Exploiting Exogenous Structure for Sample-Efficient Reinforcement Learning

Jia Wan[1*] Sean R. Sinclair[1] Devavrat Shah[1] Martin Wainwright [1]

[1] Massachusetts Institute of Technology
{jiawan,seansinc,devavrat,mjwain}@mit.edu

Abstract

We study a class of structured Markov Decision Processes (MDPs) known as Exo-MDPs, characterized by a partition of the state space into two components. The *exogenous* states evolve stochastically in a manner not affected by the agent's actions, whereas the *endogenous* states are affected by the actions, and evolve in a deterministic and known way conditional on the exogenous states. Exo-MDPs are a natural model for various applications including inventory control, finance, power systems, ride sharing, among others. Despite seeming restrictive, this work establishes that any discrete MDP can be represented as an Exo-MDP. Further, Exo-MDPs induce a natural representation of the transition and reward dynamics as linear functions of the exogenous state distribution. This linear representation leads to near-optimal algorithms with regret guarantees scaling only with the (effective) size of the exogenous state space d, independent of the sizes of the endogenous state and action spaces. Specifically, when the exogenous state is fully observed, a simple plug-in approach achieves a regret upper bound of $\tilde{O}(H^{3/2}\sqrt{dK})$, where H denotes the horizon and K denotes the total number of episodes. When the exogenous state is unobserved, the linear representation leads to a regret upper bound of $\tilde{O}(H^{3/2}d\sqrt{K})$. We also establish a nearly matching regret lower bound of $\Omega(Hd\sqrt{K})$ for the no observation regime. An experimental study for inventory control complements these theoretical findings.

1 Introduction

Reinforcement learning (RL) provides a natural framework for sequential decision-making under uncertainty. The past few decades have witnessed tremendous empirical success from RL, notably in "data-rich" areas such as competitive game-playing [20], computational advertising [22], robotics [13], and human-guided training of large language models [16]. This success relies on the availability of massive datasets, either due to large amounts of pre-collected data or via access to simulators for generating data. In contrast, there are various other application domains that are notoriously "data-poor", including finance [18], resource allocation [10], inventory control [14], supply chain management [19], as well as ridesharing systems [5]. The limited data arises from various causes, including the small quantity of pre-collected data, difficulty in collecting new data, and/or the lack of good simulators. In such data-limited settings—and in the absence of structure on the underlying Markov decision processes (MDPs)—information-theoretic lower bounds dictate that a good RL policy cannot be learned without large sample sizes. Therefore, it is essential to identify and exploit domain-specific structures so as to enable data efficient RL policy learning.

With this motivation in mind, we focus on a structured family of Markov decision processes known as Exo-MDPs (e.g., [6, 7, 17, 21, 8]). They are defined by a partition of state variables into exogenous

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

versus endogenous states. More specifically, we say that a state variable is *exogenous* if it evolves in a way that is *not influenced by the agent's actions*; otherwise the state is *endogenous*. All stochasticity in the system dynamics is captured via the exogenous states, while the endogenous state variables evolve according to a known deterministic function of the endogenous states, the agent's action, as well as the exogenous state variables. For example, in the classic inventory control (i.e., newsvendor) problem in supply chain, the external demand represents the exogenous state, the inventory in the system represents the endogenous state, the action corresponds to placing a new purchase order, and the inventory in the system evolves as a function of existing inventory, the exogenous demand, and the purchase orders placed; see Section 2.1 for details on this example. Similarly, the efficiency of a ridesharing system is tied to fluctuating demand levels exogenous to the system itself [8].

Exo-MDPs hold promise for designing data-efficient simulators and hence the identification of optimal RL policies, due to the fact that all randomness is captured through exogenous states that are not impacted by the actions or the endogenous states as well as known dynamics of the endogenous states. This insight has been in recent prior works such as Sinclair et al. [21], Mao et al. [15]. However, they crucially assume the exogenous variables are completely observed. This assumption is simplistic and does not hold in many real-world systems. For example, inventory models often encounter lost sales, where the true exogenous demand is unobserved due to stockouts. Similarly, ridesharing systems exhibit demand shortfalls when drivers are unavailable, resulting in users leaving the platform. Accordingly, this paper tackles the following question:

Challenge: *How to exploit Exo-MDP structure to learn policies in a sample-efficient manner with no (or partial) observation of exogenous states?*

Contributions. Let us briefly summarize the main contributions of this paper. Our first result is structural in nature: we show that for tabular MDPs, the Exo-MDP assumption is actually not limiting; any tabular MDP can be represented as an Exo-MDP. Moreover, any Exo-MDP can be be viewed as an instance of a discrete linear mixture MDP. The arguments used to establish these relations reveal interesting structural properties of these classes, and also inform our subsequent study into the effective dimension of an Exo-MDP.

Second, we provide sample-efficient learning algorithms that exploit the Exo-MDP structure. We do so both in the *full observation regime* in which the exogenous states are observed, and the more challenging *no observation regime*, in which the exogenous states are entirely unobserved. When the exogenous states are fully observed, we analyze a plug-in approach and prove that it achieves a regret upper bound of $\tilde{O}(H^{3/2}\sqrt{dK})$ in terms of the horizon H, total number of episodes K, and dimension d of the exogenous state. On the other hand, for problems in which exogenous states are not unobserved, we first make use of the linear mixture representation of an Exo-MDP, thereby obtaining an algorithm with nearly-optimal regret upper bound of $\tilde{O}(H^{3/2}d\sqrt{K})$. We then introduce a notion of *effective dimension* r, and establish a sharper guarantee $\tilde{O}(H^{3/2}r\sqrt{K})$. The term r captures the effective dimension of the feature space, and can be computed a priori without any samples. We complement our upper bounds by proving a lower bound for the no observation regime which scales as $\Omega(Hd\sqrt{K})$, thereby matching our upper bound up to a factor of $\sqrt{H}$. We show that for a more general version of Exo-MDPs—in which the exogenous dynamics can differ from stage to stage—it is possible to achieve the stronger lower bound with the additional $\sqrt{H}$ factor. Combined with our upper bound which also solves the more general case, this characterizes the minimax optimal rate for non-stationary Exo-MDPs up to polylogarithmic factors.

