Proceedings of Machine Learning Research vol 178:1–66, 2022 35th Annual Conference on Learning Theory

Sample-Efficient Reinforcement Learning in the Presence of Exogenous Information

Yonathan Efroni *Microsoft Research, NYC*

Dylan J. Foster *Microsoft Research, NYC*

Dipendra Misra *Microsoft Research, NYC*

Akshay Krishnamurthy *Microsoft Research, NYC*

John Langford *Microsoft Research, NYC*

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

In real-world reinforcement learning applications the learner's observation space is ubiquitously high-dimensional with both relevant and irrelevant information about the task at hand. Learning from high-dimensional observations has been the subject of extensive investigation in supervised learning and statistics (e.g., via sparsity), but analogous issues in reinforcement learning are not well understood, even in finite state/action (tabular) domains. We introduce a new problem setting for reinforcement learning, the Exogenous Markov Decision Process (ExoMDP), in which the state space admits an (unknown) factorization into a small controllable (or, *endogenous*) component and a large irrelevant (or, *exogenous*) component; the exogenous component is independent of the learner's actions, but evolves in an arbitrary, temporally correlated fashion. We provide a new algorithm, ExoRL, which learns a near-optimal policy with sample complexity polynomial in the size of the endogenous component and nearly independent of the size of the exogenous component, thereby offering a *doubly-exponential* improvement over off-the-shelf algorithms. Our results highlight for the first time that sample-efficient reinforcement learning is possible in the presence of exogenous information, and provide a simple, user-friendly benchmark for investigation going forward.

1. Introduction

Most applications of machine learning and statistics involve complex inputs such as images or text, which may contain spurious information for the task at hand. A traditional approach to this problem is to use feature engineering to identify relevant information, but this requires significant domain expertise, and can lead to poor performance if relevant information is missed. As an alternative, representation learning and feature selection methodologies developed over the last several decades address these issues, and enable practitioners to directly operate on complex, high-dimensional inputs with minimal domain knowledge. In the context of supervised learning and statistical estimation, these methods are particularly well-understood (Hastie et al., 2015; Wainwright, 2019) and—in some cases—can be shown to provably identify relevant information for the task at hand in the presence of a vast amount of irrelevant or spurious features. As such, these approaches have emerged as the methods of choice for many practitioners.

Complex, high-dimensional inputs are also ubiquitous in Reinforcement Learning (RL) applications. However, due to the interactive, multi-step nature of the RL problem, naive extensions of representation learning techniques from supervised learning do not seem adequate. Empirically, this can be seen in the brittleness of deep RL algorithms and, the large body of work on stabilizing these methods (Gelada et al., 2019; Zhang et al., 2020). Theoretically, this can be seen by the prevalence of strong function approximation assumptions that preclude introducing spurious features (Wang et al., 2021; Weisz et al., 2021). As a result, developing representation learning methodology for RL is a central topic of investigation.

Recently, a line of theoretical works have developed structural conditions under which RL with complex inputs is statistically tractable (Jiang et al., 2017; Jin et al., 2021; Du et al., 2021; Foster et al., 2021), along with a complementary set of algorithms for addressing these problems via representation learning (Du et al., 2019; Misra et al., 2020; Agarwal et al., 2020; Misra et al., 2021; Uehara et al., 2021). While these works provide some clarity into the challenges of high-dimensionality in RL, the models considered do not allow for spurious, temporally correlated information (e.g., exogenous information that evolves over time through a complex dynamical system). On the other hand, this structure is common in applications; for example, when a human is navigating a forest trail, the flight of birds in the sky is temporally correlated, but irrelevant for the human's decision making. Motivated by the success of high-dimensional statistics in developing and understanding feature selection methods for supervised learning, we ask:

Can we develop provably efficient algorithms for RL in the presence of a large number of dynamic, yet irrelevant features?

Efroni et al. (2021b) initiated the study of this question in a rich-observation setting with function approximation. However, their results require deterministic dynamics, and their approach crucially uses determinism to sidestep many challenges that arise in the presence of exogenous information.

Our contributions. In this paper, we take a step back from the function approximation setting considered by Efroni et al. (2021b), and introduce a simplified problem setting in which to study representation learning and exploration with high-dimensional, exogenous information. Our model, the *Exogenous Markov Decision Process* or ExoMDP, involves a discrete d-dimensional state space (with each dimension taking values in $\{1, \dots, S\}$) in which an *unknown* subset of $k \ll d$ dimensions of the state can be controlled by the agent's actions. The remaining $d - k$ state variables are irrelevant for the agent's task, but may exhibit complex temporal structure.

Our main result is a new algorithm, ExoRL, that learns a policy which is (i) near-optimal and (ii) does not depend on the exogenous and irrelevant factors, while requiring only $\mathrm{poly}(S^k, \log(d))$ trajectories. Here, the dominant S^k term represents the size of the controllable (or, endogenous) state space, and the $\log(d)$ term represents the price incurred for feature selection (analogous to guarantees for sparse regression (Hastie et al., 2015; Wainwright, 2019)). Our result represents a *doubly-exponential* improvement over naive application of existing tabular RL methods to the ExoMDP setting, which results in $\mathrm{poly}(S^d)$ sample complexity. Our algorithm and analysis involve many new ideas for addressing exogenous noise, and we believe our work may serve as a building block for addressing these issues in more practical settings.

2. Overview of Results

In this section we introduce the ExoMDP setting and give an overview of our algorithmic results, highlighting the key challenges they overcome. Before proceeding, we formally describe the basic RL setup we consider.

Markov decision processes. We consider a finite-horizon Markov decision process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, H, d_1)$, in which $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $H \in \mathbb{N}$ is the horizon, and $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution. Given a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$, where $\pi_h : \mathcal{S} \to \mathcal{A}$, an episode in the MDP $\mathcal{M}$ proceeds as follows, beginning from $s_1 \sim d_1$: For $h = 1, \ldots, H$: $a_h = \pi_h(s_h)$, $r_h = R(s_h, a_h)$, and $s_{h+1} \sim T(\cdot \mid s_h, a_h)$. We let $\mathbb{E}_\pi[\cdot]$ and $\mathbb{P}_\pi(\cdot)$ denote the expectation and probability for the trajectory $(s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)$ when π is executed, respectively, and define $J(\pi) = \mathbb{E}_\pi\left[\sum_{h=1}^{H} r_h\right]$ as the average reward.

The objective of the learner is to learn an ϵ-optimal policy online: Given N episodes to execute a policy and observe the resulting trajectory, find a policy $\widehat{\pi}$ such that $J(\widehat{\pi}) \geq \max_{\pi \in \Pi_{\mathrm{NS}}} J(\pi) - \epsilon$, where Π_{NS} denotes the set of all non-stationary policies $\pi = (\pi_1, \ldots, \pi_H)$.

2.1. The Exogenous MDP (ExoMDP) Setting

The ExoMDP is a Markov decision process in which the state space factorizes into an endogenous component that is (potentially) affected by the learner's actions, and an exogenous component that is independent of the learner's actions, but evolves in an arbitrary, temporally correlated fashion. Formally, given a parameter $d \in \mathbb{N}$ (the number of factors), the state space $\mathcal{S}$ takes the form $\mathcal{S} = \otimes_{i=1}^{d} \mathcal{S}_i$, so that each state $s \in \mathcal{S}$ has the form $s = (s_1, \ldots, s_d)$, with $s_i \in \mathcal{S}_i$; we refer to $\mathcal{S}_i$ (equivalently, i) as the i^{th} *factor*. We take $\mathcal{I}_\star \subset [d]$ to represent the *endogenous factors* and $\mathcal{I}_\star^{\mathrm{c}} := [d] \setminus \mathcal{I}_\star$ to represent the *exogenous factors*, which are unknown to the learner. Letting $s[\mathcal{I}] := (s_i)_{i \in \mathcal{I}}$, we assume the dynamics and rewards factorize across the endogenous and exogenous components as follows:

$$
\begin{aligned}
T\left(s' \mid s, a\right) &= T_{\mathrm{en}}(s'\left[\mathcal{I}_\star\right] \mid s\left[\mathcal{I}_\star\right], a) \cdot T_{\mathrm{ex}}(s'\left[\mathcal{I}_\star^{\mathrm{c}}\right] \mid s\left[\mathcal{I}_\star^{\mathrm{c}}\right]), \\
R(s, a) &= R_{\mathrm{en}}(s\left[\mathcal{I}_\star\right], a), \\
d_1(s) &= d_{1,\mathrm{en}}\left(s\left[\mathcal{I}_\star\right]\right) \cdot d_{1,\mathrm{ex}}\left(s\left[\mathcal{I}_\star^{\mathrm{c}}\right]\right),
\end{aligned}
\tag{1}
$$

for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. That is, the endogenous factors $\mathcal{I}_\star$ are (potentially) affected by the agent's actions and are sufficient to model the reward, while the exogenous factors $\mathcal{I}_\star^{\mathrm{c}}$ evolve independently of the learner's actions and do not influence the reward.

In this paper, we focus on a finite-state/action (tabular) variant of the ExoMDP setting in which $\mathcal{S}_i = [S]$ and $\mathcal{A} = [A]$, with $S \in \mathbb{N}$ representing the number of states per factor and $A \in \mathbb{N}$ representing the number of actions. We assume that $|\mathcal{I}_\star| \leq k$, where $k \ll d$ is a known upper bound on the number of endogenous factors.[1] In the absence of the structure in Eq. (1), this is a generic tabular RL problem with $|\mathcal{S}| = S^d$, and the optimal sample complexity scales as $\mathrm{poly}(S^d, A, H, \epsilon^{-1})$ (Azar et al., 2017), which has exponential dependence on the number of factors d. On the other hand, if $\mathcal{I}_\star$ were known a-priori, applying off-the-shelf algorithms for tabular RL to the endogenous subset of the state space would lead to sample complexity $\mathrm{poly}(S^k, A, H, \epsilon^{-1})$ (Azar et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019; Kaufmann et al., 2021), which is independent of d and offers significant improvement when $k \ll d$. This motivates us to ask: *With no prior knowledge, can we learn an ϵ-optimal policy for the ExoMDP with sample complexity polynomial in S^k and sublinear in d?*

1. Extending our results to settings in which different factors have different sizes (i.e., $\mathcal{S}_i = [S_i]$) is straightforward.

2.2. Challenges of RL in the Presence of Exogenous Information

Sample-efficient learning in the absence of prior knowledge poses significant algorithmic challenges.

(C1) *Hardness of identifying endogenous factors.* In general, the endogenous factors may not be identifiable (that is, multiple choices for $\mathcal{I}_\star$ may obey the structure in Eq. (1)). Even when $\mathcal{I}_\star$ is identifiable, *certifying* whether a particular factor $i \in [d]$ is exogenous can be statistically intractable (e.g., if the effect of the agent's action on the state component s_i is small relative to ϵ).

(C2) *Necessity of exploration.* The agent's action might have a large effect on an endogenous factor $i \in \mathcal{I}_\star$, but only in a particular state $s \in \mathcal{S}$ that requires deliberate planning to reach. As such, any approach that attempts to recover the endogenous factors must be interleaved with exploration, resulting in a chicken-and-egg problem. "Test-then-explore" approaches do not suffice.

(C3) *Entanglement of endogenous and exogenous factors.* The factorized dynamics in (1) lead to a number of useful structural properties for ExoMDPs, such as factorization of state occupancy measures (cf. Appendix B). However, these properties generally only hold for policies that act on the endogenous portion of the state. When an agent executes a policy whose actions depend on the exogenous state factors, the evolution of the endogenous and exogenous components becomes entangled. This entanglement makes it difficult to apply supervised learning or estimation methods to extract information from trajectories gathered from such policies, and can lead to error amplification. As a result, significant care is required in gathering data.

Failure of existing algorithms. Existing RL techniques do not appear to be sufficient to address the challenges above and generally have sample complexity requirements scaling with $\Omega(d)$ or worse. For example, tabular methods do not exploit factored structure, resulting in $\Omega(S^d)$ sample complexity, and we can show that complexity measures like the Bellman rank (Jiang et al., 2017) and its variants scale as $\Omega(d)$, so they do not lead to sample-efficient learning guarantees. Moreover, algorithms for factored MDPs (e.g., Rosenberg and Mansour (2020)) obtain guarantees that depend on sparsity in the transition operator, but this operator is dense in the ExoMDP setting, leading to sample complexity that is exponential in d. See further discussion in Section 5 and Appendix B.1.

2.3. Main Result

We present a new algorithm, ExoRL, which learns a near-optimal policy for the ExoMDP with sample complexity polynomial in the number of endogenous states and *logarithmic* in the number of exogenous components. Following previous approaches to representation learning in RL (Du et al., 2019; Misra et al., 2020; Agarwal et al., 2020), our results depend on a *reachability parameter*.

Definition 2.1. *The endogenous state space is η-reachable if for all $h \in [H]$ and $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$, either*

$$\max_{\pi \in \Pi_{\mathrm{NS}}} \mathbb{P}_\pi\left(s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) \geq \eta, \quad \text{or} \quad \max_{\pi \in \Pi_{\mathrm{NS}}} \mathbb{P}_\pi\left(s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) = 0.$$

Crucially, this notation of reachability considers only the endogenous portion of the state space, not the full state space. We assume access to a lower bound η on the optimal reachability parameter.

Our main result is as follows.

Theorem 4.1 (informal). *With high probability, ExoRL learns an ϵ-optimal policy for the ExoMDP using* $\mathrm{poly}(S^k, A, H, \log(d)) \cdot \left(\epsilon^{-2} + \eta^{-2}\right)$ *trajectories.*

This constitutes a *doubly-exponential* improvement over the S^d sample complexity for naive tabular RL in terms of dependence on the number of factors d, and it provides a RL analogue of sparsity-dependent guarantees in high-dimensional statistics (Hastie et al., 2015; Wainwright, 2019). Importantly, the result does not require any statistical assumptions beyond the factored structure in Eq. (1) and reachability (for example, we do not require deterministic dynamics). Beyond polynomial factors, the dependence on the size of the state space cannot be improved further.

2.4. Our Approach: Exploration with a Certifiably Endogenous Policy Cover

ExoRL is built upon the notion of an *endogenous policy cover*. Define an endogenous policy as follows.

Definition 2.2 (Endogenous policy). *A policy $\pi = (\pi_1, \ldots, \pi_H)$ is* endogenous *if it acts only on the endogenous component of the state space: For all $h \in [H]$ and $s \in \mathcal{S}$, we have $\pi_h(s) = \pi_h(s[\mathcal{I}_\star])$.*

An endogenous policy cover is a (small) collection of endogenous policies that ensure each state is reached with near-maximal probability.

Definition 2.3 (Endogenous policy cover). *A set of non-stationary policies Ψ is an endogenous (ϵ-approximate) policy cover for timestep h if:*

1. *For all $s \in \mathcal{S}$, $\max_{\psi\in\Psi} \mathbb{P}_\psi\left(s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) \geq \max_{\pi\in\Pi_{\mathrm{NS}}} \mathbb{P}_\pi\left(s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) - \epsilon$.*
2. *The set Ψ contains only endogenous policies.*

While the coverage property of Definition 2.3 is stated in terms of occupancy measures for the endogenous portion of the state space, the factored structure of the ExoMDP implies that this yields a cover for the entire state space (cf. Appendix B.2):

$$\max_{\psi\in\Psi} \mathbb{P}_\psi\left(s_h = s\right) \geq \max_{\pi} \mathbb{P}_\pi\left(s_h = s\right) - \epsilon, \quad \forall s \in \mathcal{S}.$$

In particular, even though $|\mathcal{S}| = S^d$, this guarantees that for each timestep h, there exists a *small* endogenous policy cover with $|\Psi| \leq S^k$. ExoRL constructs such a policy cover and uses it for sample-efficient exploration in two phases. First, in Phase I (OSSR), the algorithm builds the policy cover in a manner guaranteeing endogeneity; this accounts for the majority of the algorithm design and analysis effort. Then, in Phase II (ExoPSDP), the algorithm uses the policy cover to optimize rewards.

Finding a certifiably endogenous policy cover: OSSR. The main component of ExoRL is an algorithm, OSSR, which iteratively learns a sequence of endogenous policy covers $\Psi^{(1)}, \ldots, \Psi^{(H)}$ with

$$\max_{\psi\in\Psi^{(h)}} \mathbb{P}_\psi\left(s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) \geq \max_{\pi} \mathbb{P}_\pi\left(s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) - \epsilon$$

for all $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$. For each $h \in [H]$, given the policy covers $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$ for preceding timesteps, OSSR builds the policy cover $\Psi^{(h)}$ using a novel statistical test. The test constructs a factor set $\mathcal{I} \subset [d]$ which is (i) endogenous, in the sense that $\mathcal{I} \subset \mathcal{I}_\star$, yet (ii) ensures sufficient coverage, in the sense that there exists a near-optimal policy cover operating only on $s[\mathcal{I}]$. The analysis of this test relies on a unique structural property of the ExoMDP setting called the *restriction lemma* (Lemma B.2), which provides a mechanism to "regularize" the factor set under consideration toward endogeneity in a data-driven fashion.

This approach circumvents challenges (C1) and (C2): It does not rely on explicit identification of the endogenous factors and instead iteratively builds a *subset* of factors that is certifiably endogenous, but nonetheless sufficient to explore. Endogeneity of the resulting policy cover $\Psi^{(h)}$ ensures the success of subsequent tests at rounds $h+1,\dots,H$, and circumvents the issue of entanglement raised in challenge (C3). To summarize, the following guarantee constitutes our main technical result.

Theorem 3.1 (informal). *With high probability, OSSR finds an endogenous $\frac{\eta}{2}$-approximate policy cover using $\mathrm{poly}\big(S^k, A, H, \log(d)\big)\cdot\eta^{-2}$ trajectories.*

2.5. Organization

The remainder of the paper is organized as follows. In Section 3, we introduce the OSSR algorithm, highlight the key algorithm design techniques and analysis ideas, and state its formal guarantee (Theorem 3.1) for finding a policy cover. Building on this result, in Section 4 we introduce the ExoRL algorithm, and provide the main sample complexity guarantee for RL in ExoMDPs (Theorem 4.1). We close with discussion of additional related work (Section 5) and open problems (Section 6).

2.6. Preliminaries

We let Π denote the set of all one-step policies $\pi:\mathcal{S}\to\mathcal{A}$. We use the term $t\to h$ *policy* to refer to a non-stationary policy $\pi=(\pi_t,\dots,\pi_h)$ defined over a subset of timesteps $t\le h$.

For a non-stationary policy $\pi\in\Pi_{\mathrm{NS}}$, we define the state-action and state value functions: $Q^\pi_h(s,a) := \mathbb{E}_\pi\left[\sum_{h'=h}^H r_{h'} \mid s_h=s, a_h=a\right]$, and $V^\pi_h(s) := Q^\pi_h(s,\pi_h(s))$. We denote the expected value of a policy π from time step t to h by $V_{t,h}(\pi) := \mathbb{E}_\pi\left[\sum_{t'=t}^h r_{t'}\right]$. We adopt the shorthand $d_h(s\,;\,\pi) := \mathbb{P}_\pi(s_h=s)$ for the induced state occupancy measure. Likewise, for $\mathcal{I}\subseteq[d]$, we define $d_h(s[\mathcal{I}]\,;\,\pi) := \mathbb{P}_\pi(s_h[\mathcal{I}]=s[\mathcal{I}])$.

For algorithm design purposes, we consider *mixture policies* of the form $\mu\in\Pi_{\mathrm{mix}} := \Delta(\Pi_{\mathrm{NS}})$. To run a mixture policy $\mu\in\Pi_{\mathrm{mix}}$, we sample $\pi\sim\mu$, then execute π for an entire episode. We further denote $\Pi_{\mathrm{mix}}[\mathcal{I}] := \Delta(\Pi_{\mathrm{NS}}[\mathcal{I}])$ as the set of mixture policies over the policy set $\Pi_{\mathrm{NS}}[\mathcal{I}]$, where $\Pi_{\mathrm{NS}}[\mathcal{I}]$ denotes the set of policies that act on the factor set $\mathcal{I}$. We let $\mathbb{E}_\mu[\cdot]$ and $\mathbb{P}_\mu(\cdot)$ denote the expectation and probability under this process, and we define $J(\mu)=\mathbb{E}_{\pi\sim\mu}[J(\pi)]=\mathbb{E}_\mu\left[\sum_{h=1}^H r_h\right]$ and $d_h(s\,;\,\mu) := \mathbb{P}_\mu(s_h=s)$ analogously. We say that $\mu\in\Pi_{\mathrm{mix}}$ is endogenous if it is supported over endogenous policies in Π_{NS}. Finally, for $\mu\in\Pi_{\mathrm{mix}}$ and $\pi\in\Pi$ we let $\mu\circ_t\pi$ be the policy that follows μ for the first $t-1$ timesteps, and at the t^{th} timestep it switches to π. For sets of policies Ψ_1 and Ψ_2 we let $\Psi_1\circ_t\Psi_2 := \{\psi_1\circ_t\psi_2 \mid \psi_1\in\Psi_1,\psi_2\in\Psi_2\}$.

