. Instead, we compute the exact value of c_γ numerically in a specific example to assess how restrictive this requirement is for ensuring Lipschitz continuity: when $L_r=2$, taking $\gamma>0.96$

suffices. The exact value of c_γ for other hyperparameter settings can be computed in the same manner. Recall that most SKUs in real-world inventories are slow-moving. Thus, our approach still yields $\sqrt{T}$ regret for the majority of those items.

We believe that one can further relax the constraints on L_r and γ by carefully analyzing the growth of the inventory gap between any two policies. Since each exponential jump requires on the order of $\Omega(L_r)$ increments, the effective growth rate becomes slower, weakening the necessary bounds on L_r and γ. Nevertheless, because the gap still grows exponentially, a fully general Lipschitz proof remains elusive for arbitrary L_r and γ. We defer a complete characterization of sufficient conditions for Lipschitz continuity for future work.

D.2.3. Existence of Distributional Ordering

DEFINITION 4. Consider a fixed $\alpha \in (0,1]$ and a trajectory $\mathcal{H}_\theta^T$ under policy θ. Define the hitting times

$$\tau_i = \inf\{t > \tau_{i-1} : I_t^\theta = z_r^\theta\}, \quad \tau_0 = 0,$$

and let $\mathcal{I}_\theta = \{\tau_1, \tau_2, \dots\}$. Denote $\tilde{D}_t$ as the sales observed at time t. For any $\theta' \in \Theta$, if $|\mathcal{I}_\theta| \geq \alpha T$, we define the counterfactual trajectory $\tilde{\mathcal{H}}_{\theta'}^{\alpha T}$ by simulating policy θ' from the initial state s_1 assuming the true demand sequence is $\{\tilde{D}_{\tau_i}\}_{i=1}^{\alpha T}$. Then the counterfactual estimate $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ is defined as:

$$\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) = \begin{cases} G_{\theta'}(\tilde{\mathcal{H}}_{\theta'}^{\alpha T}), & \text{if } |\mathcal{I}_\theta| \geq \alpha T, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Recall that $\gamma = \mathbb{P}(D_t = 0)$. We now present a sequence of lemmas demonstrating how the estimator $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ defined in Equation (10) can be leveraged to establish a distributional policy order.

LEMMA D.4. *Let $\{I_1^\theta, I_2^\theta, \dots, I_T^\theta\}$ denote an observed sequence of on-hand inventories under any policy $\theta \in \Theta$ starting from s_1. Then for any $\delta > 0$, with probability at least $1-\delta$,*

$$\left|\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta) - \mathbb{E}[\lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta)]\right| \leq \frac{1}{1-\gamma}\left(\frac{1}{\gamma^{L_r} T} + \frac{1}{\gamma^{L_r}}\sqrt{\frac{2\log(2/\delta)}{T}}\right),$$

where $\mathbb{I}(\cdot)$ refers to the indicator function.

Proof. Recall from Lemma A.2 that the empirical estimator concentrates provided the gain is uniform across states and the span of the bias is bounded. We verify these conditions for the cost function defined by $\mathbb{I}(I_t^\theta = z_r^\theta)$, which indicates whether the on-hand inventory equals z_r^θ at time t. To apply Lemma A.2 and prove Lemma D.4, it suffices to show: (i) the gain function is uniform across states, and (ii) the bias span is bounded by $\frac{1}{(1-\gamma)\gamma^{L_r}}$.

However, Lemma 3 only relied on the costs being bounded in $[0,1]$, which applies under this cost function. Thus, the two conditions are satisfied directly. Applying Lemma A.2 yields the bound here. □

The previous lemma establishes that for an observed trajectory $\mathcal{H}_\theta^T$ of length T under policy θ, we have a lower bound of the number of observations where the on-hand inventory is z_r^θ. Formally, we have:

LEMMA D.5. *Let $\theta \in \Theta$ be any policy. For any $\delta > 0$, suppose $T \geq T_h(\delta)$, where $T_h(\delta) = \Omega(\frac{\log(1/\delta)}{(1-\gamma)^4\gamma^{2L_r}})$. Then, with probability at least $1-\delta$, we have:*

$$\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta) \geq \frac{1-\gamma}{2}\gamma^{L_r}T.$$