Finally, we complement our theoretical results with an experimental study applying our Exo-MDP algorithms to an inventory control problem with lost sales and positive lead time. Our results highlight the robustness of our algorithms, where despite being more general solvers, they achieve performances competitive with state-of-the-art algorithms tailored for inventory control.

Related work. The past years have witnessed an evolving line of work on Exo-MDPs (e.g., [6, 7, 17, 14, 2, 21, 8]). Some researchers [6, 7] have studied the case when the rewards or transitions factorize so that the exogenous process can be filtered out. While doing so simplifies development, it can lead to sub-optimality, since policies agnostic to the exogenous states need not be optimal. Other work studies the use of hindsight optimization, showing that the regret for hindsight optimization policies can be bounded by the hindsight bias, a problem-dependent term [21, 8]. This work assumes

full observation of the exogeous states, whereas (in addition to this case), we also study the more challenging problem of solving Exo-MDPs with unobserved exogenous states.

As shown in this paper, any Exo-MDP can be cast as a linear mixture MDP (and vice versa), so our analysis establishes connections to the literature on linear mixture MDPs [11, 4, 23]. Ayoub et al. [3] proved an $\Omega(H\sqrt{dK})$ lower bound on the regret for *stationary* linear mixture MDPs. Despite Exo-MDPs being a subclass of this problem class, we are able to prove a regret lower bound that is tighter by a factor of $\sqrt{d}$.

Lastly, we provide experimental results on inventory control with lost sales, censored demand, and positive lead times. Agrawal and Jia [1] design an online learning algorithm to learn the optimal base-stock policy, a well-known heuristic policy that is optimal under restrictive settings. Our empirical results show that our algorithms can surpass the sub-optimality of this heuristic class and instead converge to the true optimal policy. Other authors [14, 2] studied specializations of Exo-MDPs to these inventory settings, along with associated regret analysis. Their analysis is predicated on observing an unbiased signal from which the true demand can be recovered; in sharp contrast, our algorithms apply even when the demand is fully unobserved.

Notation. For a positive integer n, we denote $[n] \coloneqq \{1, 2, \ldots, n\}$. For a finite set $\mathcal{S}$, let $|\mathcal{S}|$ denote its cardinality. We use calligraphic letters to denote sets, e.g., $\mathcal{S}, \mathcal{A}$; capital letters denote random variables, e.g., S, A, R; lower case letters denote specific realization of random variables, e.g., s, a, r; and for a distribution over a discrete set of elements, we use bolded lower case letters to denote the probability vector corresponding to the multinomial distribution, e.g., $\mathbf{p}_x$. We use lower case letters with superscripts, e.g., $x^j \in \mathcal{X}$ to denote elements of a set $\mathcal{X}$ indexed by j. For vector x, we use $[x]_j$ to denote its j-th entry. We use $\tilde{O}(\cdot)$ to denote rates omitting absolute constants and polylogarithmic factors. Fixing an episode k, $h \in [H]$ denotes the h-th stage of the MDP. Lastly, we let $(x)^+ = \max\{x, 0\}$.

2 Background and Problem Set-up

Throughout this paper, we consider stationary episodic tabular Markov decision processes (MDPs) with finite state and action spaces. We define any MDP with a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, \mathbb{P}, R)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, horizon H is the number of stages in each episode, s_1 is a fixed initial state, $\mathbb{P}(\cdot \mid s_h, a_h)$ gives the probability distribution over the next state s_{h+1} based on the state action pair s_h, a_h at stage h in an episode, and assume bounded stochastic reward $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ at each stage h. Without loss of generality, throughout this paper we assume a fixed starting state s_1.

2.1 Exo-MDP: Markov Decision Processes with Exogenous States

We now consider a specialized class of MDPs with exogenous states (Exo-MDPs), where the state space can be partitioned into two parts: the *endogenous states* $\mathcal{S}$, and the *exogenous states* $\mathcal{X}$ [7, 21]. Both the endogenous and exogenous states affect the dynamics of the system, but the agent's actions only influence the dynamics of the endogenous states, *not* the exogenous states. See Figure 1 for an illustration of the distinctions between a standard MDP and an Exo-MDP.

More precisely, any Exo-MDP is represented by a tuple $\mathcal{M}[\mathbb{P}_x, \mathbf{f}, \mathbf{g}] = (\mathcal{S} \times \mathcal{X}, \mathcal{A}, H, s_1, \mathbb{P}, R)$. In an Exo-MDP, the state vector at stage h takes the form (S_h, X_h), where S_h and X_h are endogenous and exogenous, respectively. The exogenous state evolves in a stationary way independent of (S_h, A_h), where each X_h is an i.i.d. sample from an unknown distribution $\mathbb{P}_x$. We fix indexings $\mathcal{X} = \{x^j\}_{j=1}^d$ and let $\mathbf{p}_x$ denote the probability vector corresponding to $\mathbb{P}_x$, where $\mathbf{p}_x = (\mathbb{P}_x(X = x^1), \ldots, \mathbb{P}_x(X = x^{|\mathcal{X}|})) \in [0, 1]^{|\mathcal{X}|}$. We denote $d = |\mathcal{X}|$ as the cardinality of the exogenous state space. As we will show later, one can view d as the effective dimension to summarize the Exo-MDP, leading to sample complexity results that only depend on d regardless of the sizes of the endogenous state and action spaces $|\mathcal{S}|, |\mathcal{A}|$.