ExoMDP notation. Recall that for a factor set $\mathcal{I}\subseteq[d]$, we define $\mathcal{I}^{\mathrm{c}} := [d]\setminus\mathcal{I}$ as the complement, and define $s[\mathcal{I}] := (s_i)_{i\in\mathcal{I}}$ and $\mathcal{S}[\mathcal{I}] := \otimes_{i\in\mathcal{I}}\mathcal{S}_i$ as the corresponding components of the state and state space. We make frequent use of the fact that for any pair of factors $\mathcal{I}_1$ and $\mathcal{I}_2$ with $\mathcal{I}=\mathcal{I}_1\cup\mathcal{I}_2$ and $\mathcal{I}_1\cap\mathcal{I}_2=\varnothing$, any state $s[\mathcal{I}]\in\mathcal{S}[\mathcal{I}]$ can be uniquely split as $s[\mathcal{I}]=(s[\mathcal{I}_1],s[\mathcal{I}_2])$, with $s[\mathcal{I}_1]\in\mathcal{S}[\mathcal{I}_1]$ and $s[\mathcal{I}_2]\in\mathcal{S}[\mathcal{I}_2]$. We use a canonical ordering when indexing with factor sets.

Any factor set $\mathcal{I}\subseteq[d]$ can be written as $\mathcal{I}=(\mathcal{I}\cap\mathcal{I}_\star)\cup(\mathcal{I}\cap\mathcal{I}_\star^{\mathrm{c}})$. We denote these intersections by $\mathcal{I}_{\mathrm{en}} := \mathcal{I}\cap\mathcal{I}_\star$ and $\mathcal{I}_{\mathrm{ex}} := \mathcal{I}\cap\mathcal{I}_\star^{\mathrm{c}}$, which represent the endogenous and exogenous components of $\mathcal{I}$.

We say that a policy π *acts on a factor set* $\mathcal{I}$ if it selects actions as a measurable function of $\mathcal{S}[\mathcal{I}]$. We let $\Pi[\mathcal{I}]$ denote the set of all one-step policies $\pi:\mathcal{S}[\mathcal{I}]\to\mathcal{A}$ that act on $\mathcal{I}$, and let $\Pi_{\mathrm{NS}}[\mathcal{I}]$ denote the set of all non-stationary policies that act on $\mathcal{I}$.

Lastly, if $\mathcal{I} \subseteq \mathcal{I}_\star^c$, i.e., the factor $\mathcal{I}$ is a subset of the exogenous factors, we omit the dependence in the policy π from its occupancy measure, $d_h(s[\mathcal{I}] \,;\, \pi) = d_h(s[\mathcal{I}])$. Indeed, for any $\pi, \pi' \in \Pi_{\mathrm{NS}}$ it holds that $d_h(s[\mathcal{I}] \,;\, \pi) = d_h(s[\mathcal{I}] \,;\, \pi')$, and hence the occupancy measure of $s[\mathcal{I}]$ is independent of the policy.

Collections of factor sets. For a factor set $\mathcal{I} \subseteq [d]$, we let $\mathscr{I}_{\leq k}(\mathcal{I}) := \{\mathcal{I}' \subseteq [d] \mid \mathcal{I} \subseteq \mathcal{I}',\ |\mathcal{I}'| \leq k\}$ denote a collection of all factor sets of size at most k that contain $\mathcal{I}$, and analogously define $\mathscr{I}_k(\mathcal{I}) := \{\mathcal{I}' \subseteq [d] \mid \mathcal{I} \subseteq \mathcal{I}',\ |\mathcal{I}'| = k\}$. We adopt the shorthand $\mathscr{I}_{\leq k} := \mathscr{I}_{\leq k}(\varnothing)$ and $\mathscr{I}_k := \mathscr{I}_k(\varnothing)$. With some abuse of notation, for a given collection of factor sets $\mathscr{I}$, we define $\Pi[\mathscr{I}] := \cup_{\mathcal{I} \in \mathscr{I}} \Pi[\mathcal{I}]$ as the set of all possible policies induced by factors in $\mathscr{I}$.

We define $[N] := \{1, 2, \cdots, N\}$. $\mathrm{Unf}(\mathcal{X})$ denotes the uniform distribution over a finite set $\mathcal{X}$.

3. Learning a Near-Optimal Endogenous Policy Cover: OSSR

In this section, we present the first of our main algorithms, OSSR (Algorithm 8), which performs reward-free exploration to construct an endogenous policy cover for the ExoMDP. OSSR constitutes the main algorithmic component of ExoRL, and we believe it is of independent interest.

OSSR is a *forward-backward* algorithm. For each layer $h \in [H]$, given previous policy covers $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$, the algorithm constructs an endogenous policy cover $\Psi^{(h)}$ in a backwards fashion. Backward steps proceed from $t = h-1, \ldots, 1$, with each step consisting of (i) an *optimization* phase, in which we find a (potentially large) collection of policies for choosing actions at step t that lead to good coverage for all possible target factors sets $\mathcal{I}$ at layer h, and (ii) a *selection* phase, in which we narrow the collection of policies from the first phase down to a small set of policies that act on a single (endogenous) factor set $\mathcal{I}$, yet still ensure coverage for all states at step h.

Instead of directly diving into OSSR, we build up to the algorithm through two warm-up exercises:

- In Section 3.1, we consider a simplified version of OSSR (OSSR.OneStep, or Algorithm 1) which computes an endogenous policy cover under the assumption that (i) $H = 2$, and (ii) certain occupancy measures for the underlying ExoMDP can be computed exactly.
- Building on this result, in Section 3.2 we provide another simplified algorithm (OSSR.Exact, or Algorithm 2) which computes an endogenous policy cover for general H, but still requires exact access to certain occupancy measures for the ExoMDP.

Finally, in Section 3.3 we present the full OSSR algorithm and its main sample complexity guarantee.

3.1. Warm-Up I: Finding an Endogenous Policy Cover with Exact Queries ($H = 2$)

Algorithm 1 presents OSSR.OneStep, a simplified version of OSSR that computes a (small) endogenous policy cover for horizon two, assuming exact access to the state occupancies $d_2(s \,;\, \pi)$. This algorithm highlights the mechanism through which OSSR is able to simultaneously ensure both endogeneity and coverage.

OSSR.OneStep learns an endogenous policy cover in two phases. In the *optimization phase* (Lines 1 and 2) the algorithm computes a *partial policy cover* $\Gamma[\mathcal{J}]$ for each factor set $\mathcal{J} \in \mathscr{I}_{\leq k}$, which ensures that for all state factor values $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$ there exists a policy $\pi_{s[\mathcal{J}]} \in \Gamma[\mathcal{J}]$ which maximizes the probability to reach the state factor value $s[\mathcal{J}]$ at the 2^{nd} timestep.

Algorithm 1 OSSR.OneStep: Optimization-Selection State Refinement for ExoMDPs with $H = 2$

Phase I: Optimization

1: Find factor set $\widetilde{\mathcal{I}} \in \mathscr{I}_{\leq k}$ with minimal cardinality such that for all $\mathcal{J} \in \mathscr{I}_{\leq k}$ and $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$,

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = \max_{\pi \in \Pi[\widetilde{\mathcal{I}}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right).$$

2: For all $\mathcal{J} \in \mathscr{I}_{\leq k}$, define $\pi_{s[\mathcal{J}]} = \arg\max_{\pi \in \Pi[\widetilde{\mathcal{I}}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right)$ for each $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$, then set

$$\Gamma\left[\mathcal{J}\right] := \left\{\pi_{s[\mathcal{J}]} : s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]\right\}.$$

Phase II: Selection

3: Find factor set $\widehat{\mathcal{I}} \in \mathscr{I}_{\leq k}$ with minimal cardinality such that for all $\mathcal{J} \in \mathscr{I}_{\leq k}$ and $s\left[\mathcal{J}\right] \in \mathcal{S}\left[\mathcal{J}\right]$,

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = d_2\left(s[\mathcal{J}]\ ;\ \pi_{s[\mathcal{J} \cap \widehat{\mathcal{I}}]}\right).$$

4: **return** $\left(\widehat{\mathcal{I}}, \Gamma[\widehat{\mathcal{I}}]\right)$

All of the partial policy covers are induced by a single factor set $\widetilde{\mathcal{I}}$; existence of such a factor set is guaranteed by Property 3.2. We show that by regularizing by cardinality, $\widetilde{\mathcal{I}}$ is guaranteed to be endogenous, and so the policy covers $(\Gamma[\mathcal{J}])_{\mathcal{J} \in \mathscr{I}_{\leq k}}$ are endogenous as well.

At this point, the only issue is size: The set $\bigcup_{\mathcal{J} \in \mathscr{I}_{\leq k}} \Gamma[\mathcal{J}]$ is an exact policy cover for $h = 2$ (in the sense of Definition 2.3), but its size scales as $\Omega(d^k)$,[2] which makes it unsuitable for exploration. To address this issue, the *selection phase* (Line 3) identifies a single endogenous factor $\widehat{\mathcal{I}}$ such that $\Gamma[\widehat{\mathcal{I}}]$ is an endogenous policy cover (note that choosing $\Gamma[\mathcal{I}_\star]$ would suffice, but $\mathcal{I}_\star$ is not known to the learner). Since $|\Gamma[\widehat{\mathcal{I}}]| \leq S^k$ by construction, this yields a small policy cover as desired.

Proposition 3.1. *The pair $\left(\widehat{\mathcal{I}}, \Gamma[\widehat{\mathcal{I}}]\right)$ returned by OSSR.OneStep has the property that (i) $\widehat{\mathcal{I}}$ is endogenous (i.e., $\widehat{\mathcal{I}} \subseteq \mathcal{I}_\star$), and (ii) $\Gamma[\widehat{\mathcal{I}}]$ is an endogenous policy cover for $h = 2$: For all $s \in \mathcal{S}$,*

$$\max_{\pi \in \Pi} d_2\big(s[\mathcal{I}_\star]\ ;\ \pi\big) = d_2\big(s[\mathcal{I}_\star]\ ;\ \pi_{s[\widehat{\mathcal{I}}]}\big), \quad \textit{where } \pi_{s[\widehat{\mathcal{I}}]} \in \Gamma[\widehat{\mathcal{I}}].$$

The ExoMDP transition structure further implies that $\max_{\pi \in \Pi} d_2\big(s\ ;\ \pi\big) = d_2\big(s\ ;\ \pi_{s[\widehat{\mathcal{I}}]}\big)\ \forall s \in \mathcal{S}$.

Proof of Proposition 3.1. We begin by highlighting two useful structural properties of the ExoMDP; both properties are specializations of more general results, Lemmas B.1 and B.2 (Appendix B).

Property 3.1 (Decoupling for endogenous policies)**.** *For any endogenous policy π, we have $d_2\left(s[\mathcal{I}]\ ;\ \pi\right) = d_2\left(s[\mathcal{I}_{\mathrm{en}}]\ ;\ \pi\right) \cdot d_2\left(s[\mathcal{I}_{\mathrm{ex}}]\right)$, for all $\mathcal{I} \subseteq [d]$ and $s \in \mathcal{S}$.*

Property 3.2 (Restriction lemma)**.** *For all factor sets $\mathcal{I}$ and $\mathcal{J}$, we have*

$$\max_{\pi \in \Pi[\mathcal{I}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = \max_{\pi \in \Pi[\mathcal{I}_{\mathrm{en}}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) \quad \forall s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]. \tag{2}$$

2. The set $\Pi[\widetilde{\mathcal{I}}]$ also gives a policy cover, but it is even larger.

Property 3.2 is perhaps the most critical structural result used by our algorithms. It implies that $\max_{\pi\in\Pi} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = \max_{\pi\in\Pi[\mathcal{I}_\star]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right)$, which in turn implies that the optimization and selection phases of Algorithm 1 are feasible (since we can show that $\mathcal{I}_\star$ is a valid choice). If $\widetilde{\mathcal{I}}$ and $\widehat{\mathcal{I}}$ are endogenous, then since $\widehat{\mathcal{I}} \subset \mathcal{I}_\star$ the selection rule ensures that $\Gamma[\widehat{\mathcal{I}}]$ is a policy cover for $\mathcal{S}[\mathcal{I}_\star]$ (by choosing $\mathcal{J} = \mathcal{I}_\star$ in Line 3 and since $\widehat{\mathcal{I}} \cap \mathcal{I}_\star = \widehat{\mathcal{I}}$). We next show that both $\widetilde{\mathcal{I}}$ and $\widehat{\mathcal{I}}$ are endogenous.

Claim 1: $\widetilde{\mathcal{I}}$ is endogenous. Observe that for any (potentially non-endogenous) factor set $\widetilde{\mathcal{I}} = \widetilde{\mathcal{I}}_{\text{en}} \cup \widetilde{\mathcal{I}}_{\text{ex}}$, Property 3.2 implies that for all $\mathcal{J} \in \mathscr{I}_{\leq k}$ and $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$,

$$\max_{\pi\in\Pi[\widetilde{\mathcal{I}}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = \max_{\pi\in\Pi[\widetilde{\mathcal{I}}_{\text{en}}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right),$$

For any factor set $\widetilde{\mathcal{I}}$ that satisfies the constraints in Line 1 but has $\widetilde{\mathcal{I}}_{\text{ex}} \neq \varnothing$, we can further reduce the cardinality without violating the constraints, so the minimum cardinality solution is endogenous.

Claim 2: $\widehat{\mathcal{I}}$ is endogenous. Consider a (potentially non-endogenous) factor set $\widehat{\mathcal{I}} = \widehat{\mathcal{I}}_{\text{en}} \cup \widehat{\mathcal{I}}_{\text{ex}}$. If $\widehat{\mathcal{I}}$ satisfies the constraint in Line 3, then for all $\mathcal{J} \in \mathscr{I}_{\leq k}$ and $s \in \mathcal{S}$, since $\mathcal{J}_{\text{en}} = \mathcal{J} \cap \mathcal{I}_\star \in \mathscr{I}_{\leq k}$,

$$\max_{\pi\in\Pi[\mathscr{I}_{\leq k}]} d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi\right) = d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}]}\right) = d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}_{\text{en}}]}\right). \tag{3}$$

Next, using Property 3.2 and Property 3.1, we have

$$\max_{\pi\in\Pi[\mathscr{I}_{\leq k}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = \max_{\pi\in\Pi[\mathcal{I}_\star]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) = \max_{\pi\in\Pi[\mathcal{I}_\star]} d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi\right) \cdot d_2(s[\mathcal{J}_{\text{ex}}]).$$

As a result, since $\pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}_{\text{en}}]}$ satisfies

$$\max_{\pi\in\Pi[\mathcal{I}_\star]} d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi\right) = d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}_{\text{en}}]}\right)$$

and it is an endogenous policy, we have

$$\begin{aligned}\max_{\pi\in\Pi[\mathscr{I}_{\leq k}]} d_2\left(s[\mathcal{J}]\ ;\ \pi\right) &= d_2\left(s[\mathcal{J}_{\text{en}}]\ ;\ \pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}_{\text{en}}]}\right) \cdot d_2(s[\mathcal{J}_{\text{ex}}]) \\ &= d_2\left(s[\mathcal{J}]\ ;\ \pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}_{\text{en}}]}\right) = d_2\left(s[\mathcal{J}]\ ;\ \pi_{s[\mathcal{J}\cap\widehat{\mathcal{I}}_{\text{en}}]}\right),\end{aligned}$$

where the second relation holds by Property 3.1, applicable since $\pi_{s[\mathcal{J}_{\text{en}}\cap\widehat{\mathcal{I}}_{\text{en}}]}$ is an endogenous policy, and the third relatin holds since $\mathcal{J}_{\text{en}} \cap \widehat{\mathcal{I}_{\text{en}}} = \mathcal{J} \cap \widehat{\mathcal{I}_{\text{en}}}$.

Thus, $\widehat{\mathcal{I}}_{\text{en}}$ satisfies the constraint in Line 3, and if $\widehat{\mathcal{I}}_{\text{ex}} \neq \varnothing$, we can reduce the cardinality while keeping the constraints satisfied, so the minimum cardinality solution is endogenous. □

3.2. Warm-Up II: Finding an Endogenous Policy Cover with Exact Occupancies ($H \geq 2$)

Algorithm 2 describes OSSR.Exact, which extends the OSSR.OneStep method to handle ExoMDPs with general horizon (rather than $H = 2$), but still requires exact access to occupancy measures. When invoked with a layer h, OSSR.Exact_h takes as input a sequence of endogenous policy covers $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$ for layers $1, \ldots, h-1$ and uses them to compute an endogenous policy cover $\Psi^{(h)}$ for layer h. The algorithm constructs $\Psi^{(h)}$ in a *backwards* fashion based on the dynamic programming principle. To describe the approach in detail, we use the notation of *$t \to h$ policy cover*.

Algorithm 2 OSSR.Exact$_h$: Optimization-Selection State Refinement with Exact Occupancies

require: Timestep $h \in [H]$, policy covers $\{\Psi^{(t)}\}_{t=1}^{h-1}$ for steps $1, \ldots, h-1$.
initialize: $\mathcal{I}^{(h,h)} \leftarrow \varnothing$ and $\Psi^{(h,h)} \leftarrow \varnothing$.
for $t = h-1, \ldots, 1$ **do**

Phase I: Optimization

Let $\mu^{(t)} := \mathrm{Unf}(\Psi^{(t)})$.

Find $\widetilde{\mathcal{I}} \in \mathscr{I}_{\leq k}$ with minimal cardinality such that for all $\mathcal{J} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$, $s\left[\mathcal{J}\right] \in \mathcal{S}\left[\mathcal{J}\right]$,

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{J}]\,;\, \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) = \max_{\pi \in \Pi[\widetilde{\mathcal{I}}]} d_h\left(s[\mathcal{J}]\,;\, \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right).$$

// Beginning from any state at layer t, $\pi^{(t)}_{s[\mathcal{J}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}$ maximizes probability that $s_h[\mathcal{J}] = s[\mathcal{J}]$.

For each factor set $\mathcal{J} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ and $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$, let

$$\pi_{s[\mathcal{J}]} \in \operatorname*{argmax}_{\pi \in \Pi[\widetilde{\mathcal{I}}]} d_h\left(s[\mathcal{J}]\,;\, \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right),$$

and define $\Gamma^{(t)}\left[\mathcal{J}\right] := \left\{\pi_{s[\mathcal{J}]} : s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]\right\}$.

Phase II: Selection

Find $\widehat{\mathcal{I}} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ with minimal cardinality s.t. for all $\mathcal{J} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$, $s\left[\mathcal{J}\right] \in \mathcal{S}\left[\mathcal{J}\right]$,

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{J}]\,;\, \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) = d_h\left(s[\mathcal{J}]\,;\, \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J} \cap \widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right).$$

Policy composition

Let $\mathcal{I}^{(t,h)} \leftarrow \widehat{\mathcal{I}}$, and for each $s[\mathcal{I}^{(t,h)}] \in \mathcal{S}[\mathcal{I}^{(t,h)}]$ define

$$\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} := \pi^{(t)}_{s[\mathcal{I}^{(t,h)}]} \circ_t \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}.$$

// Recall that $\pi^{(t)}_{s[\mathcal{I}^{(t,h)}]} \in \Gamma^{(t)}\left[\mathcal{I}^{(t,h)}\right]$ and $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]} \in \Psi^{(t+1,h)}$.

Let $\Psi^{(t,h)} \leftarrow \left\{\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} : s[\mathcal{I}^{(t,h)}] \in \mathcal{S}[\mathcal{I}^{(t,h)}]\right\}$.

return $\Psi^{(h)} := \Psi^{(1,h)}$

Definition 3.1. *For $h \in [H]$ and $t < h$, a set of non-stationary policies Ψ is said to be a (ϵ-approximate) $t \to h$* policy cover *with respect to a roll-in policy $\mu \in \Pi_{\text{mix}}$ if for all $s \in \mathcal{S}$,*

$$\max_{\psi \in \Psi} d_h\left(s[\mathcal{I}_\star]\,;\, \mu \circ_t \psi\right) \geq \max_{\pi \in \Pi_{\text{NS}}} d_h\left(s[\mathcal{I}_\star]\,;\, \mu \circ_t \pi\right) - \epsilon.$$

If all policies in Ψ are endogenous, we say that Ψ is endogenous.