Proof. From Lemma D.4, we have that with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta) &\geq T \cdot \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta)\right] - \frac{T}{1-\gamma}\left(\frac{1}{\gamma^{L_r}T} + \frac{1}{\gamma^{L_r}}\sqrt{\frac{2\log(2/\delta)}{T}}\right) \\
&\geq \frac{T}{\mathbb{E}[Y_1]} - \frac{1}{(1-\gamma)\gamma^{L_r}} - \frac{1}{(1-\gamma)\gamma^{L_r}}\sqrt{2\log(2/\delta)T} \\
&\geq (1-\gamma)\gamma^{L_r}T - \frac{1}{(1-\gamma)\gamma^{L_r}} - \frac{1}{(1-\gamma)\gamma^{L_r}}\sqrt{2\log(2/\delta)T}.
\end{aligned}$$

The second inequality uses the fact that $\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta) \geq N(T)$ almost surely (with $N(T)$ from Equation (EC.10)), as each occurrence of L_r consecutive zero demands implies at least one t with $I_t^\theta = z_r^\theta$. The third inequality applies $\mathbb{E}[Y_1] \leq \frac{1}{(1-\gamma)\gamma^{L_r}}$ from Lemma D.2.

To ensure the right-hand side exceeds $\frac{1-\gamma}{2}\gamma^{L_r}T$, it suffices that:

$$\sqrt{T} \geq \frac{2\sqrt{2\log\left(\frac{2}{\delta}\right)} + 2}{(1-\gamma)^2\gamma^{L_r}}. \tag{EC.21}$$

Thus, for all $T \geq T_h(\delta)$ satisfying Equation (EC.21), we conclude:

$$\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta) \geq \frac{1-\gamma}{2}\gamma^{L_r}T.$$

□

Combining Lemma D.5 with the definition of our counterfactual estimator in Definition 4, we formally establish the distributional equivalence that underpins the policy ordering.

LEMMA D.6. *Let $\alpha = \frac{1-\gamma}{2}\gamma^{L_r}$. For any policy $\theta \in \Theta$, any $\delta > 0$, and any $T \geq T_h(\delta)$ where the definition of $T_h(\delta)$ follows Lemma D.5, define the event*

$$\mathcal{E}_4(T,\delta) = \left\{\sum_{t=1}^{T} \mathbb{I}(I_t^\theta = z_r^\theta) \geq \frac{1-\gamma}{2}\gamma^{L_r}T\right\}.$$

Then, conditioned on $\mathcal{E}_4(T,\delta)$, for any $\theta' \in \Theta$ such that $z_r^{\theta'} \leq z_r^\theta$, the counterfactual estimator $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ has the same distribution as $G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha T})$.

Proof. This result follows directly from the definition of $\tilde{G}_{\theta'|\theta}$ in Definition 4. Under the event $\mathcal{E}_4(T,\delta)$, the index set $\mathcal{I}_\theta$ contains at least αT time steps at which the on-hand inventory equals z_r^θ, thus $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) = G_{\theta'}(\tilde{\mathcal{H}}_{\theta'}^{\alpha T})$. At these time steps, the observed sales are uncensored with respect to any policy θ' such that $z_r^{\theta'} \le z_r^\theta$, and thus can be treated as valid demand samples for estimating $G_{\theta'}$. As a result, the pseudo-trajectory $\tilde{\mathcal{H}}_{\theta'}^{\alpha T}$ constructed from these samples under policy θ' yields a counterfactual estimate $G_{\theta'}(\tilde{\mathcal{H}}_{\theta'}^{\alpha T})$ that is identically distributed to $G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha T})$. □

LEMMA 2. *Set $\gamma = \mathbb{P}(D_t = 0)$ and let $\alpha = \frac{1-\gamma}{2}\gamma^{L_r}$. Fix any policy $\theta \in \Theta$, confidence level $\delta > 0$, and time horizon $T \ge T_h(\delta)$, where $T_h(\delta) = \Omega(\frac{\log(1/\delta)}{(1-\gamma)^4\gamma^{2L_r}})$. Let $\tilde{G}_{\theta'|\theta}$ be as specified in Definition 4. Then, for any policy $\theta' \in \Theta$ such that $z_r^{\theta'} \le z_r^\theta$, it follows that $\pi_{\theta'} \preceq \pi_\theta$.*