Additionally, we assume that, conditional on the realization of $X_1, \ldots, X_H$ in an episode, the transition and reward are completely specified by known deterministic functions $\mathbf{f}$ and $\mathbf{g}$. Specifically, the next state S_{h+1} given triple (S_h, A_h, X_h) follows

$$S_{h+1} = \mathbf{f}(S_h, A_h, X_h) \qquad \text{where } \mathbf{f} : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \to \mathcal{S}. \tag{1a}$$

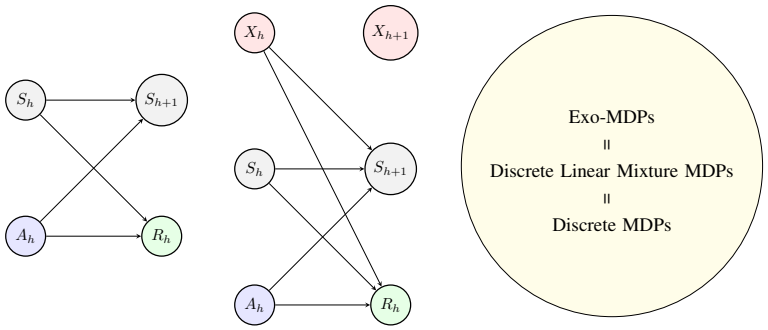

Figure 1. Directed graphical models showing an ordinary MDP (left), and an Exo-MDP (middle). In an ordinary MDP, the state space is fully endogenous, and the current state S_h and action A_h generate the next state S_{h+1} and reward R_h. In an Exo-MDP, the state vector is partitioned into two components: an endogenous component S_h and an exogenous component X_h. The exogenous state X_h at each stage is drawn i.i.d from $\mathbb{P}_x$ independent of (S_h, A_h). There are also known deterministic functions $\mathbf{f}$ and $\mathbf{g}$ such that $S_{h+1} = \mathbf{f}(S_h, A_h, X_h)$ and $R_h = \mathbf{g}(S_h, A_h, X_h)$. The right panel gives the structural equivalence relations between the class of Exo-MDPs, discrete MDPs and discrete linear mixture MDPs.

Similarly, the reward R_h at stage h is given by

$$R_h = \mathbf{g}(S_h, A_h, X_h) \qquad \text{where } \mathbf{g} : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \to [0, 1]. \tag{1b}$$

To put the Exo-MDP formulation in action, consider the following simple setting of inventory control.

Example: inventory control. Suppose a retailer needs to order products to meet exogenous independent demand at each stage h over a finite horizon H. Given current on-hand inventory Inv_h, the retailer picks an amount O_h of products to order. The inventory level then transitions to $\text{Inv}_{h+1} = \mathbf{f}(\text{Inv}_h, O_h, X_h) = (\text{Inv}_h + O_h - X_h)^+$, where X_h denotes the exogenous demand drawn i.i.d. from $\mathbb{P}_x$. The cost $\mathbf{g}(\text{Inv}_h, O_h, X_h)$ (negative reward) consists of the holding cost for remaining products $c(\text{Inv}_h + O_h - X_h)^+$, plus the penalty for lost sales $p(X_h - \text{Inv}_h - O_h)^+$. This can be formulated as an Exo-MDP where d denotes the size of the support for the demand X_h and the state and action correspond to inventory Inv_h and orders O_h respectively. The true exogenous state X_h is unobserved, only the realized sales $\min\{\text{Inv}_h + O_h, X_h\}$.

2.2 Data Setting and Learning Objective in an Exo-MDP

Observation regimes on x. In this paper, we mainly study two observation regimes on the exogenous state x: (i) the *full observation regime* (Section 4.1), where the learning agent observes (S_h, A_h, R_h, X_h) at each stage; and (ii) the *no observation regime* (Section 4.2), where the learning agent observes (S_h, A_h, R_h) with no observation on X_h. We focus on these two extremes regarding observations on the exogenous state x, leaving as an open direction other (e.g., partial or censored) observation regimes which may lead to sample complexities interpolating between the two.

Policies. We consider (stochastic) policies $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes a distribution over the action space. Importantly, in the case of Exo-MDPs, the policy is *not* allowed to depend on the exogenous state X_h. An algorithm that decides the policy depends on the historical trajectory $\{S_{h,k}, A_{h,k}, R_{h,k}, X_{h,k}\}_{h\in[H],k\in[K]}$ for the full observation regime; and $\{S_{h,k}, A_{h,k}, R_{h,k}\}_{h\in[H],k\in[K]}$ for the no observation regime.

Online learning. We focus on the online learning setting for solving Exo-MDPs. At the start of the Exo-MDP, the learning agent is given $\mathcal{S}, \mathcal{A}, \mathcal{X}, H$ and functions $\mathbf{f}, \mathbf{g}$, but does not know the vector $\mathbf{p}_x$. The agent interacts with the environment for K episodes. At the beginning of each episode $k \in [K]$, the agent fixes a policy π^k. At each stage $h \in [H]$, the agent observes state $S_{h,k}$ and picks action $A_{h,k} \sim \pi^k_h(\cdot \mid S_{h,k})$. The exogenous state $X_{h,k}$ is then sampled from $\mathbb{P}_x$, and the agent receives reward $R_{h,k} = R(S_{h,k}, A_{h,k}) = \mathbf{g}(S_{h,k}, A_{h,k}, X_{h,k})$, and transitions to the state $S_{h+1,k} = \mathbf{f}(S_{h,k}, A_{h,k}, X_{h,k})$. Under the full observation regime, the agent additionally observes $X_{h,k}$ at the end of the stage. This continues until the final transition to state $S_{H,k}$, at which point the agent chooses policy π^{k+1} for the next episode. We denote the value function $V^\pi : \mathcal{S} \times [H] \to \mathbb{R}$ of

a policy π under MDP $\mathcal{M}$[1] as $V_h^\pi(s, \mathcal{M}) := \mathbb{E}_{X_{\geq h}, \pi}[\sum_{\tau \geq h} R(S_\tau, A_\tau, X_\tau) \mid S_h = s]$, where $X_{\geq h}$ denotes vector $(X_h, \ldots, X_H)$. Let $V_h^*(s, \mathcal{M})$ denote the optimal value function, i.e., $V_h^*(s, \mathcal{M}) = V_h^{\pi^*}(s, \mathcal{M})$ where $\pi^* = \arg\max_\pi V_h^\pi(s, \mathcal{M})$ is the optimal policy.