OSSR.Exact$_h$ performs a serious of "backward" steps $t = h-1, \ldots, 1$. In each step t, the algorithm rolls in with the mixture policy $\mu^{(t)} := \mathrm{Unf}(\Psi^{(t)})$ and constructs a $t \to h$ policy cover $\Psi^{(t,h)}$ with respect to $\mu^{(t)}$. $\Psi^{(t,h)}$ acts on an endogenous factor set $\mathcal{I}^{(t,h)}$ (with $\mathcal{I}^{(t,h)} \supseteq \mathcal{I}^{(t+1,h)} \supseteq$

$\cdots \supseteq \mathcal{I}^{(h,h)} = \varnothing$), and is built from the next-step policy cover $\Psi^{(t+1,h)}$ via dynamic programming. In particular, the algorithm searches for a collection of endogenous "one-step" policies for choosing the action at time t that—when carefully composed with the $(t+1) \to h$ policy cover $\Psi^{(t+1,h)}$—result in a $t \to h$ policy cover. The algorithm ensures that the factor set $\mathcal{I}^{(t,h)}$ (upon which $\Psi^{(t,h)}$ acts) is endogenous using an optimization and selection phases analogous to those in OSSR.OneStep.

In more detail, $\mathsf{OSSR.Exact}_h$ satisfies the following invariants for $1 \leq t \leq h-1$.

(i) $\quad \mathcal{I}^{(h,h)} \subseteq \cdots \subseteq \mathcal{I}^{(t,h)} \subseteq \cdots \subseteq \mathcal{I}_\star$. ("state refinement")

(ii) $\quad$ The set $\Psi^{(t,h)}$ is an endogenous $t \to h$ policy cover with respect to $\mu^{(t)} = \mathrm{Unf}(\Psi^{(t)})$:

$$d_h\big(s[\mathcal{I}_\star]\ ;\ \mu^{(t)} \circ_t \psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]}\big) = \max_{\pi \in \Pi_{\mathrm{NS}}} d_h\big(s[\mathcal{I}_\star]\ ;\ \mu^{(t)} \circ_t \pi\big), \quad \forall s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star].$$

This implies that $\Psi^{(h)} := \Psi^{(1,h)}$ is an endogenous policy cover for layer h (Definition 2.3). In what follows we show how $\mathsf{OSSR.Exact}_h$ uses dynamic programming to satisfy these invariants.

Dynamic programming. Consider step $t < h-1$, and suppose that $(\mathcal{I}^{(t+1,h)}, \Psi^{(t+1,h)})$ satisfies invariants (i) and (ii). Because $\mu^{(t+1)}$ uniformly covers all states in layer $t+1$ (recall $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$ are policy covers), the policy $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}$ maximizes the probability that $s_h[\mathcal{I}_\star] = s[\mathcal{I}_\star]$, starting from any state in layer $t+1$. Hence, the Bellman optimality principle implies that to find a $t \to h$ policy to maximize this probability, it suffices to use the policy $\pi^{(t)} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}$, where $\pi^{(t)}$ solves the one-step problem:

$$\pi^{(t)} \in \operatorname*{argmax}_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s[\mathcal{I}_\star]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right). \tag{4}$$

At first glance, it is not apparent whether this observation is useful, because the endogenous factor set $\mathcal{I}_\star$ is not known to the learner, which prevents one from directly solving the optimization problem in Eq. (4). Fortunately, we can tackle this problem using a generalization of the optimization-selection approach of OSSR.OneStep. First, in the optimization phase (Line 5 and Line 6), we compute a collection of one-step policy covers $(\Gamma^{(t)}[\mathcal{J}])_{\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})}$, where $\Gamma^{(t)}[\mathcal{J}]$ consists of the policies that solve Eq. (4) with $\mathcal{I}_\star$ replaced by $\mathcal{J}$, for all possible choices of state in $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$. Then, in the selection phase (Line 7), we find a single factor $\mathcal{I}^{(t,h)} \supseteq \mathcal{I}^{(t+1,h)}$ such that $\Gamma^{(t)}[\mathcal{I}^{(t,h)}]$ provides good coverage (in the sense of Eq. (4)) for all factor sets $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$ simultaneously. Both steps ensure endogeneity by penalizing by cardinality in the same fashion as OSSR.OneStep. The success of this approach critically relies on the assumption that the preceding policy covers $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$ are endogenous, which ensures that the occupancy measures induced by $\mu^{(1)}, \ldots, \mu^{(h-1)}$ factorize (due to independence of the endogenous and exogenous state factors). To summarize:

Proposition 3.2. *If $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$ are endogenous policy covers for layers $1, \ldots, h-1$, then the set $\Psi^{(h)}$ returned by $\mathsf{OSSR.OneStep}_h$ is an endogenous policy cover for layer h, and has $|\Psi^{(h)}| \leq S^k$.*

We do not prove this result directly, and instead refer the reader to the proof of Theorem 3.1, which proves the sample-based version of the result using the same reasoning.

3.3. OSSR: Overview and Main Result

The full version of the OSSR algorithm ($\mathsf{OSSR}^{\epsilon,\delta}_h$) is given in Algorithm 8 (deferred to Appendix G due to space constraints). The algorithm follows the same template as OSSR.Exact: For each

Algorithm 3 ExoRL: RL in the Presence of Exogenous Information

require: precision parameter $\epsilon > 0$, reachability parameter $\eta > 0$, failure probability $\delta \in (0, 1)$.
initialize: $\Psi^{(1)} = \varnothing$.
for $h = 2, 3, \cdots, H$ **do**
  $\Psi^{(h)} \leftarrow \mathsf{OSSR}_h^{\eta/2,\delta}\big(\{\Psi^{(t)}\}_{t=1}^{h-1}\big)$. // Learn policy cover via OSSR (Algorithm 8 in Appendix G).
$\widehat{\pi} \leftarrow \mathsf{ExoPSDP}^{\epsilon,\delta}\big(\{\Psi^{(h)}\}_{h=1}^{H}\big)$. // Apply ExoPSDP (Algorithm 7 in Appendix F) to optimize rewards.
return $\widehat{\pi}$

$h \in [H]$, given policy covers $\Psi^{(1)}, \ldots, \Psi^{(h-1)}$, the algorithm builds a policy cover $\Psi^{(h)}$ for layer h in a backwards fashion using dynamic programming. There are two differences from the exact algorithm. First, since the MDP is unknown, the algorithm estimates the relevant occupancy measures for each backwards step using Monte Carlo rollouts. Second, the optimization and selection phases from OSSR.Exact are replaced by error-tolerant variants given by subroutines EndoPolicyOptimization and EndoFactorSelection (Algorithm 5 in Appendix D and Algorithm 6 in Appendix E, respectively).

Briefly, the EndoPolicyOptimization and EndoFactorSelection subroutines are based on approximate versions of the constraints used in the optimization and selection phase for OSSR.Exact (Line 5 and Line 7 of Algorithm 2), but ensuring endogeneity of the resulting factors is more challenging due to approximation errors, and it no longer suffices to simply search for the factor set with minimum cardinality. Instead, we search for factor sets that satisfy approximate versions of Line 5 and Line 7 with an *additive regularization term* based on cardinality. We show that as long as this penalty is carefully chosen as a function of the statistical error in the occupancy estimates, the resulting factor sets will be endogenous with high probability.

The main guarantee for Algorithm 8 is as follows.

Theorem 3.1 (Sample complexity of OSSR)**.** *Suppose that OSSR*$_h^{\epsilon,\delta}$ *is invoked with* $\{\Psi^{(t)}\}_{t=1}^{h-1}$, *where each* $\Psi^{(t)}$ *is an endogenous,* $\eta/2$*-approximate policy cover for layer t. Then with probability at least* $1-\delta$*, the set* $\Psi^{(h)}$ *returned by OSSR*$_h^{\epsilon,\delta}$ *is an endogenous* ϵ*-approximate policy cover for layer* h*, and has* $|\Psi^{(h)}| \le S^k$*. The algorithm uses at most* $O\left(AS^{4k}H^2k^3 \log\left(\frac{dSAH}{\delta}\right) \cdot \epsilon^{-2}\right)$ *episodes.*

By iterating the process $\Psi^{(h)} \leftarrow \mathsf{OSSR}_h^{\eta/2,\delta}(\{\Psi^{(t)}\}_{t=1}^{h-1})$, we obtain a policy cover for every layer.

4. Main Result: Sample-Efficient RL in the Presence of Exogenous Information

In this section we provide our main algorithm, ExoRL (Algorithm 3). ExoRL first applies OSSR iteratively to learn an endogenous, $\eta/2$-approximate policy cover for each layer, then applies a novel variant of the classical Policy Search by Dynamic Programming method of (Bagnell et al., 2004) (ExoPSDP), which uses the covers to optimize rewards; the original PSDP method cannot be applied to the ExoMDP setting as-is due to subtle statistical issues (cf. Appendix F for background). The main guarantee for ExoRL is as follows; see Appendix H for a proof and overview of analysis techniques.

Theorem 4.1 (Sample complexity of ExoRL)**.** *ExoRL, when invoked with parameter,* $\epsilon \in (0,1)$ *and* $\delta \in (0,1)$*, returns an* ϵ*-optimal policy with probability at least* $1-\delta$*, and does so using at most* $O\left(AS^{3k}H^2(S^k + H^2)k^3 \log\left(\frac{dSAH}{\delta}\right) \cdot \left(\epsilon^{-2} + \eta^{-2}\right)\right)$ *episodes.*

Recall that $S^k = |\mathcal{S}[\mathcal{I}_\star]|$ may thought of as the cardinality of the endogenous state space so—up to polynomial factors, logarithmic dependence on d, and dependence on the reachability parameter η, the sample complexity of ExoRL matches the optimal sample complexity when $\mathcal{I}_\star$ is known in advance.

Remark 4.1 (Computational Complexity of ExoRL)**.** The runtime for ExoRL scales with $\sum_{k'=0}^{k} \binom{d}{k'} = \Theta(d^k)$ due to brute force enumeration over factors sets of size at most k. While this improves over the S^d runtime required to run a tabular RL algorithm over the full state space, an interesting question that remains is whether the runtime can be improved to $O(d^c)$ for some constant c independent of k.

5. Related Work

In this section we highlight additional related work not already covered by our discussion.

Reinforcement learning with exogenous information. The ExoMDP setting is a special case of the Exogenous Block MDP (EX-BMDP) setting introduced by Efroni et al. (2021b), who initiated the study of sample-efficient reinforcement learning with temporally correlated exogenous information. In particular, one can view the ExoMDP as an EX-BMDP with $\mathcal{S}$ as the observation space and $\mathcal{S}[\mathcal{I}_\star]$ as the latent state space, and with the set $\Phi := \{s \mapsto s[\mathcal{I}] \mid |\mathcal{I}| \leq k\}$ as the class of decoders. Efroni et al. (2021b) provide an EX-BMDP algorithm whose sample complexity scales with the size of the latent state space and with $\log|\Phi|$, which translates to $\mathrm{poly}(S^k, \log(d))$ sample complexity for the ExoMDP setting, but the algorithm requires that the endogenous state space has deterministic transitions and initial state. The motivation for the present work was to take a step back and provide a simplified testbed in which to study the problem of learning with stochastic transitions, as well as other refined issues (e.g., minimax rates). Also related to this line of research is Efroni et al. (2021a), which considers a linear control setting with exogenous observations. Unlike our work, Efroni et al. (2021a) assumes that the inherent system noise induces sufficient exploration, and hence does not address the exploration problem.

Empirical and theoretical works that aim to filter exogenous noise in RL include Pathak et al. (2017); Zhang et al. (2020); Gelada et al. (2019) and Dietterich et al. (2018), but these methods do not come with finite sample guarantees nor tackle the exploration problem.

Tabular reinforcement learning. As discussed earlier, existing approaches to tabular reinforcement learning (Azar et al., 2017; Jin et al., 2018; Zanette and Brunskill, 2019; Kaufmann et al., 2021) incur $\Omega(S^d)$ sample complexity if applied to the ExoMDP setting naively. One can improve this sample complexity to $\mathrm{poly}(S^k, d^k, A, H)$ using a simple reduction. This falls short of the $\mathrm{poly}(S^k, A, H, \log(d))$ sample complexity our algorithms obtain, we sketch the reduction for completeness.

- For each $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| \leq k$, run any optimal tabular RL algorithm with precision parameter ϵ over the state space $\mathcal{S}[\mathcal{I}]$, and let $\pi_\mathcal{I}$ be the resulting policy.
- Evaluate each policy $\pi_\mathcal{I}$ to precision ϵ using Monte-Carlo rollouts, and take the best one.

The first phase has $\mathrm{poly}(S^k, A, H)$ sample complexity for each set $\mathcal{I}$, and there are at most $\binom{d}{k} = O(d^k)$ subsets. The algorithm that runs on $\mathcal{S}[\mathcal{I}_\star]$ will succeed in finding an ϵ-optimal policy with high probability, so the policy returned in the second phase will be at least 2ϵ-optimal.

Factored Markov decision processes. The ExoMDP setting is related to the Factored MDP model (Kearns and Koller, 1999). Factored MDPs assume a factored state space whose transition dynamics obey the following structure:

$$\forall s, s' \in \mathcal{S}^d, a \in \mathcal{A}, \qquad T(s' \mid s, a) = \prod_{i=1}^{d} T_i(s'[i] \mid s[\mathrm{pt}(i)], a),$$

where $\mathrm{pt} : [d] \to 2^d$ is a *parent function* and $T_i : \mathcal{S}^{|\mathrm{pt}(i)|} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition distribution of the ith factor. Many algorithms have been proposed for Factored MDPs, including for the setting where the parent function is unknown (Strehl et al., 2009; Diuk et al., 2009; Hallak et al., 2015; Guo and Brunskill, 2017; Rosenberg and Mansour, 2020; Misra et al., 2021). These algorithms assume that the parent factor size is bounded, i.e., $|\mathrm{pt}(i)| \leq \kappa$ for all $i \in [d]$, and their sample complexity typically scales with $O(|\mathcal{S}|^{c\kappa})$ for a numerical constant c. The ExoMDP setting cannot be solved using off-the-shelf factored MDP algorithms for two reasons. First, we do not assume that each factor evolves independently of other factors given the previous state and action. Second, the size of the parent set for an exogenous factor can be as large as $d - k$. Therefore, even if factors were evolving independently, applying off-the-shelf Factored MDPs algorithms would lead to exponential sample in d sample complexity.

6. Conclusion

We have introduced the ExoMDP setting and provided ExoRL, the first algorithm for sample-efficient reinforcement learning in stochastic systems with high-dimensional, exogenous information. Going forward, we believe that the ExoMDP setting will serve as a useful testbed to understand refined aspects of learning with exogenous information. Natural questions we hope to see addressed include:

- *Minimax rates.* While our results provide polynomial sample complexity, it remains to understand the precise minimax rate for the ExoMDP as a function on S^k, H, and so on. Additionally, either removing the dependence on the reachability parameter or establishing a lower bound remains for its necessity is an issue which deserves further investigation.

- *Computation.* Both ExoRL and OSSR rely on brute force enumeration over subsets, which results in $\Omega(d^k)$ runtime. While this provides an improvement over naive tabular RL, it remains to see whether it is possible to develop an algorithm with runtime $O(d^c)$, where $c > 0$ is a constant independent of k.

- *Regret.* Naively lifting our ϵ-PAC results to regret results in $T^{2/3}$-type dependence on the time horizon T. Developing algorithms with $\sqrt{T}$-type regret will require new techniques.

- *Parameter-free algorithms.* The OSSR algorithm requires an upper bound on $|\mathcal{I}_\star|$ and a lower bound on η. It is relatively straightforward to remove access to these quantities when the value of the optimal policy ($\max_\pi J(\pi)$) is known, by an application of the doubling trick. However, developing truly parameter-free algorithms is an interesting direction.

Finally, the problem of learning in the ExoMDP model is related to the notion of out-of-distribution generalization and learning in the presence of acausal features (Peters et al., 2016; Arjovsky et al., 2019; Kim et al., 2019; Wald et al., 2021). It would be interesting to explore these connections in more detail. Beyond these questions, we hope that our techniques will find further use beyond the tabular setting.

References

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Advances in Neural Information Processing Systems*, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

J Andrew Bagnell, Sham M Kakade, Jeff G Schneider, and Andrew Y Ng. Policy search by dynamic programming. In *Advances in Neural Information Processing Systems*, 2004.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Thomas G Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. *arXiv preprint arXiv:1806.01584*, 2018.

Carlos Diuk, Lihong Li, and Bethany R Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 249–256, 2009.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Simon S Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019.

Yonathan Efroni, Sham Kakade, Akshay Krishnamurthy, and Cyril Zhang. Sparsity in partially controllable linear systems. *arXiv preprint arXiv:2110.06150*, 2021a.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provable RL with exogenous distractors via multistep inverse dynamics. *arXiv preprint arXiv:2110.08847*, 2021b.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2019.

Zhaohan Daniel Guo and Emma Brunskill. Sample efficient feature selection for factored MDPs. *arXiv preprint arXiv:1703.03454*, 2017.

Assaf Hallak, François Schnitzler, Timothy Mann, and Shie Mannor. Off-policy model-based learning under unknown factored dynamics. In *International Conference on Machine Learning*, pages 711–719. PMLR, 2015.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143:143, 2015.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR, 2021.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 740–747, 1999.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Dipendra Misra, Qinghua Liu, Chi Jin, and John Langford. Provable rich observation reinforcement learning with combinatorial latent states. In *International Conference on Learning Representations*, 2021.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient reinforcement learning in factored MDPs with unknown structure. *arXiv preprint arXiv:2009.05986*, 2020.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10(11), 2009.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. *arXiv:2110.04652*, 2021.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Ruosong Wang, Dean Foster, and Sham M Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2020.

Contents of Appendix

Organization and Notation

The appendix contains three parts, Part I, Part II, and Part III.

Part I: Preliminaries. In Part I we provide basic technical results used in our analysis. Appendix A contains technical lemmas for reinforcement learning (Appendix A.1), concentration inequalities (Appendix A.2), and basic analysis tools (Appendix A.3). In Appendix A.4, we provide a simple, yet useful result which shows that the collection $\mathscr{I}_{\leq k}(\mathcal{I})$ is a π-system for any factor set $\mathcal{I}$ with $|\mathcal{I}| \leq k$.

In Appendix B we present structural results for the ExoMDP model. We begin by establishing a negative result (Appendix B.1) which shows that the Bellman rank (Jiang et al., 2017) of for the ExoMDP model scales with the number of exogenous factors. In Appendix B.2 and Appendix B.3, we prove key structural results for the ExoMDP model, including a decoupling property (Lemma B.1) and restriction lemma (Lemma B.2) for occupancy measures, a restriction lemma for endogenous rewards (Lemma B.7), and a performance difference lemma for endogenous policies (Lemma B.6).

In Appendix C, we present an algorithmic template, AbstractFactorSearch, which forms the basis for the subroutines in OSSR.

Notation used throughout the main paper and appendix is collected in Table 1.

Part II: Omitted subroutines. In Part II, we describe and analyze subroutines used by OSSR and ExoRL. Appendix D presents and analyzes the EndoPolicyOptimization subroutine used in OSSR and ExoPSDP. Appendix E we presents and analyzes the EndoFactorSelection subroutine used in OSSR. Finally, Appendix F presents and analyzes ExoPSDP algorithm, which is used by ExoRL.

Part III: Additional details and proofs for main results. In Part III, we present our main results and their proofs. In Appendix G, we present and analyze the full version of the OSSR algorithm, and in Appendix H, we combine the results for OSSR and ExoPSDP to establish the main sample complexity bound for ExoRL.