Proof. We denote $G_{\theta'}(\mathcal{H}_{\theta'}^{\alpha T})$ as $G_{\theta'}$, and $\mathcal{E}_4(T,\delta)$ as $\mathcal{E}_4$ in this proof for simplicity. For any Borel measurable set B, we have that

$$\begin{aligned}
\mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B) &= \mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B|\mathcal{E}_4)\mathbb{P}(\mathcal{E}_4) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B|\mathcal{E}_4^c)\mathbb{P}(\mathcal{E}_4^c) \\
&\le \mathbb{P}(G_{\theta'} \in B) - (1-\delta)\mathbb{P}(G_{\theta'} \in B) \\
&= \delta\mathbb{P}(G_{\theta'} \in B) \\
&\le \delta,
\end{aligned}$$

where the second line is due to $\mathbb{P}(\mathcal{E}_4) \ge 1-\delta$, Lemma D.5 and Lemma D.6. Similarly,

$$\begin{aligned}
\mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B) &= \mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B|\mathcal{E}_4)\mathbb{P}(\mathcal{E}_4) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B|\mathcal{E}_4^c)\mathbb{P}(\mathcal{E}_4^c) \\
&\ge \mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B|\mathcal{E}_4^c)\mathbb{P}(\mathcal{E}_4^c) \\
&\ge -\mathbb{P}(\mathcal{E}_4^c) \\
&\ge -\delta,
\end{aligned}$$

where the second line is due to $\mathbb{P}(\mathcal{E}_4) \le 1$, and the last line is due to $\mathbb{P}(\mathcal{E}_4) \ge 1-\delta$. Thus we have

$$|\mathbb{P}(G_{\theta'} \in B) - \mathbb{P}(\tilde{G}_{\theta'|\theta} \in B)| \le \delta,$$

thus $\theta' \preceq \theta$. □

D.3. M/M/1/L Queuing Model with Service Rate Control (Section 4.3)

D.3.1. Policy Order and Counterfactual Estimation In this case, we modify the information-order definition from Definition 3, as we do not use empirical estimators (sample average); instead, we directly estimate the long-term average cost $g_{\theta'}$ for each θ'. Thus for any two policies θ and θ', we write $\pi_{\theta'} \preceq \pi_\theta$ if one can construct $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$ such that with probability at least $1-\delta$,

$$|\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) - g_{\theta'}| \le O(H\sqrt{1/\alpha T}).$$

Using this definition of the information order, one can simply replace this result with Lemma A.1 in Theorem 1 to obtain the same regret bound. In this system, we exploit the fact that the performance of any policy is exactly characterized by λ and μ. Hence, we use a "plug-in" approach to calculate our estimators $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$. This contrasts with the empirical average cost from a trajectory sampled under a policy, as was our default estimator to establish a policy order in Definition 3. In other words, this case study exemplifies that the policy order in Definition 3 does not explicitly rely on the empirical average cost, and extends to any well-defined estimator.

LEMMA 5. *For any T sufficiently large and any two policies θ and θ', there exists $\tilde{G}_{\theta'|\theta}$ such that with probability at least $1-\delta$,*

$$|\tilde{G}_{\theta'|\theta} - g_{\theta'}| \le O\left(L^3\sqrt{\frac{\log(1/\delta)}{T}}\right).$$

As a result, $\pi_{\theta'} \preceq \pi_\theta$.

Proof. With a trajectory $\mathcal{H}_\theta^T$ note that the intervals we used to estimate $\hat{\lambda}$ and $\hat{\mu}$ are independent and sub-exponential. Since U is a fixed constant in uniformization, for a fixed $\delta > 0$, with Berstein's inequality (see Theorem 2.8.1 of (Vershynin 2018)), for some $c_1, c_2 > 0$ and all $\epsilon \in (0,1)$, for T sufficiently large,

$$\mathbb{P}\Big(|\hat{\lambda}^{-1} - \lambda^{-1}| \ge \epsilon\Big) \le \exp\big(-c_1 T\epsilon^2\big) \tag{EC.22}$$