Performance metrics. The goal is to design an algorithm that minimizes *regret*, which is the cumulative difference in total reward of the sequence of policies employed by the algorithm $(\pi^k)_{k \in [K]}$ to that of the optimal policy. Specifically, $\textsc{Regret}(K) = \sum_{k=1}^K V_1^*(s_1) - V_1^{\pi^k}(s_1)$. As in Jin et al. [12], any algorithm with a regret upper bound readily converts to a final policy with value function estimation error that is $\frac{1}{K}$ times the regret upper bound.

3 Structural Relations among MDPs

Exo-MDPs by definition are a subclass of MDPs where the transition and reward dynamics are characterized by the restricted forms of Eq. (1a) and Eq. (1b). However, it turns out that Exo-MDPs can represent any discrete MDP with the addition of an exogenous state space. Intuitively, we can *lift* the randomness from transition and reward dynamics as a $2|\mathcal{S}||\mathcal{A}|$-dimensional exogenous state.

Lemma 1. *Let $\mathcal{R}$ denote the range of the reward function R. For any discrete MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, \mathbb{P}, R)$, there exists an exogenous state space $\mathcal{X} \subseteq \mathcal{S}^{|\mathcal{S}||\mathcal{A}|} \times \mathcal{R}^{|\mathcal{S}||\mathcal{A}|}$ following distribution $\mathbb{P}_x$, and transition and reward functions $\mathbf{f}$ and $\mathbf{g}$ such that $\mathcal{M}$ is equivalent to an Exo-MDP $\mathcal{M}'[\mathbb{P}_x, \mathbf{f}, \mathbf{g}] = (\mathcal{S} \times \mathcal{X}, \mathcal{A}, H, s_1, \mathbb{P}, R)$.*

We next show that Exo-MDPs have a natural *linear representation* defined by $\mathbf{f}$ and $\mathbf{g}$. This allows us to cast Exo-MDPs as a linear mixture MDP, a common subclass of MDPs from the literature in which both the transition probability and reward function are linear functions of a given feature mapping over state-action-state triples [23]. Formally,

Definition 1. *MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, \mathbb{P}, R)$ is called a linear mixture MDP if there exists vectors $\theta_p, \theta_r \in \mathbb{R}^d$ and known feature vectors $\phi_r(s, a), \phi_p(s' \mid s, a) \in \mathbb{R}^d$ such that the transition probability satisfies $\mathbb{P}(s' \mid s, a) = \phi_p(s' \mid s, a)^\intercal \theta_p$ and the expected reward satisfies $R(s, a) = \phi_r(s, a)^\intercal \theta_r$.*

Exo-MDPs are a special case of linear mixture MDPs, where the features are characterized by the given forms of $\mathbf{f}$ and $\mathbf{g}$, and the probability vector $\mathbf{p}_x$ serves as the coefficient on the d-dimensional simplex. Specifically, $\mathbb{P}(s' \mid s, a) = \sum_{x \in \mathcal{X}} \mathbb{1}_{s' = \mathbf{f}(s,a,x)} \mathbb{P}_x(x) = \sum_{i=1}^d \mathbb{1}_{s' = \mathbf{f}(s,a,x^i)} [\mathbf{p}_x]_i = \phi_p(s'|s, a)^\intercal \mathbf{p}_x$ and $R(s, a) = \sum_{x \in \mathcal{X}} \mathbf{g}(s, a, x) \mathbb{P}_x(x) = \sum_{i=1}^d \mathbf{g}(s, a, x^i)[\mathbf{p}_x]_i = \phi_r(s, a)^\intercal \mathbf{p}_x$. This leads to the following lemma representing any Exo-MDP as a linear mixture MDP.

Lemma 2. *Any Exo-MDP $\mathcal{M}[\mathbb{P}_x, \mathbf{f}, \mathbf{g}] = (\mathcal{S} \times \mathcal{X}, \mathcal{A}, H, s_1, \mathbb{P}, R)$ with a fixed indexing $\mathcal{X} = \{x^j\}_{j=1}^d$ is a linear mixture MDP $\widetilde{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, s_1, \mathbb{P}, R)$ with coefficients $\theta_p = \theta_r = \mathbf{p}_x = (\mathbb{P}_x(X = x^1), \ldots, \mathbb{P}_x(X = x^d))$. For any $s \in \mathcal{S}, a \in \mathcal{A}$, the feature vectors are given by*

$$\phi_p(s' \mid s, a) = [\mathbb{1}_{s' = \mathbf{f}(s,a,x^1)}, \ldots, \mathbb{1}_{s' = \mathbf{f}(s,a,x^d)}]^\intercal \quad \phi_r(s, a) = [\mathbf{g}(s, a, x^1), \ldots, \mathbf{g}(s, a, x^d)]^\intercal$$

Lemma 1 and Lemma 2 leads to the following interesting observation that Exo-MDPs, despite their structural assumptions, capture a rich class of MDPs as large as both the class of discrete MDPs and the class of discrete linear mixture MDPs. See the right panel of Fig. 1 for an illustration.

Theorem 1. *The classes of Exo-MDPs, discrete MDPs, and discrete linear mixture MDPs are equivalent.*

4 Sample-efficient Algorithms and Guarantees

We now turn to describing some sample-efficient algorithms for learning optimal policies in Exo-MDPs, along with theoretical bounds on their regret. Section 4.1 is devoted to the full observation regime, whereas Section 4.2 provides guarantees when no exogenous states are observed.

[1] We omit dependence on $\mathcal{M}$ when it is clear from the context.