Notation	Meaning
$\mathcal{I}$	an ordered set of factors (a set of distinct elements from $[d]$).
$\mathscr{I}_{\leq k}(\mathcal{I})$	$\{\mathcal{J} \subseteq [d] : \mathcal{J} \supseteq \mathcal{I}, \lvert\mathcal{J}\rvert \leq k\}$.
$\mathscr{I}_{k}(\mathcal{I})$	$\{\mathcal{J} \subseteq [d] : \mathcal{J} \supseteq \mathcal{I}, \lvert\mathcal{J}\rvert = k\}$.
$\mathscr{I}_{\leq k}$	$\{\mathcal{J} \subseteq [d] : \lvert\mathcal{J}\rvert \leq k\}$, or equivalently, $\mathscr{I}_{\leq k} = \mathscr{I}_{\leq k}(\varnothing)$.
$\mathscr{I}_{k}$	$\{\mathcal{J} \subseteq [d] : \lvert\mathcal{J}\rvert = k\}$, or equivalently, $\mathscr{I}_{k} = \mathscr{I}_{k}(\varnothing)$.
$\Pi[\mathcal{I}]$	the set of policies that depend only on the factors specified in $\mathcal{I}$.
$\Pi[\mathscr{I}]$	the union of the set of policies $\cup_{\mathcal{I}\in\mathscr{I}}\Pi[\mathcal{I}]$.
$\mathcal{I}_\star$	the set of endogenous factors.
$\mathcal{I}_\star^{\mathrm{c}}$	the set of exogenous factors.
$\mathcal{S}[\mathcal{I}]$	the set of states induced by the factors in $\mathcal{I}$.
$s[\mathcal{I}]$	the state s restricted to the set of factors $\mathcal{I}$.
V_1^π	value of a policy π measured with respect to an initial distribution.
$V_h^\pi(s)$	value of a policy π measured from state s at timestep h
$V_{t,h}$	$V_{t,h}(\pi) := \mathbb{E}_\pi\left[\sum_{t'=t}^{h} r_t\right]$.
$Q_h^\pi(s,a)$	Q-function for a policy π measured from state s at timestep h.
$d_h(s[\mathcal{I}];\pi)$	shorthand for $\mathbb{P}^\pi(s_h[\mathcal{I}] = s[\mathcal{I}])$.
$d_h(s[\mathcal{I}] \mid s_t[\mathcal{I}'] = s[\mathcal{I}'];\pi)$	shorthand for $\mathbb{P}^\pi(s_h[\mathcal{I}] = s[\mathcal{I}] \mid s_t[\mathcal{I}'] = s[\mathcal{I}'])$.
$\pi_1 \circ_t \pi_2$	Policy that executes π_1 until step $t-1$ and executes π_2 from then on.
$\mathcal{I}_{\mathrm{en}}$	For a set of factors $\mathcal{I}$, $\mathcal{I}_{\mathrm{en}} := \mathcal{I} \cap \mathcal{I}_\star$.
$\mathcal{I}_{\mathrm{ex}}$	For a set of factors $\mathcal{I}$, $\mathcal{I}_{\mathrm{ex}} := \mathcal{I} \cap \mathcal{I}_\star^{\mathrm{c}}$.

Table 1: Summary of notation.

Part I
Preliminaries

Appendix A. Supporting Lemmas

A.1. Reinforcement Learning

Lemma A.1 (Performance difference lemma (Kakade and Langford (2002), Lemma 6.1))**.** *Consider a fixed MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R, H, \mu)$. For any pair of policies $\pi, \pi' \in \Pi_{\mathrm{NS}}$,*

$$J(\pi) - J(\pi') = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} Q_t^{\pi'}(s_t, \pi_t(s_t)) - Q_t^{\pi'}(s_t, \pi'_t(s_t))\right].$$

Lemma A.2 (Density ratio bound for policy cover)**.** *Let Ψ be an endogenous ϵ-approximate policy cover for timestep t and $\mu^{(t)} := \mathrm{Unf}\,(\Psi)$. Then, for any $s[\mathcal{I}_\star] \in \mathcal{S}\,[\mathcal{I}_\star]$ such that $\max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} d_t(s[\mathcal{I}_\star]\ ;\ \pi) \geq 2\epsilon$, it holds that*

$$\max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} \frac{d_t(s[\mathcal{I}_\star]\ ;\ \pi)}{d_t(s[\mathcal{I}_\star]; \mu^{(t)})} \leq 2S^k.$$

Proof of Lemma A.2. Fix $s[\mathcal{I}_\star] \in \mathcal{S}\,[\mathcal{I}_\star]$. Since Ψ is an endogenous ϵ-approximate policy cover, there exists $\psi_{s[\mathcal{I}_\star]} \in \Psi$ such that

$$\max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} d_t(s[\mathcal{I}_\star]; \pi) \leq d_t(s[\mathcal{I}_\star]; \psi_{s[\mathcal{I}_\star]}) + \epsilon. \tag{5}$$

Thus, we have that

$$\begin{aligned}
\max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} \frac{d_t(s[\mathcal{I}_\star]; \pi)}{d_t(s[\mathcal{I}_\star]; \mu^{(t)})} &\overset{\text{(a)}}{=} S^k \max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} \frac{d_t(s[\mathcal{I}_\star]; \pi)}{\sum_{s'[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} d_t(s[\mathcal{I}_\star]; \psi_{s'[\mathcal{I}_\star]})} \\
&\overset{\text{(b)}}{\leq} S^k \max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} \frac{d_t(s[\mathcal{I}_\star]; \pi)}{d_t(s[\mathcal{I}_\star]; \pi_{s[\mathcal{I}_\star]})} \\
&\overset{\text{(c)}}{\leq} S^k \frac{\max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} d_t(s[\mathcal{I}_\star]; \pi)}{\max_{\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]} d_t(s[\mathcal{I}_\star]; \pi) - \epsilon}.
\end{aligned}$$

Here, (a) holds because $\mu^{(t)} = \mathrm{Unf}\,(\Psi)$, (b) holds because $d_t(s[\mathcal{I}_\star]; \psi_{s'[\mathcal{I}_\star]}) \geq 0$ for all $\psi_{s'[\mathcal{I}_\star]} \in \Psi$, and (c) holds by Eq. (5). Finally, since $x/(x-\epsilon) \leq 2$ for $x \geq 2\epsilon$, we conclude the proof. □

A.2. Probability

Lemma A.3 (Bernstein's Inequality (e.g., Boucheron et al. (2013)))**.** *Let $X_1, .., X_N$ be a sequence of i.i.d. random variables with $\mathbb{E}\,[X_i] = \mu$, $\mathbb{E}\big[(X_i - \mu)^2\big] = \sigma^2$, and $|X_i - \mu| \leq C$ almost surely. Then for all $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)\right| \geq \sqrt{\frac{2\sigma^2 \log\left(\frac{2}{\delta}\right)}{N}} + \frac{C \log\left(\frac{2}{\delta}\right)}{N}\right) \leq \delta.$$

Lemma A.4 (Union bound for sequences)**.** *Let* $\{\mathcal{G}_t\}_{t=1}^h$ *be a sequence of events. If* $\mathbb{P}(\mathcal{G}_t \mid \cap_{t'=1}^{t-1}\mathcal{G}_{t'}) \geq 1-\delta$ *for all* $t \in [h]$*, then* $\mathbb{P}(\cap_{t=1}^h \mathcal{G}_t) \geq 1 - h\delta$.

Proof of Lemma A.4. We prove the claim by induction. The base case $h = 1$ holds by assumption. Now, suppose the claim holds for some $h' \leq h$:

$$\mathbb{P}(\cap_{t=1}^{h'}\mathcal{G}_t) \geq 1 - h'\delta.$$

By Bayes' rule, we have that

$$\begin{aligned}
&\mathbb{P}(\cap_{t=1}^{h'+1}\mathcal{G}_t) \\
&= \mathbb{P}(G_{h'+1} \mid \cap_{t=1}^{h'}\mathcal{G}_t)\mathbb{P}(\cap_{t=1}^{h'}\mathcal{G}_t) \\
&\overset{\text{(a)}}{\geq} \mathbb{P}(G_{h'+1} \mid \cap_{t=1}^{h'}\mathcal{G}_t)\left(1 - h'\delta\right) \\
&\overset{\text{(b)}}{\geq} (1-\delta)\left(1 - h'\delta\right) \\
&\geq 1 - (h'+1)\delta,
\end{aligned}$$

where (a) holds by the induction hypothesis and (b) holds by assumption of the lemma. This proves the induction step and concludes the proof. □

A.2.1. CONCENTRATION FOR OCCUPANCY MEASURES

Definition A.1 (ϵ-approximate occupancy measure collection)**.** *Let* $\widehat{\mathcal{D}} = \left\{\widehat{d}_h\left(\cdot\ ;\ \pi\right) \mid \pi \in \Pi\right\}$, *be a set of occupancy measures for timestep* h*. We say that* $\widehat{\mathcal{D}}$ *is* ϵ*-approximate with respect to* $(\Pi, \mathscr{I}, h)$ *if for all* $\pi \in \Pi, \mathcal{I} \in \mathscr{I}$ *and* $s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$ *it holds that*

$$\left|\widehat{d}_h\left(s_h\left[\mathcal{I}\right] = s\left[\mathcal{I}\right]\ ;\ \pi\right) - d_h\left(s_h\left[\mathcal{I}\right] = s\left[\mathcal{I}\right]\ ;\ \pi\right)\right| \leq \epsilon.$$

In the following lemma, we bound the sample complexity required to compute a set of ϵ-approximate occupancy measures with respect to $(\mu \circ \Pi \circ \Psi, \mathscr{I}, h)$, where μ is a fixed policy, Π is a set of 1-step policies, and Ψ is a set of non-stationary policies. The proof follows from a simple application of Bernstein's inequality and a union bound.

Lemma A.5 (Sample complexity for ϵ-approximate occupancy measures)**.** *Let* $t, h \in \mathbb{N}$ *with* $t \leq h$ *be given. Fix a mixture policy* $\mu \in \Pi_{\text{mix}}$*, a collection* $\Gamma \subseteq \Pi$ *of 1-step policies, a set* $\Psi \subseteq \Pi_{\text{NS}}$*, and a collection of factors* $\mathscr{I}$*. Assume the following bounds hold:*

1. $|\Psi| \leq S^k$.
2. $|\Gamma| \leq O\left(d^k A^{S^k}\right)$.
3. $|\mathscr{I}| \leq O\left(d^k\right)$.
4. *For any* $\mathcal{I} \in \mathscr{I}$ *it holds that* $|\mathcal{S}\left[\mathcal{I}\right]| \leq S^k$.

Consider the dataset $\mathcal{Z}_{t,h}^N = \{(s_{t,n}, a_{t,n}, \psi_n, s_{h,n})\}_{n=1}^N$ *generated by the following process:*

- *Execute* $\mu^{(t)} := \mathrm{Unf}(\Psi^{(t)})$ *up to layer* t *(resulting in state* $s_{t,n}$*).*
- *Sample action* $a_{t,n} \sim \mathrm{Unf}(\mathcal{A})$ *and play it, transitioning to* $s_{t+1,n}$ *in the process.*
- *Sample* $\psi_n^{(t+1,h)} \sim \mathrm{Unf}(\Psi^{(t+1,h)})$ *and execute it from layers* $t+1$ *to* h *(resulting in* $s_{h,n}$*).*

Define a collection of empirical occupancies

$$\widehat{\mathcal{D}} = \left\{ \widehat{d}_h \left(\cdot \; ; \; \mu \circ_t \pi \circ_{t+1} \psi^{(t+1,h)} \right) \mid \pi \in \Gamma, \psi^{(t+1,h)} \in \Psi \right\},$$

where $\widehat{d}_h \left(\cdot \; ; \; \mu \circ_t \pi \circ_{t+1} \psi^{(t+1,h)} \right)$ *is given by (see also Line 5 in Algorithm 8)*

$$\widehat{d}_h(s \; ; \; \mu \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}) = \frac{1}{N} \sum_{n=1}^{N} \frac{\mathbb{1}\left\{ a_{t,n} = \pi(s_{t,n}), \psi_n^{(t+1,h)} = \psi^{(t+1,h)}, s_{h,n} = s \right\}}{(1/|\mathcal{A}|) \cdot (1/|\Psi|)}. \quad (6)$$

Then, whenever $N = \Omega\left(\frac{AS^{2k} k \log\left(\frac{dSA}{\delta}\right)}{\epsilon^2} \right)$ *trajectories, with probability at least* $1-\delta$ *it holds that* $\widehat{\mathcal{D}}$ *is* ϵ*-approximate with respect to* $(\mu \circ_t \Gamma \circ_{t+1} \Psi, \mathscr{I}, h)$*.*

Proof of Lemma A.5. Denote ρ as the policy that generates the data $\mathcal{Z}_{t,h}^N$. Fix $\pi \in \Gamma, \psi \in \Psi, \mathcal{I} \in \mathscr{I}, s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$. It holds that

$$\begin{aligned}
&\widehat{d}_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi) - d_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi) \\
&\overset{\text{(a)}}{=} \sum_{s[\mathcal{I}^c] \in \mathcal{S}[\mathcal{I}^c]} \widehat{d}_h(s \; ; \; \mu \circ_t \pi \circ_{t+1} \psi) - d_h(s \; ; \; \mu \circ_t \pi \circ_{t+1} \psi) \\
&= \frac{1}{N} \sum_{n=1}^{N} \frac{\mathbb{1}\{ a_{t,n} = \pi(s_{t,n}), \psi_n = \psi, s_{h,n}\left[\mathcal{I}\right] = s\left[\mathcal{I}\right] \}}{(1/|\mathcal{A}|) \cdot (1/|\Psi|)} - d_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left(X_n\left(\pi, \psi, s_h\left[\mathcal{I}\right]\right) - d_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi) \right)
\end{aligned}$$

where

$$X_n\left(\pi, \psi, s_h\left[\mathcal{I}\right]\right) := \frac{\mathbb{1}\{ a_{t,n} = \pi(s_{t,n}), \psi_n = \psi, s_{h,n}\left[\mathcal{I}\right] = s\left[\mathcal{I}\right] \}}{(1/|\mathcal{A}|) \cdot (1/|\Psi|)}.$$

Note that (a) holds by definition: both $\widehat{d}_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi)$ and $d_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi)$ are given by marginalizing all state factors in $\mathcal{I}^c$. Observe that the estimator X_n is unbiased and bounded almost surely:

$$\mathbb{E}_\rho[X_n\left(\pi, \psi, s\left[\mathcal{I}\right]\right)] = d_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi), \quad \text{and } 0 \leq X_n\left(\pi, \psi, s\left[\mathcal{I}\right]\right) \leq A\left|\Psi\right|. \quad (7)$$

As a result, we can control the quality of approximation of $\widehat{d}_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi)$ using Bernstein's inequality (Lemma A.3). First, observe that the variance of each term in the sum can be bounded as follows:

$$\sigma^2 := \mathbb{E}_\rho[(X_n\left(\pi, \psi, s\left[\mathcal{I}\right]\right) - d_h(s\left[\mathcal{I}\right] \; ; \; \mu \circ_t \pi \circ_{t+1} \psi))]$$

$$\begin{aligned}
&\overset{\text{(a)}}{\leq} \mathbb{E}_\rho[X_n\left(\pi, \psi, s\left[\mathcal{I}\right]\right)^2] \\
&\overset{\text{(b)}}{\leq} A\left|\Psi\right| \mathbb{E}_\rho[X_n\left(\pi, \psi, s\left[\mathcal{I}\right]\right)] \\
&\overset{\text{(c)}}{=} A\left|\Psi\right| d_h(s\left[\mathcal{I}\right] \ ; \ \mu \circ_t \pi \circ_{t+1} \psi) \\
&\leq A\left|\Psi\right|. \qquad (8)
\end{aligned}$$

Here (a) holds since $d_h(s\left[\mathcal{I}\right] \ ; \ \mu \circ_t \pi \circ_{t+1} \psi) \geq 0$, (b) holds since $0 \leq X_n\left(\pi, \psi, s\left[\mathcal{I}\right]\right) \leq A|\Psi|$, and (c) holds by Eq. (7). As a result, using Bernstein's inequality, we have that for any fixed $\pi \in \Gamma, \psi \in \Psi, \mathcal{I} \in \mathscr{I}, s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$, with probability at least $1 - \delta$,

$$\begin{aligned}
&\left|\widehat{d}_h(s\left[\mathcal{I}\right] \ ; \ \mu \circ_t \pi \circ_{t+1} \psi) - d_h(s\left[\mathcal{I}\right] \ ; \ \mu \circ_t \pi \circ_{t+1} \psi)\right| \\
&\overset{\text{(a)}}{\leq} O\left(\sqrt{\frac{\sigma^2 \log\left(\frac{1}{\delta}\right)}{N}} + \frac{A\left|\Psi\right| \log\left(\frac{1}{\delta}\right)}{N}\right) \\
&\overset{\text{(b)}}{\leq} O\left(\sqrt{\frac{A\left|\Psi\right| \log\left(\frac{1}{\delta}\right)}{N}} + \frac{A\left|\Psi\right| \log\left(\frac{1}{\delta}\right)}{N}\right),
\end{aligned}$$

where (a) holds byLemma A.3 and (b) holds by Eq. (8). Setting $N = \Theta\left(\frac{A|\Psi| \log 1 \backslash \delta}{\epsilon^2}\right)$ and using that $\epsilon^2 \leq \epsilon$ for $\epsilon \in (0, 1)$, we find that

$$\left|\widehat{d}_h(s\left[\mathcal{I}\right] \ ; \ \mu \circ_t \pi \circ_{t+1} \psi) - d_h(s\left[\mathcal{I}\right] \ ; \ \mu \circ_t \pi \circ_{t+1} \psi)\right| \leq O\left(\epsilon + \epsilon^2\right) \leq \epsilon.$$

Finally, taking a union bound over all $\pi \in \Gamma, \psi \in \Psi, \mathcal{I} \in \mathscr{I}, s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$ and using assumptions $(1) - (4)$, we conclude the proof. □

A.3. Analysis

The following elementary result shows that if two functions $\widehat{f}, f : \mathcal{X} \to \mathbb{R}$ are point-wise close, any approximate optimizer for $\widehat{f}$ is an approximate optimizer for f.

Lemma A.6. *Let $\mathcal{X}$ be a compact set, and let $f, \widehat{f} : \mathcal{X} \to \mathbb{R}$ be such that*

$$||\widehat{f} - f||_\infty := \max_{x \in \mathcal{X}} \left|\widehat{f}(x) - f(x)\right| \leq \epsilon.$$

Then, for any $\epsilon' > 0$, the following results hold:

1. *If $\max_{x \in \mathcal{X}} \widehat{f}(x) > \min_{x \in \mathcal{X}} \widehat{f}(x) + \epsilon'$, then $\max_{x \in \mathcal{X}} f(x) > \min_{x \in \mathcal{X}} f(x) + \epsilon' - 2\epsilon$.*
2. *If $\max_{x \in \mathcal{X}} \widehat{f}(x) \leq \min_{x \in \mathcal{X}} \widehat{f}(x) + \epsilon'$, then $\max_{x \in \mathcal{X}} f(x) \leq \min_{x \in \mathcal{X}} f(x) + \epsilon' + 2\epsilon$.*
3. *For any $\widehat{x} \in \mathcal{X}$, if $\max_{x \in \mathcal{X}} \widehat{f}(x) > \widehat{f}(\widehat{x}) + \epsilon'$, then $\max_{x \in \mathcal{X}} f(x) > f(\widehat{x}) + \epsilon' - 2\epsilon$.*
4. *For any $\widehat{x} \in \mathcal{X}$, if $\max_{x \in \mathcal{X}} \widehat{f}(x) \leq \widehat{f}(\widehat{x}) + \epsilon'$, then $\max_{x \in \mathcal{X}} f(x) \leq f(\widehat{x}) + \epsilon' + 2\epsilon$.*

Proof of Lemma A.6. Denote the maximizer and minimizer of f by

$$x_{\min,f} := \arg\min_{x\in\mathcal{X}} f(x), \quad x_{\max,f} := \arg\max_{x\in\mathcal{X}} f(x),$$

and denote the maximizer and minimizer of $\widehat{f}$ by

$$x_{\min,\widehat{f}} := \arg\min_{x\in\mathcal{X}} \widehat{f}(x), \quad x_{\max,\widehat{f}} := \arg\max_{x\in\mathcal{X}} \widehat{f}(x).$$

Note that these points exist by compactness of $\mathcal{X}$.