$$\mathbb{P}\Big(|(L\hat{\mu}+\theta_L)^{-1} - (L\mu+\theta_L)^{-1}| \ge \epsilon\Big) \le \exp\big(-c_2 T\epsilon^2\big), \tag{EC.23}$$

where some absolute constant factors of U, $\lambda_{\max}$, and $\mu_{\max}$ are omitted. Note that we implicitly use the fact that the visits to state 0 and state L are $\Theta(T)$, as the system is ergodic (Chapter 13 of Meyn and Tweedie (2012)). Since (λ, μ) are bounded by $\lambda_{\max}$ and $\mu_{\max}$, with T sufficiently large, we always have that for any $\delta > 0$, with probability at least $1 - 2\delta$,

$$|\hat{\lambda} - \lambda| = O(\epsilon), |\hat{\mu} - \mu| = O(L\epsilon),$$

where we denote $\epsilon = \sqrt{\frac{1}{T}\log(\frac{1}{\delta})}$. We denote this event as $\mathcal{E}_q$:

$$\mathcal{E}_q = \{|\hat{\lambda} - \lambda| = O(\epsilon), |\hat{\mu} - \mu| = O(L\epsilon)\}.$$

Thus for any policy $\theta' \in \Theta$, with $\hat{\lambda}$ and $\hat{\mu}$ in hand, we can derive a unique stationary distribution $\hat{m}_{\theta'}(s)$ (Puterman 2014). For the remainder of the proof we condition on the event $\mathcal{E}_q$.

In the true underlying continuous time Markov chain, denote by $Q(\lambda, \mu, \theta')$ as the Q generator matrix under rates λ and μ following policy $\pi_{\theta'}$. Then we see that $\Delta Q = Q(\hat{\lambda}, \hat{\mu}, \theta') - Q(\lambda, \mu, \theta')$ has at most three nonzero entries per row. Using this we have:

$$\|\Delta Q\|_\infty = \max_i \sum_j |\Delta Q_{ij}| \le 2|\hat{\lambda} - \lambda| + 2\max_s |s\hat{\mu} + \theta'_s - s\mu - \theta'_s| = O(L^2\epsilon).$$

By Theorem 2.1 of Mitrophanov (2003),

$$d_{\mathrm{TV}}\big(m_{\theta'}, \hat{m}_{\theta'}\big) = O(L^2\epsilon). \tag{EC.24}$$

Hence by the definition of $\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T)$:

$$\begin{aligned}
|\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) - g_{\theta'}| &= |\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) - \sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)] + \sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)] - g_{\theta'}| \\
&\le |\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) - \sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)]| + |\sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)] - g_{\theta'}| \\
&= |\sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}\Big[\hat{C}(s,\pi_{\theta'}(s)\Big] - \sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)]| \\
&\quad + |\sum_{s=0}^{L} \hat{m}_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)] - \sum_{s=0}^{L} m_{\theta'}(s)\mathbb{E}[C(s,\pi_{\theta'}(s)]| \\
&\le \sum_{s=0}^{L} \hat{m}_{\theta'}(s)|\mathbb{E}\Big[\hat{C}(s,\pi_{\theta'}(s)\Big] - \mathbb{E}[C(s,\pi_{\theta'}(s)]| \\
&\quad + C_{\max}\sum_{s=0}^{L}|\hat{m}_{\theta'}(s) - m_{\theta'}(s)| \\
&\le 2C\frac{L}{U}|\mu - \hat{\mu}| \cdot U + 2C_{\max} d_{\mathrm{TV}}(m_{\theta'}, \hat{m}_{\theta'}) \\
&\le O(L^3\epsilon).
\end{aligned}$$

The second inequality is due to Hölder's inequality, and $C_{\max} = w(A_{\max}) + CL\mu_{\max}$ is an upper bound on the expected cost per timestep. The term U in the last second line is due to the gap between unit time and unit jump. Hence we have under event $\mathcal{E}_q$,

$$|\tilde{G}_{\theta'|\theta}(\mathcal{H}_\theta^T) - g_{\theta'}| \le \tilde{O}\left(L^3\sqrt{\frac{2\log(2/\delta)}{T}}\right).$$

□