4.1 Plug-In Method for the Full Observation Regime

In the full observation regime, the agent observes the quadruple (S_h, A_h, R_h, X_h) at each stage $h \in [H]$. Recall that all randomness in an Exo-MDP lies in the exogenous component x, and the functions $(\mathbf{f}, \mathbf{g})$ are known. As a key consequence, estimating the probability vector $\mathbf{p}_x \in \mathbb{R}^d$ is sufficient for estimating the Exo-MDP itself, from which we can compute an optimal policy estimate. These observations motivate a natural plug-in approach for the fully observed case, in which we perform the following two steps: (i) first compute an empirical estimate $\widehat{\mathbf{p}}_x$ using the observations of the exogenous variables; and (ii) use this estimated probability vector to form an estimate $\widehat{\mathcal{M}}$ of the Exo-MDP; and (iii) compute an optimal policy via standard dynamic programming.

More precisely, at the start of each episode $k = 2, 3, \ldots, K$, the agent has access to $(k-1)$-trajectories of exogenous states, each of length H; denote this data set by $\mathcal{D}_k = \{X_{h,k'}\}_{h\in[H],k'<k}$, and observe that it contains a total of $H(k-1)$ samples. We use this data set to compute the empirical distribution

$$\widehat{\mathbf{p}}_x^k := \frac{1}{H(k-1)} \sum_{h\in[H],k'<k} \mathbb{1}_{x=X_{h,k'}} \qquad \text{for } x \in \mathcal{X},$$

and let $\widehat{\mathbf{p}}_x^1$ for $k = 1$ be the uniform distribution. At each episode $k \in [K]$, we construct the estimated MDP $\widehat{\mathcal{M}}^k$ with transition dynamics $S_{h+1} = \mathbf{f}(S_h, A_h, X_h)$, and stochastic rewards $R_h = \mathbf{g}(S_h, A_h, X_h)$, where $X_h \sim \widehat{\mathbf{p}}_x^k$. Finally, we compute the optimal policy

$$\widehat{\pi}^k = \arg\max_{\pi\in\Pi} V_1^\pi(s, \widehat{\mathcal{M}}^k),$$

via standard dynamic programming , with computational complexity polynomial in $|\mathcal{S}|$ and $|\mathcal{A}|$. This procedure yields regret that grows with the exogenous dimension d, as opposed to the cardinalities $|\mathcal{S}|$ and $|\mathcal{A}|$ of the endogenous state and action spaces. We summarize as follows:

Theorem 2. *For any error tolerance $\delta \in (0,1)$ and H-horizon Exo-MDP with exogenous dimension d, the plug-in method, when applied over K episodes, achieves regret at most*

$$\textsc{Regret}(K) \le 9H^{3/2}\sqrt{\{d + 2\log(2K/\delta)\}K}, \tag{2}$$

with probability at least $1-\delta$.

Ignoring the error probability δ and logarithmic factors, we can summarize that the regret is at most $\textsc{Regret}(K) \le \tilde{O}(H^{3/2}\sqrt{dK})$.

4.2 Guarantees for the No Observation Regime

In practice, assuming full observations of the exogenous states may not be realistic. As one concrete example, in inventory control the true demand X_h is not directly observable. Instead, one can only infer a censored signal from the sales $\min(X_h, \text{Inv}_h + O_h)$. Accordingly, in this section, we address the challenge of designing algorithms when the exogenous states are unobserved. For ease of exposition, we assume the expected reward is known and focus on unknown transition dynamics.

In the full observation setting, the plug-in method hinges on the idea that estimating the probability vector $\mathbf{p}_x$ is sufficient to estimate the full Exo-MDP. When x is not observed, it is no longer possible to estimate $\mathbf{p}_x$, but at the same time, it is not always necessary. For instance, in a trivial Exo-MDP with a single state and constant reward, any policy is optimal, and estimating $\mathbf{p}_x$ confers no advantage.

How to capture the difficulty of learning optimal policies in an Exo-MDP? Enumerating the exogenous state space as $\mathcal{X} = \{x^1, \ldots, x^d\}$, recall from Lemma 2 the feature vectors

$$\phi_p(s' \mid s, a) = [\mathbb{1}_{s'=\mathbf{f}(s,a,x^1)}, \ldots, \mathbb{1}_{s'=\mathbf{f}(s,a,x^d)}] \in \mathbb{R}^d. \tag{3}$$

Using these feature vectors, we define the *full information matrix* $F \in \mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|\times d}$ with row $F_{(s',s,a),\cdot} := \phi_p(s' \mid s, a)$ for each triple $(s', s, a) \in |\mathcal{S}|^2|\mathcal{A}|$. The key complexity parameter in our analysis is the *rank* $r := \mathsf{Rank}(F)$ of this full information matrix. Note that this rank can be computed *a priori*—that is, without collecting any data— based on the known sets $\mathcal{X}, \mathcal{S}, \mathcal{A}$ and functions $\mathbf{f}, \mathbf{g}$. Note that r is upper bounded by

$$r := \mathsf{Rank}(F) \le \min\{d := |\mathcal{X}|, |\mathcal{S}|^2\,|\mathcal{A}|\}. \tag{4}$$

Both inequalities can be conservative, and of interest to us in this section is the fact that there exist many Exo-MDPs for which $r \ll d$.

Recall that the feature vectors (3) arose as part of establishing the connection between Exo-MDP and linear MDPs in Lemma 2. This connection is a key enabler: it allows us to leverage algorithms developed for linear mixture MDPs [4, 23]. While other connections are possible, here we adapt the UCRL-VTR$^+$ algorithm [23] to our setting.[2] Our main result is to show that an algorithm that exploits the SVD of the full information matrix can achieve regret that scales with the rank r, as opposed to the ambient dimension d.