Observe that the following relations hold by the assumption that $||\widehat{f} - f||_\infty \leq \epsilon$:

$$\max_{x\in\mathcal{X}} \widehat{f}(x) = \widehat{f}(x_{\max,\widehat{f}}) \leq f(x_{\max,\widehat{f}}) + \epsilon \leq \max_{x\in\mathcal{X}} f(x) + \epsilon, \tag{9}$$

$$\min_{x\in\mathcal{X}} \widehat{f}(x) = \widehat{f}(x_{\min,\widehat{f}}) \geq f(x_{\min,\widehat{f}}) - \epsilon \geq \min_{x\in\mathcal{X}} f(x) - \epsilon, \tag{10}$$

$$\max_{x\in\mathcal{X}} \widehat{f}(x) \geq \widehat{f}(x_{\max,f}) \geq f(x_{\max,f}) - \epsilon = \max_{x\in\mathcal{X}} f(x) - \epsilon, \tag{11}$$

$$\min_{x\in\mathcal{X}} \widehat{f}(x) \leq \widehat{f}(x_{\min,f}) \leq f(x_{\min,f}) + \epsilon = \min_{x\in\mathcal{X}} f(x) + \epsilon. \tag{12}$$

Proof of the first claim. Combining relations Eq. (9), Eq. (10) and rearranging, we have

$$\max_{x\in\mathcal{X}} \widehat{f}(x) > \min_{x\in\mathcal{X}} \widehat{f}(x) + \epsilon' \implies \max_{x\in\mathcal{X}} f(x) > \min_{x\in\mathcal{X}} f(x) + \epsilon' - 2\epsilon.$$

Proof of the second claim. Combining relations Eq. (11), Eq. (12) and rearranging, we have

$$\max_{x\in\mathcal{X}} \widehat{f}(x) \leq \min_{x\in\mathcal{X}} \widehat{f}(x) + \epsilon' \implies \max_{x\in\mathcal{X}} f(x) \leq \min_{x\in\mathcal{X}} f(x) + \epsilon' + 2\epsilon.$$

Proof of the third claim. By Eq. (9) and the assumption that $||f - \widehat{f}||_\infty \leq \epsilon$, we have

$$\max_{x\in\mathcal{X}} \widehat{f}(x) > \widehat{f}(\widehat{x}) + \epsilon' \implies \max_{x\in\mathcal{X}} f(x) > f(\widehat{x}) + \epsilon' - 2\epsilon.$$

Proof of the fourth claim. By Eq. (11) and the assumption that $\|f - \widehat{f}\|_\infty \leq \epsilon$, we have

$$\max_{x\in\mathcal{X}} \widehat{f}(x) \leq \widehat{f}(\widehat{x}) + \epsilon' \implies \max_{x\in\mathcal{X}} f(x) \leq f(\widehat{x}) + \epsilon' + 2\epsilon.$$

□

Lemma A.7 (Equivalence of Maximizers for Scaled Positive Functions)**.** *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{A}$ be finite sets. Let $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}_+$ and let $\mathbb{P}$ be a probability measure over $\mathcal{X} \times \mathcal{Y}$. Let $\Pi_{\mathcal{X}\times\mathcal{Y}}$ and $\Pi_{\mathcal{X}}$ be the sets of all mappings from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{A}$ and $\mathcal{X}$ to $\mathcal{A}$, respectively. Then,*

$$\max_{\pi\in\Pi_{\mathcal{X},\mathcal{Y}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x,y))g(y)\right] = \max_{\pi\in\Pi_{\mathcal{X}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x))g(y)\right].$$

Proof of Lemma A.7. By the skolemization lemma (Lemma A.9), we can exchange maximization and expectation by writing

$$\max_{\pi \in \Pi_{\mathcal{X},\mathcal{Y}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x,y))g(y)\right] = \mathbb{E}_{x,y\sim\mathbb{P}}\left[\max_{a\in\mathcal{A}}\left(f(x,a)g(y)\right)\right]. \tag{13}$$

Let $\pi_f^\star \in \Pi_{\mathcal{X}}$ be defined via

$$\pi_f^\star(x) \in \max_a f(x,a).$$

Observe that for any $x, y \in \mathcal{X}\times\mathcal{Y}$ it holds that

$$\max_a\left(f(x,a)g(y)\right) \overset{\text{(a)}}{=} g(y)\max_a f(x,a) = g(y)f(x,\pi_f^\star(x)), \tag{14}$$

where (a) holds because $g(y)\geq 0$. Plugging Eq. (14) back into Eq. (13) we find that

$$\max_{\pi\in\Pi_{\mathcal{X},\mathcal{Y}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x,y))g(y)\right] \overset{\text{(a)}}{=} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi_f^\star(x))g(y)\right] \overset{\text{(b)}}{\leq} \max_{\pi\in\Pi_{\mathcal{X}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x))g(y)\right], \tag{15}$$

where (a) holds by Eq. (14), and (b) holds since $\pi_f^\star\in\Pi_{\mathcal{X}}$. Finally, observe that we trivially have

$$\max_{\pi\in\Pi_{\mathcal{X},\mathcal{Y}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x,y))g(y)\right] \geq \max_{\pi\in\Pi_{\mathcal{X}}} \mathbb{E}_{x,y\sim\mathbb{P}}\left[f(x,\pi(x))g(y)\right], \tag{16}$$

since $\Pi_{\mathcal{X}}\subseteq\Pi_{\mathcal{X},\mathcal{Y}}$. Combining Eq. (15) and Eq. (16) yields the result. □

Lemma A.8. *Let $k, k_1, k_2 \in \mathbb{N}$ satisfying $1 \leq k_2 \leq k_1 - 1 \leq k$ be given. Then, for all $\epsilon > 0$,*

$$\left(1+1/k\right)^{k-k_1}\epsilon + \epsilon/3k < \left(1+1/k\right)^{k-k_2}\epsilon.$$

This further implies that $\left(1+1/k\right)^{k-k_1}c\epsilon + \epsilon/3k < \left(1+1/k\right)^{k-k_2}c\epsilon$ for all $c\geq 1$.

Proof of Lemma A.8. We prove the result by explicitly bounding the difference:

$$\begin{aligned}
\left(1+1/k\right)^{k-k_1}\epsilon + \epsilon/3k - \left(1+1/k\right)^{k-k_2}\epsilon &= \left(\left(1+1/k\right)^{k_2-k_1} - 1\right)\left(1+1/k\right)^{k-k_2}\epsilon + \epsilon/3k \\
&\overset{\text{(a)}}{\leq} \left(\left(1+1/k\right)^{-1} - 1\right)\left(1+1/k\right)^{k-k_2}\epsilon + \epsilon/3k \\
&= -\left(1+1/k\right)^{k-k_2}\epsilon/(1+k) + \epsilon/3k \\
&\overset{\text{(b)}}{\leq} -\epsilon/(1+k) + \epsilon/3k.
\end{aligned}$$

Here, relation (a) holds since $k_2 - k_1 \leq -1$ and $(1+1/k)\geq 1$, and relation (b) holds since $k - k_2 \geq 1$ which implies that $\left(1+1/k\right)^{k-k_2}\geq 1$. Observe that $3k > 1+k$ for $k\geq 1$ which implies that

$$-\epsilon/(1+k) + \epsilon/3k < 0$$

for $\epsilon > 0$. Thus, under the assumptions of the lemma, we have that $\left(1+1/k\right)^{k-k_1}\epsilon + \epsilon/3k - \left(1+1/k\right)^{k-k_2}\epsilon < 0$, which implies that

$$\left(1+1/k\right)^{k-k_1}\epsilon + \epsilon/3k < \left(1+1/k\right)^{k-k_2}\epsilon.$$

□

The following result is standard, so we omit the proof.

Lemma A.9 (Skolemization). *Let $\mathcal{S}$ and $\mathcal{A}$ be finite sets and Π be the set of mappings from $\mathcal{S}$ to $\mathcal{A}$. Then for any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\max_{\pi \in \Pi} \mathbb{E}[f(s, \pi(s))] = \mathbb{E}[\max_a f(s, a)]$.*

A.4. $\mathscr{I}_{\leq k}(\mathcal{I})$ is a π-System

We now prove that $\mathscr{I}_{\leq k}(\mathcal{I})$ is a π-system (that is, a set system that is closed under intersection). Importantly, this implies that if $\mathcal{I}_\star \in \mathscr{I}_{\leq k}(\mathcal{I})$, then for any $\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I})$, $\mathcal{I}_\star \cap \mathcal{I} := \mathcal{I}_{\text{en}} \in \mathscr{I}_{\leq k}(\mathcal{I})$. This fact is repeatedly being in the design and analysis of OSSR in Section 3.3.

Lemma A.10 ($\mathscr{I}_{\leq k}(\mathcal{I})$ is a π system). *For any $\mathcal{I} \in \mathscr{I}_{\leq k}$, $\mathscr{I}_{\leq k}(\mathcal{I})$ is a π-system:*

1. *$\mathscr{I}_{\leq k}(\mathcal{I})$ is non-empty.*
2. *For any $\mathcal{I}_1, \mathcal{I}_2 \in \mathscr{I}_{\leq k}(\mathcal{I})$, we have $\mathcal{I}_1 \cap \mathcal{I}_2 \in \mathscr{I}_{\leq k}(\mathcal{I})$.*

Proof of Lemma A.10. Since $\mathcal{I} \in \mathscr{I}_{\leq k}$, we have $|\mathcal{I}| \leq k$. Furthermore, it trivially holds that $\mathcal{I} \subseteq \mathcal{I}$. Thus, $\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I})$, which implies that $\mathscr{I}_{\leq k}(\mathcal{I})$ is non-empty.

We now prove the second claim. By definition, every $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I})$ has $\mathcal{I} \subseteq \mathcal{J}$. Thus, for any $\mathcal{I}_1, \mathcal{I}_2 \in \mathscr{I}_{\leq k}(\mathcal{I})$,

$$\mathcal{I} \subseteq \mathcal{I}_1 \cap \mathcal{I}_2. \tag{17}$$

Furthermore, since, both $|\mathcal{I}_1| \leq k$ and $|\mathcal{I}_2| \leq k$, we have

$$|\mathcal{I}_1 \cap \mathcal{I}_2| \leq \min\{|\mathcal{I}_1|, |\mathcal{I}_2|\} \leq k. \tag{18}$$

Combining Eq. (17) and Eq. (18) implies that $\mathcal{I}_1 \cap \mathcal{I}_2 \in \mathscr{I}_{\leq k}(\mathcal{I})$. □

Appendix B. Structural Results for ExoMDPs

B.1. Bellman Rank for the ExoMDP Setting

In this section we show that in general, the ExoMDP setting does not admit low Bellman rank (Jiang et al., 2017), which is a standard structural complexity measure that enables tractable reinforcement learning in large state spaces. We expect that similar arguments apply for the related complexity measures (Jin et al., 2021; Du et al., 2021) and other variations. We note that Efroni et al. (2021b) showed that the more general Exogenous Block MDP model does not admit low Bellman rank. Here, we show that the same conclusion holds for the specialized ExoMDP model.

Recall that Bellman rank is a complexity measure that depends on the underlying MDP and on a class of action-value functions $\mathcal{F}$ used to approximate $Q^\star$. For a policy π, denote the average Bellman error of function $f \in \mathcal{F}$ by

$$\mathcal{E}_h(\pi, f) := \mathbb{E}_{s_h \sim \pi, a_h \sim \pi_f}\left[f(s_h, a_h) - r_h - f(s_{h+1}, \pi_f(s_{h+1}))\right].$$

With $\Pi_\mathcal{F} := \{\pi_f : f \in \mathcal{F}\}$ we define $\mathcal{E}_h(\Pi_\mathcal{F}, \mathcal{F}) = \{\mathcal{E}_h(\pi, f)\}_{\pi \in \Pi_\mathcal{F}, f \in \mathcal{F}}$ as the matrix of Bellman residuals indexed by policies and value functions. The Bellman rank is defined as $\max_h \text{rank}(\mathcal{E}_h(\Pi_\mathcal{F}, \mathcal{F}))$.

Proposition B.1. *For every $d = 2^i$ for $i \in \mathbb{N}$, there exists (i) an ExoMDP with $S = 3$, $A = 2$, $H = 2$, d exogenous factors and 1 endogenous factor, and (ii) a function class $\mathcal{F}$ containing of d functions, one of which is $Q^\star$ and the rest of which induce policies that are $1/8$ sub-optimal, such that such that the Bellman rank is at least $d - 1$.*

Proof. We construct a ExoMDP with $H = 2$, $\mathcal{A} = \{1, 2\}$ (so that $A = 2$), a single endogenous factor with values in $\{1, 2, 3\}$, and d binary exogenous factors with values in $\{0, 1\}$.

Let $e_i \in \mathbb{R}^d$ denote the i^{th} standard basis element. We take the first factor to be endogenous, and construct the initial distribution, transition dynamics, and rewards as follows:

- $d_1 = \text{Unif}(\{(1, e_i)\}_{i \in [d]})$.
- $T((2, e_i) \mid (1, e_i), 1) = 1$, and $T((3, e_i) \mid (1, e_i), 2) = 1$.
- $R((2, e_i), \cdot) = 1/2$, and $R((3, e_i), \cdot) = 3/4$.

There is only a single, terminal action at states $(2, e_i), (3, e_i)$, which we suppress from the notation. It is straightforward to verify that this is an ExoMDP. Note that the optimal policy takes action 2 at the initial state, and we have $V^\star = 3/4$.

We first construct the class $\mathcal{F}$. Since d is a power of 2, there exist subsets $A_1, \ldots, A_{d-1} \subset [d]$ such that:[3]

$$\forall j \in [d-1] : |A_j| = d/2, \qquad \forall j \neq k \in [d-1] : |A_j \cap A_k| = d/4.$$

We define $\mathcal{F} = \{f_0, f_1, \ldots, f_{d-1}\}$, with $f_0 = Q^\star$ and each f_j associated with subset A_j as follows:

$$\begin{aligned} f_j((1, e_i), 2) &= 3/4, \qquad & f_j((3, e_i), \cdot) &= 3/4 \\ f_j((1, e_i), 1) &= \mathbf{1}\{i \in A_j\}, \qquad & f_j((2, e_i), \cdot) &= \mathbf{1}\{i \in A_j\} \end{aligned}$$

Observe that since there is no reward, each function has zero Bellman error at the first timestep (that is, $\mathcal{E}_1(\pi_{f_i}, f_j) = 0 \ \ \forall i, j \in \{0, \ldots, d-1\}$). On the other hand for $j, k \in [d-1]$ we have

$$\begin{aligned} \mathcal{E}_2(\pi_{f_j}, f_k) &= \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}\{i \in A_j\}(f_k((2, e_i), \cdot) - 1/2) + \mathbf{1}\{i \notin A_j\}(f_k((3, e_i), \cdot) - 3/4) \\ &= \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}\{i \in A_j\}(f_k((2, e_i), \cdot) - 1/2) \\ &= \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}\{i \in A_j \cap A_k\}(1 - 1/2) + \mathbf{1}\{i \in A_j \cap \bar{A}_k\}(0 - 1/2) \\ &= \frac{1}{2} \mathbf{1}\{j = k\}, \end{aligned}$$

where we have used that $|A_j \cap A_k| = |A_j \cap \bar{A}_k| = d/4$ when $j \neq k$. This shows that we can embed a $(d-1) \times (d-1)$ identity matrix in $\mathcal{E}_2(\Pi_{\mathcal{F}}, \mathcal{F})$, so we have $\text{rank}(\mathcal{E}_2(\Pi_{\mathcal{F}}, \mathcal{F})) \geq d - 1$. $\square$

3. This can be seen by associating the sets with rows of a Walsh matrix.

B.2. Structural Results for State Occupancies

In this section we provide structural results concerning the state occupancy measures in the ExoMDP model. These results refine certain results derived for the more general EX-BMDP model in Efroni et al. (2021b).

For the first result, we adopt the shorthand

$$d_h^\pi\left(s[\mathcal{I}]\right) := d_h\left(s[\mathcal{I}]\ ;\ \pi\right) := \mathbb{P}_\pi(s_h[\mathcal{I}] = s[\mathcal{I}]).$$

Lemma B.1 (Decoupling of state occupancy measures)**.** *Fix $t, h \in [H]$ such that $t \le h$. Let $\pi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$ be an endogenous policy and let $\mathcal{I}$ be any factor set. Then for any $s'\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$ and $s \in \mathcal{S}, a \in \mathcal{A}$ the following claims hold.*

1. $d_h^\pi\left(s'[\mathcal{I}] \mid s_t = s, a_t = a\right) = d_h^\pi\left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_t[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_t = a\right) \cdot d_h\left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_t[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right).$
2. $d_h^\pi\left(s'[\mathcal{I}] \mid s_t = s\right) = d_h^\pi\left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_t[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) \cdot d_h\left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_t[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right).$
3. *For any endogenous mixture policy $\mu \in \Pi_{\mathrm{mix}}[\mathcal{I}_\star]$ and factor set $\mathcal{I}$,*

$$d_h^\mu(s[\mathcal{I}]) = d_h^\mu(s[\mathcal{I}_{\mathrm{en}}]) \cdot d_h(s[\mathcal{I}_{\mathrm{ex}}]).$$

Hence, the random variables $(s_h[\mathcal{I}_{\mathrm{en}}], s_h[\mathcal{I}_{\mathrm{ex}}])$ are independent under μ.

Proof of Lemma B.1. The proof follows a simple backwards induction argument.

Proof of Claims 1 and 2. We prove the two claims by induction on $t' = h-1, .., t$.

Base case: $t' = h - 1$. The base case holds as an immediate consequence of the ExoMDP structure. In more detail, we have the following results.

1. **Claim 1.**

$$\begin{aligned}
&d_h^\pi\left(s'[\mathcal{I}] \mid s_{h-1} = s, a_{h-1} = a\right)\\
&= \sum_{s'[\mathcal{I}^{\mathrm{c}}] \in \mathcal{S}[\mathcal{I}^{\mathrm{c}}]} T(s'[\mathcal{I}] \mid s, a)\\
&= \sum_{s'[\mathcal{I}_\star \setminus \mathcal{I}_{\mathrm{en}}] \in \mathcal{S}[\mathcal{I}_\star \setminus \mathcal{I}_{\mathrm{en}}]} \sum_{s'[\mathcal{I}_\star^{\mathrm{c}} \setminus \mathcal{I}_{\mathrm{ex}}] \in \mathcal{S}[\mathcal{I}_\star^{\mathrm{c}} \setminus \mathcal{I}_{\mathrm{ex}}]} T(s'[\mathcal{I}_\star] \mid s[\mathcal{I}_\star], a) T(s'[\mathcal{I}_\star^{\mathrm{c}}] \mid s[\mathcal{I}_\star^{\mathrm{c}}])\\
&= \sum_{s[\mathcal{I}_\star \setminus \mathcal{I}_{\mathrm{en}}] \in \mathcal{S}[\mathcal{I}_\star \setminus \mathcal{I}_{\mathrm{en}}]} T(s'[\mathcal{I}_\star] \mid s[\mathcal{I}_\star], a) \sum_{s[\mathcal{I}_\star^{\mathrm{c}} \setminus \mathcal{I}_{\mathrm{ex}}] \in \mathcal{S}[\mathcal{I}_\star^{\mathrm{c}} \setminus \mathcal{I}_{\mathrm{ex}}]} T(s'[\mathcal{I}_\star^{\mathrm{c}}] \mid s[\mathcal{I}_\star^{\mathrm{c}}])\\
&= d_h^\pi\left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{h-1}[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_{h-1} = a\right) d_h\left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{h-1}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right). \qquad (19)
\end{aligned}$$

2. **Claim 2.**

$$\begin{aligned}
&d_h^\pi\left(s'[\mathcal{I}] \mid s_{h-1} = s\right)\\
&\overset{\text{(a)}}{=} \sum_{a \in \mathcal{A}} d_h^\pi\left(s'[\mathcal{I}] \mid s_{h-1} = s, a_{h-1} = a\right) \pi_{h-1}(a \mid s\left[\mathcal{I}_\star\right])\\
&\overset{\text{(b)}}{=} \sum_{a \in \mathcal{A}} d_h^\pi\left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{h-1}[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_{h-1} = a\right) d_h\left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{h-1}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right) \pi_{h-1}(a \mid s\left[\mathcal{I}_\star\right])
\end{aligned}$$

$$= d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{h-1}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right) \sum_{a\in\mathcal{A}} d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{h-1}[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_{h-1} = a\right) \pi_{h-1}(a \mid s\,[\mathcal{I}_\star])$$
$$\overset{(c)}{=} d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{h-1}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right) d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{h-1}[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right).$$

Here (a) holds by Bayes' rule and because $\pi \in \Pi[\mathcal{I}_\star]$ is endogenous policy, (b) holds by Eq. (19), and (c) holds by Bayes' rule and the law of total probability.

Induction step Fix $t' < h-1$ and assume the induction hypothesis holds for $t'+1$.

1. **Claim 1.**

$$\begin{aligned}
&d_h^\pi \left(s'[\mathcal{I}] \mid s_{t'} = s, a_{t'} = a\right) \\
&= \sum_{\bar{s}\in\mathcal{S}} d_h^\pi \left(s'[\mathcal{I}] \mid s_{t'+1} = \bar{s}\right) \mathbb{P}(s_{t'+1} = \bar{s} \mid s_{t'} = s, a_{t'} = a) \\
&\overset{(a)}{=} \sum_{\bar{s}\in\mathcal{S}} d_h^\pi \left(s'[\mathcal{I}] \mid s_{t'+1} = \bar{s}\right) T(\bar{s}[\mathcal{I}_\star] \mid s[\mathcal{I}_\star], a) T(\bar{s}[\mathcal{I}_\star^{\mathrm{c}}] \mid s[\mathcal{I}_\star^{\mathrm{c}}]) \\
&\overset{(b)}{=} \sum_{\bar{s}[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{t'+1}[\mathcal{I}_\star] = \bar{s}[\mathcal{I}_\star]\right) T(\bar{s}[\mathcal{I}_\star] \mid s[\mathcal{I}_\star], a) \\
&\quad \times \sum_{\bar{s}[\mathcal{I}_\star^{\mathrm{c}}]\in\mathcal{S}[\mathcal{I}_\star^{\mathrm{c}}]} d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{t'+1}[\mathcal{I}_\star^{\mathrm{c}}] = \bar{s}[\mathcal{I}_\star^{\mathrm{c}}]\right) T(\bar{s}[\mathcal{I}_\star^{\mathrm{c}}] \mid s[\mathcal{I}_\star^{\mathrm{c}}]) \\
&= d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{t'}[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_{t'} = a\right) d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{t'}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right), \qquad (20)
\end{aligned}$$

where (a) holds by the ExoMDP model assumption (Section 2), and (b) holds by the induction hypothesis.