Theorem 3. *For any H-horizon Exo-MDP with effective dimension $\mathsf{Rank}(F) = r$, applying a rank-reduced* UCRL-VTR$^+$ *algorithm over K episodes yields a sequence of policies $\{\pi^k\}_{k=1}^K$ with regret at most*

$$\textsc{Regret}(K) \leq \tilde{O}\left(\sqrt{r^2H^2 + rH^3}\sqrt{KH} + r^2H^3 + r^3H^2\right). \tag{5}$$

Note that when $r \geq H$ and $K \geq r^4H + r^3H^2$, we can restate the regret bound (5) more succinctly as $\textsc{Regret}(K) \leq \tilde{O}(rH^{3/2}\sqrt{K})$. Thus, up to poly-logarithmic factors, it grows linearly in the rank r of the full information matrix F. When no rank reduction occurs (i.e., $r = d$), then we recover a regret bound that scales linearly with the cardinality d of the exogenous state space at $\tilde{O}(H^{3/2}d\sqrt{K})$.

Proof sketch. The rowspace of the information matrix F entirely captures all possible transition features $\phi_p(s' \mid s, a)\}_{s,s' \in \mathcal{S}, a \in \mathcal{A}}$ across all state-action-state triples in the Exo-MDP. So the feature space has a low-rank structure if and only if the row-space of F is low-rank. Let $F = U\Sigma V^\intercal$ be the r-dimensional singular value decomposition of F, so that $U \in \mathbb{R}^{S^2A \times r}, \Sigma \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{d \times r}$. Note that by construction, $\mathbb{P}(s' \mid s, a) = \phi_p(s' \mid s, a)\mathbf{p}_x = e_{s'|s,a}F\mathbf{p}_x$, where $e_{s'|s,a}$ is the unit vector with a one in the corresponding entry to (s', s, a). By projecting the feature vector to the r-ranked row space of F, we can rewrite the transition probability as the inner product of the transformed r-dimensional feature and coefficients $\tilde{\phi}_p, \tilde{\theta}_p$ where $\mathbb{P}(s' \mid s, a) = e_{s'|s,a}F\mathbf{p}_x = \left(e_{s'|s,a}U\Sigma\right)\left(V^\intercal \mathbf{p}_x\right) = \tilde{\phi}_p(s' \mid s, a)^\intercal \tilde{\theta}_p$. Running the UCRL-VTR$^+$ algorithm on the linear mixture MDP with feature $\tilde{\phi}_p(s' \mid s, a) = \left(e_{s'|s,a}U\Sigma\right)^\intercal \in \mathbb{R}^r$ and $\tilde{\theta}_p = V^\intercal \mathbf{p}_x$ gives the stated performance.

Note that the full information matrix F only depends on $\mathcal{S}, \mathcal{A}, \mathcal{X}, \mathbf{f}, \mathbf{g}$, all of which are known a priori to the agent, therefore requires no samples to compute. The singular decomposition of F can be computed in time polynomial in $|\mathcal{S}|, |\mathcal{A}|, d$.

5 Regret Lower Bound under the No Observation Regime

In this section, we present a regret lower bound of $\Omega(Hd\sqrt{K})$ for Exo-MDPs under the no observation regime. This almost matches our upper bound of $\tilde{O}(H^{3/2}d\sqrt{K})$ in Section 4.2, showing that the dependence on dimension d and number of episodes K is optimal, while the dependence on horizon H differs by a factor of $\sqrt{H}$. Following prior work, our lower bound is on expected regret, calculated over both the distribution $\mathbb{P}_x$ and the chosen policy. We formally state our regret lower bound for Exo-MDPs below in Theorem 4.

Prior work such as Ayoub et al. [3] provides a lower bound of $\Omega(H\sqrt{dK})$ for *stationary* linear mixture MDPs, that is, the transition and reward dynamics are the same across each stage of an episode. Despite Exo-MDPs being a subclass of stationary linear mixture MDPs, our regret lower bound is tighter by a factor of $\sqrt{d}$. In the appendix, we provide a lower bound for Exo-MDPs when the dynamics of the exogenous state $\mathbb{P}_x^h$ can differ across each stage of $\Omega(H^{3/2}d\sqrt{K})$. Since the upper bound in Section 4.2 of $\tilde{O}(H^{3/2}d\sqrt{K})$ applies to this more general setting, we achieve the minimax optimal rate for nonstationary Exo-MDPs up to polylogarithmic factors.

Theorem 4. *Assume $K \geq \frac{1}{10}d^2$, then for any Exo-MDP algorithm $\mathcal{B}$, there exists an Exo-MDP $\mathcal{M}$ such that the expected regret of $\mathcal{B}$ over K episodes on the Exo-MDP $\mathcal{M}$ is lower bounded by $\gamma Hd\sqrt{K}$ for some universal constant γ.*

Proof sketch. Our lower bound construction builds upon the hardness of learning a single-horizon Exo-MDP, which we call an Exo-Bandit. Specifically, we construct an Exo-Bandit instance which

[2] To the best of our knowledge, it has the best known guarantees.

reduces to learning a linear bandit on a hypercube action set that achieves a lower bound of $\Omega(d\sqrt{K})$. We then use this Exo-Bandit to construct a hard instance of Exo-MDP, denoted as $\mathcal{M}$.
At stage $h = 1$, $\mathcal{M}$ follows the same reward dynamics as the Exo-Bandit. For stages $h = 2, 3, \ldots, H$, the specific forms of $\mathbf{f}$ and $\mathbf{g}$ force the reward from the first stage to repeat H times regardless of the actions or exogenous states, without revealing any additional information on $\mathbb{P}_x$. This directly leads to a lower bound of $\Omega(Hd\sqrt{K})$. We outline the hard instance of Exo-MDP $\mathcal{M}[\mathbb{P}_x(\tilde{Z}), \mathbf{f}, \mathbf{g}] = (\mathcal{S} \times \mathcal{X}, \mathcal{A}, H, s_1, \mathbb{P}, R)$ below.

The state space of $\mathcal{M}$ is given by $\mathcal{S} = s_1 \cup \{(h, r) \mid h \in \{2, 3, \ldots, H\}, r \in \{-1, 1\}\}$. That is, $\mathcal{S}$ consists of s_1, a single starting state, and each of the next $H - 1$ states are indexed by the stage $h \in [H]$ as well as a single number $r \in \{-1, 1\}$. The exogenous state space is given by $\mathcal{X} = [d] = \{1, 2, \ldots, d\}$. The action set $\mathcal{A}$ sits on a subset of the d-dimensional hypercube, where

$$\mathcal{A} = \{([Z]_1, -[Z]_1, [Z]_2, -[Z]_2, \ldots, [Z]_{\frac{d}{2}}, -[Z]_{\frac{d}{2}}) \mid Z \in \{-1, 1\}^{d/2}\} \subset \{-1, 1\}^d.$$

Each action $a \in \mathcal{A}$ is completely characterized by a vector $Z \in \{-1, 1\}^{d/2}$ where $a(Z) = ([Z]_1, -[Z]_1, [Z]_2, -[Z]_2, \ldots, [Z]_{\frac{d}{2}}, -[Z]_{\frac{d}{2}})$. The (unknown) distribution $\mathbb{P}_x$ for the exogenous state X, parameterized by $\tilde{Z} \in \{-1, 1\}^{d/2}$ and constant $c = \frac{1}{10}\sqrt{\frac{2}{5K}}$, is given by

$$\mathbf{p}_x(\tilde{Z}) = (\mathbb{P}_x(1), \ldots, \mathbb{P}_x(d)) = (\frac{1}{d} + c[\tilde{Z}]_1, \frac{1}{d} - c[\tilde{Z}]_1, \ldots, \frac{1}{d} + c[\tilde{Z}]_{\frac{d}{2}}, \frac{1}{d} - c[\tilde{Z}]_{\frac{d}{2}}).$$

In other words, $\mathbf{p}_x$ is almost a uniform distribution except each coordinate is perturbed from $\frac{1}{d}$ by a small constant c or $-c$ depending on the value of $\tilde{Z}$. Intuitively, the hardness comes from correctly guessing the coordinates of these small perturbations by choosing action $a(Z)$ that matches $\tilde{Z}$ closely.

The known state transition function is given by

$$s_{h+1} = \mathbf{f}(s_h, a_h, x_h) = \begin{cases} (h+1, r) & \text{if } s_h = (h, r), h = 2, 3, \ldots, H-1 \\ (2, r = [a_1]_{x_1}) & \text{if } h = 1, s_h = s_1 \end{cases}$$

The action a_h has no effect on the state transition, except, in the first stage, action a_1 assigns value $r = [a_1]_{x_1}$ to the second coordinate of the state, which is then retained and shared across all stages afterwards. The known reward function is given by

$$R_h = \mathbf{g}(s_h, a_h, x_h) = \begin{cases} [a_1]_{x_1} & \text{if } h = 1, s_h = s_1 \\ r & \text{if } s_h = (h, r). \end{cases}$$

At stage $h = 1$, taking action a_1 incurs reward $[a_1]_{x_1}$, where $x_1 \sim \mathbb{P}_x$. For all $H - 1$ stages afterwards, the same reward at the first stage is repeated, leading to a total reward of $H \cdot [a_1]_{x_1}$.

6 Simulations on Inventory Control

We compare the empirical performance of the PLUG-IN (Section 4.1) and UCRL-VTR$^+$ (Section 4.2) algorithms on an extension to the inventory control example in Section 2. We show that under certain parameter settings, UCRL-VTR$^+$ achieves comparable performance with state-of-the-art algorithms tailored for inventory control despite being a more general solver.

Inventory control with lead time. In our experiments, we consider the online inventory control problem from Section 2 with the addition of a *lead time* L [9]. Suppose that instead of each order arriving immediately, orders take L timesteps to arrive. At the beginning of each stage h, the retailer observes the current inventory level Inv_h as well as the L previous orders that have not yet arrived, denoted $O_{h-L}, \ldots, O_{h-1}$. At each stage h, the order O_{h-L} arrives and the final on-hand inventory becomes $(\text{Inv}_h + O_{h-L} - X_h)^+$, where X_h denotes the independent exogenous demand drawn from $\mathbb{P}_x$. The cost consists of the holding cost for remaining products $c(\text{Inv}_h + O_{h-L} - X_h)^+$, plus the penalty for lost sales $p(X_h - \text{Inv}_h - O_{h-L})^+$. This can also be formulated as an Exo-MDP, where the state space scales exponentially with the lead time L.

We focus on the performances of UCRL-VTR$^+$ and the PLUG-IN approach. A direct application of Theorem 2 and Theorem 3 yields regret guarantees of $\tilde{O}(H^{3/2}\sqrt{dK})$ and $\tilde{O}(H^{3/2}d\sqrt{K})$ respectively, where d is the support of the demand distribution. We grant the PLUG-IN method additional access

Table 1. Performance of baselines at the final episode $K = 1000$, ($V_1^{\pi^K}$). $\star$ indicates significant improvement and $\circ$ significant decrease over ONLINE BASE-STOCK by Welch's t-test with a p value of 0.05. In parenthesis we show the relative performance to the cost of the optimal policy, $(V_1^{\pi^K} - V_1^*)/V_1^*$.

Algorithm	Scenario I	Scenario II
Optimal Policy (V_1^*)	41.8 (0%)	33.0 (0%)
Optimal Base-Stock Policy ($V_1^{b^*}$)	79.0 (89%)	33.0 (0%)
PLUG-IN	$\mathbf{41.8}^\star$ (1%)	$\mathbf{33.6}^\star$ (2%)
RANDOM	87.8° (110%)	76.6° (132%)
ONLINE BASE-STOCK	81.5 (95%)	40.6 (23%)
UCRL-VTR$^+$	$\mathbf{42.3}^\star$ (1%)	38.4 (16%)

to past demand trajectories for comparison, even though X_h is unobserved since the algorithms only observe the sales of $\min\{X_h, \text{Inv}_h + O_{h-L}\}$.