2. **Claim 2.**

$$\begin{aligned}
&d_h^\pi \left(s'[\mathcal{I}] \mid s_{t'} = s\right) \\
&\overset{(a)}{=} \sum_{a\in\mathcal{A}} d_h^\pi \left(s'[\mathcal{I}] \mid s_{t'} = s, a_{t'} = a\right) \pi_{t'}(a \mid s\,[\mathcal{I}_\star]) \\
&\overset{(b)}{=} \sum_{a\in\mathcal{A}} d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{t'}[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_{t'} = a\right) d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{t'}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right) \pi_{t'}(a \mid s\,[\mathcal{I}_\star]) \\
&= d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{t'}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right) \sum_{a\in\mathcal{A}} d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{t'}[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_{t'} = a\right) \pi_{t'}(a \mid s\,[\mathcal{I}_\star]) \\
&\overset{(c)}{=} d_h \left(s'[\mathcal{I}_{\mathrm{ex}}] \mid s_{t'}[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right) d_h^\pi \left(s'[\mathcal{I}_{\mathrm{en}}] \mid s_{t'}[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right).
\end{aligned}$$

Here (a) holds by Bayes' rule and because $\pi \in \Pi[\mathcal{I}_\star]$ is endogenous policy, (b) holds by Eq. (20), and (c) holds by Bayes' rule and law of total probability.

This proves the induction step and both claims.

Proof of Claim 3. We first prove the claim holds for $\pi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$. That is, for any $\pi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$, factor set $\mathcal{I}$ and $s[\mathcal{I}]$, we have

$$d_h^\pi(s[\mathcal{I}]) = d_h^\pi(s[\mathcal{I}_{\mathrm{en}}]) \cdot d_h(s[\mathcal{I}_{\mathrm{ex}}]). \tag{21}$$

This yields the result, since for $\mu \in \Pi_{\mathrm{mix}}[\mathcal{I}_\star]$, Eq. (21) implies that

$$\begin{aligned} d_h^\mu(s[\mathcal{I}]) &= \mathbb{E}_{\pi\sim\mu}[d_h^\pi(s[\mathcal{I}])] \\ &= \mathbb{E}_{\pi\sim\mu}[d_h^\pi(s[\mathcal{I}_{\mathrm{en}}]) \cdot d_h(s'[\mathcal{I}_{\mathrm{ex}}])] \\ &= \mathbb{E}_{\pi\sim\mu}[d_h^\mu(s[\mathcal{I}_{\mathrm{en}}])]d_h(s[\mathcal{I}_{\mathrm{ex}}]) = d_h^\mu(s[\mathcal{I}_{\mathrm{en}}]) \cdot d_h(s[\mathcal{I}_{\mathrm{ex}}]). \end{aligned}$$

We now prove Eq. (21). Fix $\pi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$, and observe that

$$\begin{aligned} d_h^\pi(s) &\overset{\text{(a)}}{=} \mathbb{E}_{s_1\sim d_1}\left[d_h^\pi\left(s \mid s_1\right)\right] \\ &\overset{\text{(b)}}{=} \mathbb{E}_{s_1\sim d_1}\left[d_h^\pi\left(s[\mathcal{I}_\star] \mid s_1[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right) d_h^\pi\left(s[\mathcal{I}_\star^{\mathrm{c}}] \mid s_1[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right)\right] \\ &\overset{\text{(c)}}{=} \mathbb{E}_{s_1[\mathcal{I}_\star]\sim d_h}\left[d_h^\pi\left(s[\mathcal{I}_\star] \mid s_1[\mathcal{I}_\star] = s[\mathcal{I}_\star]\right)\right] \mathbb{E}_{s_1[\mathcal{I}_\star^{\mathrm{c}}]\sim d_h}\left[d_h^\pi\left(s[\mathcal{I}_\star^{\mathrm{c}}] \mid s_1[\mathcal{I}_\star^{\mathrm{c}}] = s[\mathcal{I}_\star^{\mathrm{c}}]\right)\right] \\ &= d_h^\pi(s\left[\mathcal{I}_\star\right])d_h(s\left[\mathcal{I}_\star^{\mathrm{c}}\right]). \end{aligned} \tag{22}$$

Relation (a) holds by the tower property, and relation (b) holds by the second claim of the lemma, because π is an endogenous policy. Relation (c) holds because $s_1[\mathcal{I}_\star]$ and $s_1[\mathcal{I}_\star^{\mathrm{c}}]$ are independent (by the ExoMDP model assumption, we have $d_1(s) = d_1(s[\mathcal{I}_\star])d_1(s[\mathcal{I}_\star^{\mathrm{c}}])$).

The relation in Eq. (22) now implies the result:

$$\begin{aligned} d_h^\pi(s[\mathcal{I}]) &\overset{\text{(a)}}{=} \sum_{s[\mathcal{I}_\star\setminus\mathcal{I}_{\mathrm{en}}]\in\mathcal{S}[\mathcal{I}_\star\setminus\mathcal{I}_{\mathrm{en}}]} \sum_{s[\mathcal{I}_\star^{\mathrm{c}}\setminus\mathcal{I}_{\mathrm{ex}}]\in\mathcal{S}[\mathcal{I}_\star^{\mathrm{c}}\setminus\mathcal{I}_{\mathrm{ex}}]} d_h^\pi(s) \\ &\overset{\text{(b)}}{=} \sum_{s[\mathcal{I}_\star\setminus\mathcal{I}_{\mathrm{en}}]\in\mathcal{S}[\mathcal{I}_\star\setminus\mathcal{I}_{\mathrm{en}}]} \sum_{s[\mathcal{I}_\star^{\mathrm{c}}\setminus\mathcal{I}_{\mathrm{ex}}]\in\mathcal{S}[\mathcal{I}_\star^{\mathrm{c}}\setminus\mathcal{I}_{\mathrm{ex}}]} d_h^\pi(s\left[\mathcal{I}_\star\right])d_h(s\left[\mathcal{I}_\star^{\mathrm{c}}\right]) \\ &= \sum_{s[\mathcal{I}_\star\setminus\mathcal{I}_{\mathrm{en}}]\in\mathcal{S}[\mathcal{I}_\star\setminus\mathcal{I}_{\mathrm{en}}]} d_h^\pi(s\left[\mathcal{I}_\star\right]) \sum_{s[\mathcal{I}_\star^{\mathrm{c}}\setminus\mathcal{I}_{\mathrm{ex}}]\in\mathcal{S}[\mathcal{I}_\star^{\mathrm{c}}\setminus\mathcal{I}_{\mathrm{ex}}]} d_h(s\left[\mathcal{I}_\star^{\mathrm{c}}\right]) \\ &= d_h^\pi(s'[\mathcal{I}_{\mathrm{en}}])d_h(s'[\mathcal{I}_{\mathrm{ex}}]), \end{aligned}$$

where (a) holds by the law of total probability and (b) holds by Eq. (22). □

Lemma B.2 (Restriction lemma). *Fix $h, t \in [H]$ where $t \leq h-1$. Let $\mu \in \Pi_{\mathrm{mix}}[\mathcal{I}_\star]$ and $\rho \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$ be endogenous policies. Let $\mathcal{J}$ and $\mathcal{I}$ be two factor sets. Then, for all $s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$ it holds that*

$$\max_{\pi\in\Pi[\mathcal{J}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) = \max_{\pi\in\Pi[\mathcal{J}_{\mathrm{en}}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right).$$

Let us briefly sketch the proof. To begin, we marginalize over the factor set $\mathcal{J}^{\mathrm{c}} := [d] \setminus \mathcal{J}$ at layer t. We then show that if μ and ρ are endogenous policies, then for all $\pi \in \Pi$ and $s\left[\mathcal{I}\right] \in \mathcal{S}[\mathcal{I}]$,

$$d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) = \mathbb{E}_{s_t\sim d_t(s[\mathcal{J}]\ ;\ \pi)}\left[f(s_t\left[\mathcal{J}_{\mathrm{en}}\right], \pi\left(s_t\left[\mathcal{J}\right]\right))\bar{g}(s_t\left[\mathcal{J}_{\mathrm{ex}}\right])\right] \tag{23}$$

where both f and $\bar{g}$ are maps to $\mathbb{R}_+$. We observe that the policy

$$\pi_f(s\left[\mathcal{J}_{\text{en}}\right]) \in \underset{a}{\operatorname{argmax}}\, f(s\left[\mathcal{J}_{\text{en}}\right], \pi\left(s\left[\mathcal{J}\right]\right))$$

also maximizes Eq. (23). The result follows by observing that $\pi_f \in \Pi[\mathcal{J}_{\text{en}}]$.
Proof of Lemma B.2. Fix $s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$. The following relations hold.

$$\begin{aligned}
&d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\\
&\overset{\text{(a)}}{=} \mathbb{E}_{s[\mathcal{J}]\sim d_t(\cdot\ ;\ \mu)}\left[\mathbb{E}_{s[\mathcal{J}^c]\sim d_t(\cdot|s_t[\mathcal{J}]=s[\mathcal{J}]\ ;\ \mu)}\left[d_h\left(s\left[\mathcal{I}\right]\mid s_t = s\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\right]\right]\\
&\overset{\text{(b)}}{=} \mathbb{E}_{s[\mathcal{J}]\sim d_t(\cdot\ ;\ \mu)}\left[\mathbb{E}_{s[\mathcal{J}^c]\sim d_t(\cdot|s_t[\mathcal{J}]=s[\mathcal{J}]\ ;\ \mu)}\left[d_h\left(s\left[\mathcal{I}\right]\mid s_t = s\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\right]\right],
\end{aligned} \tag{24}$$

where (a) holds by the tower property, and (b) holds by the Markov assumption of the dynamics: conditioning on the full state s at timestep t, the future is independent of the history.

$$\begin{aligned}
(\star) &:= d_h\left(s\left[\mathcal{I}\right]\mid s_t = s\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right),\\
(\star\star) &:= \mathbb{E}_{s[\mathcal{J}^c]\sim d_t(\cdot|s_t[\mathcal{J}]=s[\mathcal{J}]\ ;\ \mu)}\left[d_h\left(s\left[\mathcal{I}\right]\mid s_t = s\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\right].
\end{aligned}$$

Analysis of term $(\star)$. Let $\pi \in \Pi[\mathcal{J}]$. Fix $s \in \mathcal{S}$ at the t^{th} timestep, and observe that $a = \pi(s\left[\mathcal{J}\right])$ is also fixed, since the policy π is a deterministic function of $s\left[\mathcal{J}\right]$.

$$\begin{aligned}
&d_h\left(s\left[\mathcal{I}\right]\mid s_t = s\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\\
&\overset{\text{(a)}}{=} d_h\left(s\left[\mathcal{I}\right]\mid s_t = s, a_t = \pi\left(s[\mathcal{J}]\right)\ ;\ \rho\right)\\
&\overset{\text{(b)}}{=} \underbrace{d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\mid s_t[\mathcal{I}_\star] = s[\mathcal{I}_\star], a_t = \pi\left(s[\mathcal{J}]\right)\ ;\ \rho\right)}_{=:\bar{f}(s_t[\mathcal{I}_\star],\pi(s[\mathcal{J}]))}\cdot\underbrace{d_h\left(s\left[\mathcal{I}_{\text{ex}}\right]\mid s_t[\mathcal{I}_\star^{\text{c}}] = s[\mathcal{I}_\star^{\text{c}}]\right)}_{=:\bar{g}(s_t[\mathcal{I}_\star^{\text{c}}])}.
\end{aligned} \tag{25}$$

Relation (a) holds by the Markov property for the MDP, and relation (b) holds by the first statement of Lemma B.1, which shows the the endogenous and exogenous state factors are decoupled; note that the assumptions of Lemma B.1 hold because ρ is endogenous policy and $a = \pi\left(s[\mathcal{J}]\right)$ is fixed. In addition, both $\bar{f}(\cdot)$ and $\bar{g}(\cdot)$ are mappings to $\mathbb{R}_+$.

Analysis of term $(\star\star)$. We consider term $(\star\star)$ and analyze it by marginalizing over the state factors not contained in $s\left[\mathcal{J}\right]$. Observe that $d_t\left(s[\mathcal{J}^{\text{c}}]\mid s_t\left[\mathcal{J}\right] = s\left[\mathcal{J}\right]\ ;\ \mu\right)$ also factorizes between the endogenous and exogenous factors due to decoupling lemma (Lemma B.1, Claim 3):

$$\begin{aligned}
&d_t\left(s[\mathcal{J}^{\text{c}}]\mid s_t\left[\mathcal{J}\right] = s\left[\mathcal{J}\right]\ ;\ \mu\right)\\
&= d_t\left(s[\mathcal{I}_\star \setminus \mathcal{J}_{\text{en}}]\mid s_t\left[\mathcal{J}_{\text{en}}\right] = s\left[\mathcal{J}_{\text{en}}\right]\ ;\ \mu\right) d_t\left(\mathcal{I}_\star^{\text{c}} \setminus \mathcal{J}_{\text{ex}}\mid s_t\left[\mathcal{J}_{\text{ex}}\right] = s\left[\mathcal{J}_{\text{ex}}\right]\right).
\end{aligned} \tag{26}$$

Hence, we have

$$\begin{aligned}
&\mathbb{E}_{s[\mathcal{J}^c]\sim d_t(\cdot|s_t[\mathcal{J}]=s[\mathcal{J}]\ ;\ \mu)}\left[d_h\left(s\left[\mathcal{I}\right]\mid s_t = s\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\right]\\
&\overset{\text{(a)}}{=} \mathbb{E}_{s[\mathcal{J}^c]\sim d_t(\cdot|s_t[\mathcal{J}]=s[\mathcal{J}]\ ;\ \mu)}\left[\bar{f}(s\left[\mathcal{I}_\star\right], \pi\left(s\left[\mathcal{J}\right]\right))\bar{g}(s\left[\mathcal{I}_\star^{\text{c}}\right])\right]
\end{aligned}$$

$$\overset{\text{(b)}}{=} \underbrace{\mathbb{E}_{s[\mathcal{I}_\star \setminus \mathcal{J}_{\text{en}}] \sim d_t(\cdot | s_t[\mathcal{J}_{\text{en}}] = s[\mathcal{J}_{\text{en}}]\ ;\ \mu)} \left[\bar{f}(s\left[\mathcal{I}_\star\right], \pi\left(s\left[\mathcal{J}\right]\right))\right]}_{=: f(s[\mathcal{J}_{\text{en}}], \pi(s[\mathcal{J}]))} \underbrace{\mathbb{E}_{s[\mathcal{I}_\star^{\text{c}} \setminus \mathcal{J}_{\text{ex}}] \sim d_t(\cdot | s_t[\mathcal{J}_{\text{ex}}] = s[\mathcal{J}_{\text{ex}}])} \left[\bar{g}(s\left[\mathcal{I}_\star^{\text{c}}\right])\right]}_{=: g(s[\mathcal{J}_{\text{ex}}])}, \tag{27}$$

where (a) holds by the calculation of term $(\star)$ in Eq. (25), and (b) holds by the decoupling of the occupancy measure $d_t\left(s[\mathcal{J}^{\mathrm{c}}] \mid s_t\left[\mathcal{J}\right] = s\left[\mathcal{J}\right]\ ;\ \mu\right)$ in Eq. (26).

Combining the results. Plugging the expression in Eq. (27) back into Eq. (24) yields

$$d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) = \mathbb{E}_{s[\mathcal{J}] \sim d_t(\cdot\ ;\ \mu)}\left[f(s_t\left[\mathcal{J}_{\text{en}}\right], \pi\left(s_t\left[\mathcal{J}\right]\right)) g(s_t\left[\mathcal{J}_{\text{ex}}\right])\right]. \tag{28}$$

We conclude the proof by invoking Lemma A.7, which gives

$$\begin{aligned}
\max_{\pi \in \Pi[\mathcal{J}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) &\overset{\text{(a)}}{=} \max_{\pi \in \Pi[\mathcal{J}]} \mathbb{E}_{s[\mathcal{J}] \sim d_t(\cdot\ ;\ \mu)}\left[\left(f(s\left[\mathcal{J}_{\text{en}}\right], \pi\left(s\left[\mathcal{J}\right]\right))\right) g(s\left[\mathcal{J}_{\text{ex}}\right])\right] \\
&\overset{\text{(b)}}{=} \max_{\pi \in \Pi[\mathcal{J}_{\text{en}}]} \mathbb{E}_{s[\mathcal{J}] \sim d_t(\cdot\ ;\ \mu)}\left[\left(f(s\left[\mathcal{J}_{\text{en}}\right], \pi\left(s\left[\mathcal{J}_{\text{en}}\right]\right)) g(s\left[\mathcal{J}_{\text{ex}}\right])\right)\right] \\
&\overset{\text{(c)}}{=} \max_{\pi \in \Pi[\mathcal{J}_{\text{en}}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right).
\end{aligned}$$

Relations (a) and (c) hold by Eq. (28). Relation (b) holds by invoking Lemma A.7 with $\mathcal{X} = \mathcal{S}[\mathcal{J}_{\text{en}}]$, $\mathcal{Y} = \mathcal{S}\left[\mathcal{J}_{\text{ex}}\right], \mathcal{X} \times \mathcal{Y} = \mathcal{S}[\mathcal{J}]$, $f(x, a) = f(s[\mathcal{J}_{\text{en}}], a)$, $g(y) = g(s[\mathcal{J}_{\text{ex}}])$, $\Pi_{\mathcal{X} \times \mathcal{Y}} = \Pi[\mathcal{J}]$ and $\Pi_{\mathcal{X}} = \Pi[\mathcal{J}_{\text{en}}]$. □

The result is proven as a consequence of the restriction lemma (Lemma B.2).

Lemma B.3 (Existence of endogenous policy cover)**.** *Fix $h, t \in [H]$ with $t \leq h-1$. Let $\mu \in \Pi_{\text{mix}}[\mathcal{I}_\star]$ and $\rho \in \Pi_{\text{NS}}[\mathcal{I}_\star]$ be endogenous policies. Let $\mathcal{I}$ be a factor set and $\mathscr{I}$ be a collection of factor sets with $\mathcal{I}_\star \in \mathscr{I}$. Then for all $s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$,*

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) = \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right).$$

Proof of Lemma B.3. For all $\mathcal{J} = \mathcal{J}_{\text{en}} \cup \mathcal{J}_{\text{ex}} \in \mathscr{I}$ and $s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$, we have

$$\begin{aligned}
\max_{\pi \in \Pi[\mathcal{J}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) &\overset{\text{(a)}}{=} \max_{\pi \in \Pi[\mathcal{J}_{\text{en}}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \\
&\overset{\text{(b)}}{\leq} \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right),
\end{aligned} \tag{29}$$

where (a) holds by Lemma B.2, and (b) holds because $\Pi[\mathcal{J}_{\text{en}}] \subseteq \Pi[\mathcal{I}_\star]$ (since $\mathcal{J}_{\text{en}} \subseteq \mathcal{I}_\star$). Since Eq. (29) holds for all $\mathcal{J} \in \mathscr{I}$, we conclude that

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \leq \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right). \tag{30}$$

On the other hand, since $\Pi[\mathcal{I}_\star] \subseteq \Pi[\mathscr{I}]$ it trivially holds that

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \geq \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right). \tag{31}$$

Combining Eq. (30) and Eq. (31) yields the result. □

Consider the problem of finding a policy π that maximizes

$$d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right), \tag{32}$$

where both μ and ρ are endogenous policies. Our next result (Lemma B.4) shows that if $\widehat{\pi}$ is an endogenous policy that is approximately optimal for reaching $s\left[\mathcal{I}_{\text{en}}\right]$ in the sense that

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \leq d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) + \epsilon, \tag{33}$$

then it is also approximately optimal for Eq. (32), in the sense that

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \leq d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) + \epsilon.$$

Lemma B.4 (Optimizing for endogenous factors is sufficient). *Fix $h, t \in [H]$ with $t \leq h-1$. Let $\mu \in \Pi_{\text{mix}}, \widehat{\pi} \in \Pi$ and $\rho \in \Pi_{\text{NS}}$ be given. Let $\mathcal{I}$ be a factor set and $\mathscr{I}$ be a collection of factor sets such that $\mathcal{I}_\star \in \mathscr{I}$. Fix $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$ and assume that:*

(A1) *μ, ρ and $\widehat{\pi}$ are endogenous .*

(A2) *$\widehat{\pi}$ is approximately optimal for $s\left[\mathcal{I}_{\text{en}}\right]$:*

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \leq d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) + \epsilon.$$

Then

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) \leq d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) + \epsilon.$$

Proof of Lemma B.4. By assumption (A1), μ and ρ are endogenous policies, so Lemma B.3 yields

$$\max_{\pi \in \Pi[\mathscr{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) = \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right). \tag{34}$$

Next, we observe that the following relations hold

$$\begin{aligned}
\max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right) &\overset{\text{(a)}}{=} \left(\max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\ ;\ \mu \circ_t \pi \circ_{t+1} \rho\right)\right) d_h\left(s\left[\mathcal{I}_{\text{ex}}\right]\right) \\
&\overset{\text{(b)}}{\leq} d_h\left(s\left[\mathcal{I}_{\text{en}}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) d_h\left(s\left[\mathcal{I}_{\text{ex}}\right]\right) + \epsilon \\
&\overset{\text{(c)}}{=} d_h\left(s\left[\mathcal{I}_{\text{en}}\right], s\left[\mathcal{I}_{\text{ex}}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) + \epsilon \\
&= d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu \circ_t \widehat{\pi} \circ_{t+1} \rho\right) + \epsilon.
\end{aligned} \tag{35}$$

Relation (a) holds by Lemma B.1, as $\mu \circ_t \pi \circ_{t+1} \rho$ is an endogenous policy. Relation (b) holds by assumption (A2) and because $d_h\left(s\left[\mathcal{I}_{\text{ex}}\right]\right) \leq 1$. Relation (c) holds by Lemma B.1; note that assumptions of the lemma are satisfied because $\mu \circ_t \widehat{\pi} \circ_{t+1} \rho$ is endogenous. Combining Eq. (34) and Eq. (35) concludes the proof. □

B.3. Structural Results for Value Functions

In this section we provide a structural results concerning the values functions for endogenous policies in the ExoMDP model. These results leverage the assumption that the rewards depend only on endogenous components. We repeatedly invoke the notion of an *endogenous MDP* $\mathcal{M}_{\text{en}} = (\mathcal{S}[\mathcal{I}_\star], \mathcal{A}, T_{\text{en}}, R_{\text{en}}, H, d_{1,\text{en}})$, which corresponds to the restriction of an ExoMDP $\mathcal{M}$ to the endogenous component of the state space. Note that only endogenous policies are well-defined in the endogenous MDP. We also denote the state-action and state value functions of an endogenous policy measured in $\mathcal{M}_{\text{en}}$ as $Q^\pi_{h,\text{en}}(s[\mathcal{I}_\star], a)$, and $V^\pi_{h,\text{en}}(s[\mathcal{I}_\star])$.