Baseline algorithms. We compare the performance of our algorithms against a widely-used heuristic, *base-stock policies* [9]. These policies are defined relative to b, the so-called base-stock level. At each stage h, the policy orders an amount to ensure the total inventory position (including both on-hand inventory and outstanding orders) is at least b units. In Table 1 we use $V_1^{b^*}$ to denote the performance of the best base-stock policy. To learn the optimal base-stock policy online, inspired by Agrawal and Jia [1] we include an online convex optimization algorithm over the base-stock level, which we denote as ONLINE BASE-STOCK. This approach yields a regret guarantee of $O(H\sqrt{K})$, where regret is defined relative to the optimal base-stock $V_1^{b^*}$. At first glance this guarantee seems stronger (scaling independent of d). However, since base-stock policies are not optimal in general, even the optimal base-stock policy can lead to regret of $\Omega(K)$ relative to the performance of the optimal policy.

Simulation results. In Table 1 we compare the performance of our algorithms. Under Scenario I we note a large optimality gap between the optimal (V_1^*) and best performing base-stock policy ($V_1^{b^*}$). While UCRL-VTR$^+$ achieves a *slower sample complexity rate*, it outperforms ONLINE BASE-STOCK since it surpasses the sub-optimality of the heuristic class of base-stock and instead converges to the true optimal policy. Moreover, UCRL-VTR$^+$ and the PLUG-IN algorithm with *idealized observations* both converge quickly to the true optimal policy V_1^*. Under Scenario II, the best performing base-stock policy ($V_1^{b^*}$) is the true optimal (V_1^*). We again observe that UCRL-VTR$^+$ achieves similar performance to ONLINE BASE-STOCK. This highlights the robustness of UCRL-VTR$^+$ to different regimes, achieving convergence to the true optimal policy even in settings where optimal base-stock is sub-optimal, and additionally achieves similar statistical power as ONLINE BASE-STOCK.

7 Conclusion and Future Work

In this paper we study a special class of Markov decision processes called Markov decision processes with exogenous states (Exo-MDPs), which arise from real-world MDPs where some state variables are exogenous and outside of the control of the decision maker. We highlight that Exo-MDPs, despite their structural assumptions, represent a rich class of MDPs equivalent to both the class of discrete MDPs and discrete linear mixture MDPs. We provide algorithms under both the full observation and no observation regimes on the exogenous states, as well as nearly-matching lower bounds. Importantly, all our algorithms achieve regret bounds scaling only with the dimension of the exogenous state d regardless of the endogenous state and action spaces. One interesting open direction would be to investigate intermediate observation regimes with sample complexities interpolating between the full and no observation regimes. Another future direction is to study the upper bounds outside the high dimension ($d \geq H$) and large sample ($K \geq d^4H + d^3H^2$) regime required for the upper bounds.

References

[1] Shipra Agrawal and Randy Jia. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Operations Research*, 70(3):1646–1664, 2022.

[2] Matias Alvo, Daniel Russo, and Yash Kanoria. Neural inventory control in networks via hindsight differentiable policy optimization. *arXiv preprint arXiv:2306.11246*, 2023.

[3] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[4] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ayoub20a.html.

[5] Jim G Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 2021.

[6] Thomas Dietterich, George Trimponias, and Zhitang Chen. Discovering and Removing Exogenous State Variables and Rewards for Reinforcement Learning. In *ICML*, 2018.

[7] Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, pages 5062–5127. PMLR, 2022.

[8] Jiekun Feng, Mark Gluzman, and Jim G Dai. Scalable deep reinforcement learning for ride-hailing. In *2021 American Control Conference (ACC)*, pages 3743–3748. IEEE, 2021.

[9] David A. Goldberg, Martin I. Reiman, and Qiong Wang. A Survey of Recent Progress in the Asymptotic Analysis of Inventory Systems. *Production and Operations Management*, 30(6), 2021.

[10] Ori Hadary, Luke Marshall, Ishai Menache, Abhisek Pan, Esaias E Greeff, David Dion, Star Dorminey, Shailesh Joshi, Yang Chen, Mark Russinovich, et al. Protean:{VM} allocation service at scale. In *14th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 20)*, pages 845–861, 2020.

[11] Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 666–686. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/jia20a.html.

[12] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[13] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[14] Dhruv Madeka, Kari Torkkola, Carson Eisenach, Anna Luo, Dean P Foster, and Sham M Kakade. Deep inventory management. *arXiv preprint arXiv:2210.03137*, 2022.

[15] Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. *arXiv preprint arXiv:1807.02264*, 2018.

[16] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13:1, 2022.

[17] W.B. Powell. *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. Wiley, 2022. ISBN 9781119815037. URL https://books.google.com/books?id=y8V6EAAAQBAJ.

[18] Ashwin Rao and Tikhon Jelvis. *Foundations of reinforcement learning with applications in finance*. Chapman and Hall/CRC, 2022.

[19] Benjamin Rolf, Ilya Jackson, Marcel Müller, Sebastian Lang, Tobias Reggelin, and Dmitry Ivanov. A review on reinforcement learning algorithms and applications in supply chain management. *International Journal of Production Research*, 61(20):7151–7179, 2023.

[20] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[21] Sean R Sinclair, Felipe Vieira Frujeri, Ching-An Cheng, Luke Marshall, Hugo De Oliveira Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, and Adith Swaminathan. Hindsight learning for mdps with exogenous inputs. In *International Conference on Machine Learning*, pages 31877–31914. PMLR, 2023.

[22] Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin. " deep reinforcement learning for search, recommendation, and online advertising: a survey" by xiangyu zhao, long xia, jiliang tang, and dawei yin with martin vesely as coordinator. *ACM sigweb newsletter*, 2019(Spring): 1–15, 2019.

[23] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.