Our first result is a straightforward extension of Proposition 5 in Efroni et al. (2021b). It shows that the value function for any endogenous policy in an ExoMDP is an *endogenous function* in the sense that it only depends on the endogenous state factors.

Lemma B.5 (Value functions for endogenous policies are endogenous)**.** *Let $\pi \in \Pi_{\text{NS}}[\mathcal{I}_\star]$ be an endogenous policy, and assume that the reward function is endogenous. Then, for any $t \in [H]$ and $s \in \mathcal{S}$, we have*

$$V^\pi_t(s) = V^\pi_{t,\text{en}}(s[\mathcal{I}_\star]) \textit{ and } Q^\pi_t(s,a) = Q^\pi_{t,\text{en}}(s[\mathcal{I}_\star], a),$$

where $V^\pi_{t,\text{en}}$ and $Q^\pi_{t,\text{en}}$ are value functions for π in the endogenous MDP $\mathcal{M}_{\text{en}} = (\mathcal{S}[\mathcal{I}_\star], \mathcal{A}, T_{\text{en}}, R_{\text{en}}, H, d_{1,\text{en}})$.

Proof of Lemma B.5. Let $R = \{R_h\}_{h=1}^H$ denote the reward function. We prove the result via induction. The base case $t = H$ holds by the assumption that the reward is endogenous. Next, assume the claim is correct for $t+1$, and let us prove it for t. Since R_t is endogenous, the inductive hypothesis yields

$$\begin{aligned}
&V^\pi_t(s) \\
&= \mathbb{E}_\pi\left[R_{\text{en},t}(s[\mathcal{I}_\star], \pi_t(s[\mathcal{I}_\star])) + V^\pi_{t,\text{en}+1}(s_{t+1}[\mathcal{I}_\star]) | s_t = s, a = \pi_{t+1}(s[\mathcal{I}_\star])\right] \\
&\overset{\text{(a)}}{=} R_{\text{en},t}(s[\mathcal{I}_\star], \pi_t(s[\mathcal{I}_\star])) \\
&\quad + \sum_{s'[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} T_{\text{en}}\left(s'[\mathcal{I}_\star] \mid s[\mathcal{I}_\star], \pi_{t+1}(s[\mathcal{I}_\star])\right) V^\pi_{t,\text{en}+1}(s'[\mathcal{I}_\star]) \sum_{s'[\mathcal{I}^c_\star]\in\mathcal{S}[\mathcal{I}^c_\star]} T_{\text{en}}\left(s'[\mathcal{I}^c_\star] \mid s[\mathcal{I}^c_\star]\right) \\
&\overset{\text{(b)}}{=} R_{\text{en},t}(s[\mathcal{I}_\star], \pi_t(s[\mathcal{I}_\star])) + \sum_{s'[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} T_{\text{en}}\left(s'[\mathcal{I}_\star] \mid s[\mathcal{I}_\star], \pi_{t+1}(s[\mathcal{I}_\star]))\right) V^\pi_{t,\text{en}+1}(s'[\mathcal{I}_\star]),
\end{aligned} \tag{36}$$

where (a) holds by the factorization of the transition operator (see Eq. (1)), and (b) holds by marginalizing the exogenous factors, since $\sum_{s'[\mathcal{I}^c_\star]\in\mathcal{S}[\mathcal{I}^c_\star]} T_{\text{en}}\left(s'[\mathcal{I}^c_\star] \mid s[\mathcal{I}^c_\star]\right) = 1$. Finally, observe that Eq. (36) is the precisely the value function for π in the endogenous MDP $\mathcal{M}_{\text{en}} = (\mathcal{S}[\mathcal{I}_\star], \mathcal{A}, T_{\text{en}}, R_{\text{en}}, H, d_{1,\text{en}})$, which concludes the proof. □

Lemma B.6 (Performance difference lemma for endogenous policies)**.** *Let $\pi, \pi' \in \Pi_{\text{NS}}[\mathcal{I}_\star]$ be endogenous policies. Then*

$$J(\pi) - J(\pi') = \mathbb{E}_\pi\left[\sum_{t=1}^H Q^{\pi'}_t(s_t[\mathcal{I}_\star], \pi_t(s_t[\mathcal{I}_\star])) - Q^{\pi'}_t(s_t[\mathcal{I}_\star], \pi'_t(s_t[\mathcal{I}_\star]))\right].$$

Proof of Lemma B.6. For any endogenous policy π, observe that

$$J(\pi) := \mathbb{E}_{s_1\sim d_1}[V_1^\pi(s_1)] \overset{(a)}{=} \mathbb{E}_{s_1\sim d_1}[V_1^\pi(s_1[\mathcal{I}_\star])] \overset{(b)}{=} \mathbb{E}_{s_1[\mathcal{I}_\star]\sim d_{1,\text{en}}}[V_1^\pi(s_1[\mathcal{I}_\star])] = J_{\text{en}}(\pi), \quad (37)$$

Relation (a) holds by Lemma B.5, since $J_{\text{en}}(\pi)$ is the averaged value of $V_1^\pi(s_1)$ with respect to the initial endogenous distribution. Relation (b) holds by marginalizing out $s_1[\mathcal{I}_\star^c]$, since $V_1^\pi(s_1[\mathcal{I}_\star])$ does not depend on this quantity. Using (37) and applying the standard performance difference lemma to the endogenous MDP $\mathcal{M}_{\text{en}}$ now yields

$$J(\pi) - J(\pi') = J_{\text{en}}(\pi) - J_{\text{en}}(\pi') \quad = \mathbb{E}_\pi\left[\sum_{t=1}^H Q_t^{\pi'}(s_t[\mathcal{I}_\star], \pi_t(s_t[\mathcal{I}_\star])) - Q_t^{\pi'}(s_t[\mathcal{I}_\star], \pi_t'(s_t[\mathcal{I}_\star]))\right].$$

□

Lemma B.7 (Restriction lemma for endogenous rewards)**.** *Fix $t \le h$. Let $\mu \in \Pi_{\text{mix}}[\mathcal{I}_\star]$ and $\rho \in \Pi_{\text{NS}}[\mathcal{I}_\star]$ be endogenous policies. Define*

$$V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \rho\right) := \mathbb{E}_{\mu\circ_t\pi\circ_{t+1}\rho}\left[\sum_{t'=t}^h r_{t'}\right]. \quad (38)$$

Assume that R is an endogenous reward function. Then for any factor set $\mathcal{I}$, we have

$$\max_{\pi\in\Pi[\mathcal{I}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) = \max_{\pi\in\Pi[\mathcal{I}_{\text{en}}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right).$$

To prove this result, we generalize the proof technique used in the restriction lemma for state occupancy measures (Lemma B.2).

Proof of Lemma B.7. Since $\mu \in \Pi_{\text{mix}}[\mathcal{I}_\star]$ is an endogenous policy, the occupancy measure at the t^{th} timestep factorizes. That is, by the third statement of Lemma B.1, we have that

$$d_t\left(s[\mathcal{I}]\ ;\ \mu\right) = d_t\left(s[\mathcal{I}_{\text{en}}]\ ;\ \mu\right) d_t\left(s[\mathcal{I}_{\text{ex}}]\right).$$

For each $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$, the conditional state occupancy measure factorize as well:

$$\begin{aligned} &d_t\left(s[\mathcal{I}^c] \mid s_t[\mathcal{I}] = s[\mathcal{I}]\ ;\ \mu\right) \\ &= d_t\left(s[\mathcal{I}_\star \setminus \mathcal{I}_{\text{en}}] \mid s_t[\mathcal{I}_{\text{en}}] = s[\mathcal{I}_{\text{en}}]\ ;\ \mu\right) d_t\left(s[\mathcal{I}_\star^c \setminus \mathcal{I}_{\text{ex}}] \mid s_t[\mathcal{I}_{\text{ex}}] = s[\mathcal{I}_{\text{ex}}]\right). \end{aligned} \quad (39)$$

Let $Q_{t,\text{en}}^\rho$ be the Q function on the endogenous MDP $\mathcal{M}_{\text{en}} = (\mathcal{S}[\mathcal{I}_\star], \mathcal{A}, T_{\text{en}}, R_{\text{en}}, h, d_{1,\text{en}})$ when executing policy ρ starting from timestep $t+1$. We can express the value function as follows:

$$\begin{aligned} &V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \rho\right) \\ &= \mathbb{E}_\mu[Q_t^\rho(s_t[[d]], \pi_t\left(s_t[\mathcal{I}]\right))] \\ &\overset{(a)}{=} \mathbb{E}_\mu\left[Q_{t,\text{en}}^\rho(s_t[\mathcal{I}_\star], \pi_t\left(s_t[\mathcal{I}]\right))\right] \\ &= \mathbb{E}_{s[\mathcal{I}]\sim d_t(\cdot\ ;\ \mu)}\left[\mathbb{E}_{s[\mathcal{I}^c]\sim d_t(\cdot|s_t[\mathcal{I}]=s[\mathcal{I}]\ ;\ \mu)}\left[Q_{t,\text{en}}^\rho(s[\mathcal{I}_\star], \pi_t\left(s[\mathcal{I}]\right))\right]\right] \\ &\overset{(b)}{=} \mathbb{E}_{s[\mathcal{I}]\sim d_t(\cdot\ ;\ \mu)}\left[\mathbb{E}_{s[[\mathcal{I}_\star\setminus\mathcal{I}_{\text{en}}]\sim d_t(\cdot|s_t[\mathcal{I}_{\text{en}}]=s[\mathcal{I}_{\text{en}}]\ ;\ \mu)}\left[Q_{t,\text{en}}^\rho(s[\mathcal{I}_\star], \pi_t\left(s[\mathcal{I}]\right))\right]\right]. \end{aligned} \quad (40)$$

Relation (a) holds by Lemma B.5, since ρ is an endogenous policy. Relation (b) holds by decoupling of conditional occupancy measure (Eq. (39)), and because $Q^{\rho}_{t,\mathrm{en}}\left(s[\mathcal{I}_\star], \pi_t\left(s[\mathcal{I}]\right)\right)$ does not depend on state factors in $\mathcal{I}_\star^{\mathrm{c}} \setminus \mathcal{I}_{\mathrm{ex}}$, which are marginalized out.

To proceed, define

$$f(s_t[\mathcal{I}_{\mathrm{en}}], \pi_t\left(s[\mathcal{I}]\right)) := \mathbb{E}_{s[[\mathcal{I}_\star \setminus \mathcal{I}_{\mathrm{en}}] \sim d_t(\cdot \mid s_t[\mathcal{I}_{\mathrm{en}}] = s[\mathcal{I}_{\mathrm{en}}]\,;\,\mu)}\left[Q^{\rho}_{t,\mathrm{en}}(s[\mathcal{I}_\star], \pi_t\left(s[\mathcal{I}]\right))\right].$$

With this notation, we can rewrite the expression in Eq. (40) as

$$V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \rho\right) = \mathbb{E}_{s[\mathcal{I}] \sim d_t(\cdot\,;\,\mu)}[f(s_t[\mathcal{I}_{\mathrm{en}}], \pi_t\left(s[\mathcal{I}]\right))]. \tag{41}$$

We now invoke Lemma A.7, which shows that

$$\begin{aligned}
\max_{\pi \in \Pi[\mathcal{I}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \rho\right) &\overset{(\mathrm{a})}{=} \max_{\pi \in \Pi[\mathcal{I}]} \mathbb{E}_{s[\mathcal{I}] \sim d_t(\cdot\,;\,\pi)}[f(s[\mathcal{I}_{\mathrm{en}}], \pi(s[\mathcal{I}]))] \\
&\overset{(\mathrm{b})}{=} \max_{\pi \in \Pi[\mathcal{I}_{\mathrm{en}}]} \mathbb{E}_{s[\mathcal{I}] \sim d_t(\cdot\,;\,\pi)}[f(s[\mathcal{I}_{\mathrm{en}}], \pi(s[\mathcal{I}]))] \\
&\overset{(\mathrm{c})}{=} \max_{\pi \in \Pi[\mathcal{I}_{\mathrm{en}}]} \mathbb{E}_{s[\mathcal{I}] \sim d_t(\cdot\,;\,\pi)}[f(s[\mathcal{I}_{\mathrm{en}}], \pi(s[\mathcal{I}]))]
\end{aligned}$$

Relations (a) and (c) holds by Eq. (41). Relation (b) holds by invoking Lemma A.7, with $\mathcal{X} = \mathcal{S}[\mathcal{I}_{\mathrm{en}}]$, $\mathcal{Y} = \mathcal{S}\left[\mathcal{I}_{\mathrm{ex}}\right], \mathcal{X} \times \mathcal{Y} = \mathcal{S}[\mathcal{I}]$, $f(x,a) = f(s[\mathcal{J}_{\mathrm{en}}], a), g(y) = 1$, and $\Pi_{\mathcal{X}\times\mathcal{Y}} = \Pi[\mathcal{I}]$ and $\Pi_{\mathcal{X}} = \Pi[\mathcal{I}_{\mathrm{en}}]$. $\square$

Appendix C. Noise-Tolerant Search over Endogenous Factors: Algorithmic Template

In this section we provide a general template for designing error-tolerant algorithms that search over endogenous factors sets. This template is used in both $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ and $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ (subroutines of OSSR).

Our algorithm design template, AbstractFactorSearch is presented in Algorithm 4. Let us describe the motivation. Let $\mathcal{Z}$ be an abstract "dataset" (typically, a collection of trajectories), let $\epsilon > 0$ be a precision parameter, and let $\mathrm{Condition}(\mathcal{Z}, \epsilon, \mathcal{I}) \in \{\mathsf{true}, \mathsf{false}\}$ be an abstract function defined over factor sets $\mathcal{I}$. AbstractFactorSearch addresses the problem of finding an endogenous factor set $\widehat{\mathcal{I}} \subseteq \mathcal{I}_\star$ such that

$$\mathrm{Condition}(\mathcal{Z}, C\cdot\epsilon, \widehat{\mathcal{I}}) = \mathsf{true} \tag{42}$$

for a numerical constant $C \geq 1$, assuming that the endogenous factors $\mathcal{I}_\star$ satisfy the condition themselves:

$$\mathrm{Condition}(\mathcal{Z}, \epsilon, \mathcal{I}_\star) = \mathsf{true}. \tag{43}$$

For example, within $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$, $\mathrm{Condition}(\mathcal{Z}, \epsilon, \mathcal{I})$ checks whether policies that act on the factor set $\mathcal{I}$ lead to ϵ-optimal value for a given reward function (approximated using trajectories in $\mathcal{Z}$).

AbstractFactorSearch begins with an initial set of endogenous factors $\mathcal{I}_0 \subseteq \mathcal{I}_\star$. Naturally, since $\mathcal{I}_\star \in \mathscr{I}_{\leq k}\left(\mathcal{I}_0\right)$ and $\mathcal{I}_\star$ is known to satisfy Eq. (43), a naive approach would be to enumerate over the collection $\mathscr{I}_{\leq k}\left(\mathcal{I}_0\right)$ to find a factor set $\widehat{\mathcal{I}} \in \mathscr{I}_{\leq k}\left(\mathcal{I}_0\right)$ that satisfies Eq. (42). For example, considering the following procedure:

Algorithm 4 AbstractFactorSearch

require: abstract dataset $\mathcal{Z}$, precision ϵ, initial endogenous factor $\mathcal{I}_0 \subseteq \mathcal{I}_\star$.
for $k' = |\mathcal{I}_0|, |\mathcal{I}_0| + 1, \ldots, k$ **do**
  Set $\epsilon_{k'} = (1 + 1/k)^{k-k'} \epsilon$.
  for $\mathcal{I} \in \mathscr{I}_{k'}(\mathcal{I}_0)$ **do**
    if $\mathrm{Condition}(\mathcal{Z}, \epsilon_{k'}, \mathcal{I}) =$ true **then return** $\widehat{\mathcal{I}} \leftarrow \widehat{\mathcal{I}}$.
return fail.

- For each $\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I}_0)$, check whether $\mathrm{Condition}(\mathcal{Z}, C\epsilon, \mathcal{I}) =$ true.
- If so, return $\widehat{\mathcal{I}} \leftarrow \mathcal{I}$.

It is straightforward to see that this approach returns a factor set $\widehat{\mathcal{I}} \in \mathscr{I}_{\leq k}(\mathcal{I}_0)$ that satisfies Eq. (42), but the issue is that there is nothing preventing $\widehat{\mathcal{I}}$ from containing exogenous factors. AbstractFactorSearch resolves this problem by searching for factors in a bottom-up fashion. The algorithm begins by searching over factor sets with minimal cardinality ($k' = |\mathcal{I}_0|$), and gradually increases the size until a factor set satisfying (42) is found.

In more detail, observe that we have

$$\mathscr{I}_{\leq k}(\mathcal{I}_0) = \cup_{k'=|\mathcal{I}_0|}^{k} \mathscr{I}_k(\mathcal{I}_0),$$

where

$$\mathscr{I}_k(\mathcal{I}_0) := \left\{\mathcal{I}' \subseteq [d] \mid \mathcal{I}_0 \subseteq \mathcal{I}',\ \left|\mathcal{I}'\right| = k\right\}.$$

Starting from $k' = |\mathcal{I}_0|$, AbstractFactorSearch checks whether exists a set of factors $\mathcal{I} \in \mathscr{I}_{k'}(\mathcal{I}_0)$ that satisfies $\mathrm{Condition}(\cdots)$ with respect to an accuracy parameter $\epsilon_{k'} = (1 + 1/k)^{k-k'} \epsilon$; this choice allows for larger errors for smaller k'. When a set of factors $\mathcal{I}$ satisfies Eq. (42) AbstractFactorSearch halts and returns this set; otherwise, k' is increased. For this approach to succeed, we assume that $\mathrm{Condition}$ satisfies the following property.

Assumption C.1. *For any set of factors $\mathcal{I} = \mathcal{I}_{\mathrm{en}} \cup \mathcal{I}_{\mathrm{ex}}$ with $|\mathcal{I}_{\mathrm{ex}}| \geq 1$, it holds that*

$$\mathrm{Condition}(\mathcal{Z}, \epsilon_{|\mathcal{I}|}, \mathcal{I}) = \textit{true} \implies \mathrm{Condition}(\mathcal{Z}, \epsilon_{|\mathcal{I}_{\mathrm{en}}|}, \mathcal{I}_{\mathrm{en}}) = \textit{true}. \tag{44}$$

We now describe three key steps used to prove that this scheme succeeds.

1. *AbstractFactorSearch does not return fail.* This follows immediately from the assumption that (43) is satisfied.
2. *AbstractFactorSearch returns an endogenous set of factors.* Observe that the assumption $\mathcal{I}_\star \in \mathscr{I}_{\leq k}(\mathcal{I}_0)$ implies that for any $\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I}_0)$, $\mathcal{I}_{\mathrm{en}} := \mathcal{I}_\star \cap \mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I}_0)$; this follows from Lemma A.10. Hence, if $\mathcal{I}$ satisfies Eq. (42), Assumption C.1 implies that $\mathcal{I}_{\mathrm{en}}$ satisfies Eq. (42) as well. Since AbstractFactorSearch scans $\mathscr{I}_{\leq k}(\mathcal{I}_0)$ in a bottom-up fashion, this means it must return an endogenous factor set, since it will verify that $\mathcal{I}_{\mathrm{en}}$ satisfies Eq. (42) prior to $\mathcal{I}$.
3. *AbstractFactorSearch is near-optimal.* Since $(1 + 1/k)^{k-k'} \epsilon \leq 3\epsilon$ for all $k' \in [k]$, the factor set $\widehat{\mathcal{I}}$ returned by AbstractFactorSearch satisfies $\mathrm{Condition}(\mathcal{Z}, 3\epsilon, \widehat{\mathcal{I}}) =$ true.

Part II
Omitted Subroutines

Appendix D. Finding a Near-Optimal Endogenous Policy: EndoPolicyOptimization

Algorithm 5 EndoPolicyOptimization$^{\epsilon}_{t,h}$: One-Step Endogenous Policy Optimization

// Find an endogenous policy $\pi \in \Pi[\mathscr{I}_{\leq k}]$ that approximately maximizes $V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$, where $\mu \in \Pi_{\text{mix}}$ and $\psi \in \Pi_{\text{NS}}$ are fixed policies.

1: **require:**
- Starting timestep t, end timestep h, and target precision $\epsilon \in (0,1)$.
- Collection $\left\{\widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)\right\}_{\pi \in \Pi[\mathscr{I}_{\leq k}]}$ of estimates for $V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$ for all $\pi \in \Pi[\mathscr{I}_{\leq k}]$.

2: **for** $k' = 0, 1, \cdots, k$ **do**

3: Let $\epsilon_{k'} = (1 + 1/k)^{k-k'}\,\epsilon$.

4: **for** $\mathcal{I} \in \mathscr{I}_{k'}$ **do**

5: Set is_cover = true if

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) \leq \max_{\pi \in \Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{k'}.$$

6: **if** is_cover = true **then return:** $\widehat{\pi} \in \operatorname{argmax}_{\pi \in \Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$.

7: **return:** fail.

In this section, we introduce and analyze the EndoPolicyOptimization$^{\epsilon}_{t,h}$ algorithm (Algorithm 5), which is used in the optimization phase of OSSR$^{\epsilon,\delta}_{h}$ (Appendix G) and in ExoPSDP (Appendix F). In Appendix D.1 we give a high-level description and intuition for the algorithm, and in Appendix D.2 we prove the main theorem regarding its correctness and sample complexity.

D.1. Description of EndoPolicyOptimization.

The goal of EndoPolicyOptimization$^{\epsilon}_{t,h}$ is to return a policy $\widehat{\pi} \in \Pi[\mathcal{I}]$ such that:

1. $\widehat{\pi}$ is endogenous in the sense that $\widehat{\pi} \in \Pi[\mathcal{I}]$ for some $\mathcal{I} \subseteq \mathcal{I}_{\star}$.

2. $\widehat{\pi}$ is near-optimal in the sense that

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) \leq V_{t,h}\left(\mu \circ_t \widehat{\pi} \circ_{t+1} \psi\right) + O\left(\epsilon\right),$$

   where $V_{t,h}\left(\pi\right) := \mathbb{E}_{\pi}\left[\sum_{t'=t}^{h} r_t\right]$ for a given reward function R.

EndoPolicyOptimization assumes access to approximate value functions $\widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$ that are ϵ-close to the true value functions $V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$. Given these approximate value functions, finding a near-optimal policy is trivial; it suffices to take the empirical maximizer $\widehat{\pi} \in$

$\mathrm{argmax}_{\pi\in\Pi[\mathscr{I}_{\leq k}]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi)$. However, finding a near-optimal *endogenous* policy is a more challenging task. For this, EndoPolicyOptimization applies the abstract endogenous factor search scheme described in Appendix C (AbstractFactorSearch), which regularizes toward factors with smaller cardinality.

EndoPolicyOptimization$^{\epsilon}_{t,h}$ splits the set $\mathscr{I}_{\leq k}$ as $\mathscr{I}_{\leq k} = \cup_{k'=0}^{k} \mathscr{I}_{k'}$, where $\mathscr{I}_{k'}$ is the collection of factor sets with cardinality exactly $k' \in [k]$, and follows the bottom-up search strategy in AbstractFactorSearch. Beginning from $k' = 0, \ldots, k$, the algorithm checks whether there exists a near-optimal policy in the class $\Pi[\mathscr{I}_{k'}]$. If such a policy is found, the algorithm returns it, and otherwise it proceeds to $k' + 1$.

Intuition for correctness. We prove the correctness of the EndoPolicyOptimization$^{\epsilon}_{t,h}$ procedure by following the general template in Appendix C. In particular, we view EndoPolicyOptimization$^{\epsilon}_{t,h}$ as a special case of the AbstractFactorSearch (Algorithm 4) scheme with

$$\mathrm{Condition}(\mathcal{Z}, \epsilon, \mathcal{I}) = \mathbb{1}\left\{ \max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) \leq \max_{\pi\in\Pi[\mathcal{I}]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) + \epsilon \right\}.$$

Most the effort in proving the correctness of the algorithm is in showing that this condition satisfies Assumption C.1. In particular, we need to show that if some $\mathcal{I} \in \mathscr{I}_{\leq k}$ satisfies the condition in Line 5,

$$\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) \leq \max_{\pi\in\Pi[\mathcal{I}]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) + \epsilon_{|\mathcal{I}|},$$

then $\mathcal{I}_{\mathrm{en}} := \mathcal{I} \cap \mathcal{I}_{\star}$ also satisfies the condition in the sense that

$$\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) \leq \max_{\pi\in\Pi[\mathcal{I}_{\mathrm{en}}]} \widehat{V}_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) + \epsilon_{|\mathcal{I}_{\mathrm{en}}|}.$$

This can be shown to hold as a consequence of assumptions (A1) and (A2) in Theorem D.1. Assumption (A1) asserts the following restriction property holds: For any $\mathcal{I}$,

$$\max_{\pi\in\Pi[\mathcal{I}]} V_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi) = \max_{\pi\in\Pi[\mathcal{I}_{\mathrm{en}}]} V_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi).$$

Hence, optimizing over a larger policy class that acts on exogenous factors does not improve the value. Assumption (A2) asserts that the estimates for $V_{t,h}(\mu \circ_t \pi \circ_{t+1} \psi)$ are uniformly ϵ-close, so that optimizing with respect to these estimates is sufficient.

Importance of the decoupling property. We emphasize that assumption (A1) is non-trivial. We show it holds for several choices for the reward function in the ExoMDP (Lemma B.2 and Lemma B.7), which are used when we invoke the algorithm within OSSR. However, the condition my not hold if the endogenous and exogenous factors are correlated. In this case, optimizing over exogenous state factors may improve the value, leading the algorithm to fail.

Formal guarantee for EndoPolicyOptimization. The following result shows that EndoPolicyOptimization$^{\epsilon}_{t,h}$ returns a near-optimal endogenous policy.

Theorem D.1 (Correctness of EndoPolicyOptimization$^{\epsilon}_{t,h}$). *Fix $h \in [H]$ and $t \in [h]$. Let $\mu \in \Pi_{\mathrm{mix}}$ and $\psi \in \Pi_{\mathrm{NS}}$ be fixed policies. Assume the following conditions hold:*

(A1) Restriction property: *For any set of factors $\mathcal{I}$,*

$$\max_{\pi \in \Pi[\mathcal{I}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) = \max_{\pi \in \Pi[\mathcal{I}_{\text{en}}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right).$$

(A2) Quality of estimation. *For all $\pi \in \Pi[\mathscr{I}_{\leq k}]$,*

$$\left| V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) - \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) \right| \leq \epsilon/12k.$$

Then the policy $\widehat{\pi}$ output by $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ *satisfies the following properties:*

1. *$\widehat{\pi}$ is endogenous: $\widehat{\pi} \in \Pi[\mathcal{I}]$, where $\mathcal{I} \subseteq \mathcal{I}_\star$.*
2. *$\widehat{\pi}$ is near-optimal:* $\max_{\pi \in \Pi\left[\mathscr{I}_{\leq k}\right]} V_{t,h}\left(\mu \circ_t \pi \circ \psi\right) \leq V_{t,h}\left(\mu \circ_t \widehat{\pi} \circ \psi\right) + 4\epsilon$.

D.2. Proof of Theorem D.1

We use the three-step proof recipe described in Appendix C to prove correctness of EndoPolicyOptimization.

Step 1: $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ does not return fail. By definition, there exists $\mathcal{I} \in \mathscr{I}_{\leq k}$ such that

$$\max_{\pi \in \Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) = \max_{\pi \in \Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right).$$

Thus, Line 5 is satisfied, since $\epsilon_{k'} \geq 0$.

Step 2: $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ returns an endogenous policy. Since $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ does not return fail, it returns a policy $\widehat{\pi} \in \Pi[\mathcal{I}]$ for some factor set $\mathcal{I}$. We prove that $\mathcal{I}$ is an endogenous factor set, which implies that $\widehat{\pi}$ is an endogenous policy. We show this by proving the following claim:

Claim 1. *If $\mathcal{I}$ satisfies the condition in Line 5 (is_cover $=$ true for $\mathcal{I}$), then $\mathcal{I}_{\text{en}}$ satisfies the condition as well (is_cover $=$ true for $\mathcal{I}_{\text{en}}$).*

Given this claim, it is straightforward to see that $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ returns an endogenous policy. First, observe that for any $\mathcal{I} \in \mathscr{I}_{\leq k}$, we have $\mathcal{I}_{\text{en}} := \mathcal{I} \cap \mathcal{I}_\star \in \mathcal{I} \in \mathscr{I}_{\leq k}$ by Lemma A.10 (since $\mathcal{I}_\star \in \mathscr{I}_{\leq k}$). If $|\mathcal{I}_{\text{en}}| < |\mathcal{I}|$, then $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ verifies that $\mathcal{I}_{\text{en}} \in \mathscr{I}_{\leq k}$ satisfies Line 5 prior to verifying whether $\mathcal{I} \in \mathscr{I}_{\leq k}$ satisfies the condition. It follows that the factor set returned by the algorithm must be endogenous.

Proof of Claim 1. Assume that $\mathcal{I}$ contains at least one exogenous factor, so

$$|\mathcal{I}_{\text{en}}| \leq |\mathcal{I}| - 1. \tag{45}$$

Suppose that is_cover $=$ true for $\mathcal{I}$. By construction, it holds that for $k_1 := |\mathcal{I}| \leq k$,

$$\max_{\pi \in \Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) \leq \max_{\pi \in \Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{k_1}. \tag{46}$$

This statement, which holds for the approximate value $\widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$ implies a similar statement on the true value $V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$. Specifically, Eq. (46) together with Lemma A.6 (which can be applied using assumption (A2)), implies that

$$\begin{aligned}\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) &\leq \max_{\pi\in\Pi[\mathcal{I}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{k_1} + \epsilon/6k \\ &\overset{\text{(a)}}{=} \max_{\pi\in\Pi[\mathcal{I}_{\text{en}}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{k_1} + \epsilon/6k, \end{aligned} \tag{47}$$

and (a) holds by the restriction property in assumption (A1).

We now relate the inequality in Eq. (47), which holds for the true values $V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right)$, back to an inequality on the approximate values. Using Lemma A.6 and assumption (A2) on Eq. (47), we have htat

$$\begin{aligned}\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) &\leq \max_{\pi\in\Pi[\mathcal{I}_{\text{en}}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{k_1} + \epsilon/3k \\ &\overset{\text{(a)}}{\leq} \max_{\pi\in\Pi[\mathcal{I}_{\text{en}}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{k_2}, \end{aligned} \tag{48}$$

where (a) holds for all $k_1, k_2 \in [k]$ such that $k_2 \leq k_1 - 1$, since

$$\epsilon_{k_1} + \epsilon/3k := (1+1/k)^{k-k_1}\,\epsilon + \epsilon/3k \leq (1+1/k)^{k-k_2}\,\epsilon := \epsilon_{k_2},$$

by Lemma A.8. Setting $k_2 = |\mathcal{I}_{\text{en}}| \leq k_1 - 1 = |\mathcal{I}|$ (the cardinality of $\mathcal{I}_{\text{en}}$ is strictly smaller than that of $\mathcal{I}$ by Eq. (45)) and plugging this value into Eq. (48) yields

$$\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) \leq \max_{\pi\in\Pi[\mathcal{I}_{\text{en}}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{|\mathcal{I}_{\text{en}}|}. \tag{49}$$

Hence, $\mathcal{I}_{\text{en}}$ also satisfies the conditions in Line 5.

Step 3: $\textsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ returns a near-optimal policy. When the condition of $\textsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ at Line 5 holds and is_cover = true, the factor set $\mathcal{I}$ satisfies

$$\begin{aligned}\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) &\leq \max_{\pi\in\Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) + \epsilon_{|\mathcal{I}|} \\ &= \widehat{V}_{t,h}\left(\mu \circ_t \widehat{\pi} \circ_{t+1} \psi\right) + \epsilon_{|\mathcal{I}|} \\ &\leq \widehat{V}_{t,h}\left(\mu \circ_t \widehat{\pi} \circ_{t+1} \psi\right) + 3\epsilon, \end{aligned} \tag{50}$$

where the last relation holds because $\epsilon_{|\mathcal{I}|} \leq (1+1/k)^k\,\epsilon \leq 3\epsilon$. Applying Lemma A.6 with (A2) then gives

$$\max_{\pi\in\Pi\left[\mathscr{I}_{\leq k}\right]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) \leq V_{t,h}\left(\mu \circ_t \widehat{\pi} \circ_{t+1} \psi\right) + \underbrace{3\epsilon + \epsilon/6k}_{\leq 4\epsilon}.$$

□

Appendix E. Selecting Endogenous Factors with Strong Coverage: EndoFactorSelection

Algorithm 6 EndoFactorSelection$^{\epsilon}_{t,h}$: Simultaneous Policy Cover for all Factors

// Find $\mathcal{I}$ such that reaching $\mathcal{I}$ implicitly leads to good coverage for all $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$.

require:

- Starting timestep t and end timestep h, target precision $\epsilon \in (0,1)$.
- Set of endogenous factors $\mathcal{I}^{(t+1,h)} \subseteq \mathcal{I}_\star$.
- Collection of policy sets $\{\Gamma^{(t)}[\mathcal{I}]\}_{\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})}$, where

$$\Gamma^{(t)}[\mathcal{I}] = \left\{ \pi^{(t)}_{s[\mathcal{I}]} \mid s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}] \right\}.$$

- Set of $(t+1 \to h)$ policies

$$\Psi^{(t+1,h)} = \left\{ \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]} \mid s[\mathcal{I}^{(t+1,h)}] \in \mathcal{S}[\mathcal{I}^{(t+1,h)}] \right\}.$$

- Collection $\widehat{\mathcal{D}}$ of approximate occupancy measures for layer h under the sampling process $\mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}$.

// Pick $\Psi \in \Gamma^{(t)}[\mathcal{I}] \circ_{t+1} \Psi^{(t+1:h)}$ that explores $\mathcal{I} \subseteq \mathcal{I}_\star$ and sufficiently explores other factors.

for $k' = |\mathcal{I}^{(t+1,h)}|, |\mathcal{I}^{(t+1,h)}| + 1, .., k$ **do**

  Define $\epsilon_{k'} = (1 + 1/k)^{k-k'} 5\epsilon$.

  for $\mathcal{I} \in \mathscr{I}_{k'}(\mathcal{I}^{(t+1,h)})$ **do**

    // Test whether reaching states in $\mathcal{I}$ leads to good coverage for all factors $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$.

    Set sufficient_cover = true if for all $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$ and for all $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$:

$$\begin{aligned} &\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s[\mathcal{J}] \; ; \; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\ &\quad \leq \widehat{d}_h\left(s[\mathcal{J}] \; ; \; \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J} \cap \mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k'}, \end{aligned} \tag{51}$$

    where $\pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \in \Gamma^{(t)}[\mathcal{J} \cap \mathcal{I}]$. // Recall $\pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \approx \operatorname{argmax}_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s[\mathcal{J} \cap \mathcal{I}] \; ; \; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right)$.

    if sufficient_cover = true **then**

      $\widehat{\mathcal{I}} \leftarrow \mathcal{I}$.

      return $(\widehat{\mathcal{I}}, \Gamma^{(t)}[\widehat{\mathcal{I}}])$.

return: fail. // Low probability failure event.

In this section, we describe and analyze the EndoFactorSelection$^{\epsilon}_{t,h}$ algorithm (Algorithm 6). EndoFactorSelection$^{\epsilon}_{t,h}$ is a subroutine used in the selection phase of OSSR$^{\epsilon,\delta}_{h}$, and generalizes the selection phase used in OSSR.Exact$_h$ to the setting where only approximate occupancy measures are available. In Appendix E.1, we give a high-level description EndoFactorSelection$^{\epsilon}_{t,h}$, give intuition, and state the main theorem concerning its performance. Then, in Appendix E.2 we prove this result.

E.1. Description of EndoFactorSelection

To motivate $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$, let us first recall the selection phase of $\mathsf{OSSR.Exact}_h$ (Line 7 of Algorithm 2). The selection phase assumes access to a collection of policy sets $\{\Gamma^{(t)}[\mathcal{I}]\}_{\mathcal{I}\in\mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})}$, which are calculated in the optimization step. In particular, for each set $\mathcal{I}$ and each $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$, $\pi^{(t)}_{s[\mathcal{I}]} \in \Gamma^{(t)}[\mathcal{I}]$ is an endogenous policy that maximizes the probability of reaching $s[\mathcal{I}]$ at layer h in the following sense:

$$\pi^{(t)}_{s[\mathcal{I}]} \in \underset{\pi\in\Pi[\mathscr{I}_{\leq k}]}{\operatorname{argmax}}\, d_h\left(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right).$$

The selection phase of $\mathsf{OSSR.Exact}_h$ find the factor set $\widehat{\mathcal{I}} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ of minimal size such that for all $\mathcal{J} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ and $s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]$,

$$\max_{\pi\in\Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) = d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}\cap\widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right). \tag{52}$$

At the end of the selection step, $\mathsf{OSSR.Exact}_h$ outputs the tuple $(\widehat{\mathcal{I}}, \Gamma^{(t)}[\widehat{\mathcal{I}}])$. Since $\widehat{\mathcal{I}}$ is chosen as the *minimal* factor set that satisfies Eq. (52) it can be shown it is an endogenous factors set. Furthermore, $\Gamma^{(t)}[\widehat{\mathcal{I}}]$ satisfies condition Eq. (52).

$\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ is similar to $\mathsf{OSSR.Exact}_h$, but only requires access to *approximate state occupancy measures*. Analogous to $\mathsf{OSSR.Exact}_h$, the algorithm outputs a tuple $(\widehat{\mathcal{I}}, \Gamma^{(t)}[\widehat{\mathcal{I}}])$, where $\widehat{\mathcal{I}}$ is an endogenous factors set and $\Gamma^{(t)}[\widehat{\mathcal{I}}]$ ensures good coverage at layer h.However, since $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ has only has access to approximate state occupancy measures, the policy set $\Gamma^{(t)}[\widehat{\mathcal{I}}]$ returned by the algorithm is only guaranteed to satisfy an approximate version of Eq. (52):

$$\begin{aligned}\max_{\pi\in\Pi[\mathscr{I}_{\leq k}]} & d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\ &\leq d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}\cap\widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + O(\epsilon),\end{aligned} \tag{53}$$

where $\pi^{(t)}_{s[\mathcal{J}\cap\widehat{\mathcal{I}}]} \in \Gamma^{(t)}[\widehat{\mathcal{I}}]$.

To ensure find an endogenous factor set $\widehat{\mathcal{I}}$ such that $\Gamma^{(t)}[\widehat{\mathcal{I}}]$ satisfies Eq. (53), $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ follows the AbstractFactorSearch scheme described in Appendix C. It enumerates the collection of factor sets $\mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ in a bottom-up fashion—starting from factor sets of minimal cardinality—and checks whether each factor set approximately satisfies the optimality condition.

Intuition for correctness. To establish the correctness of $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$, we view the algorithm as an instance of AbstractFactorSearch with

$$\begin{aligned}&\mathrm{Condition}(\mathcal{Z}, \epsilon, \mathcal{I}) \\ &= \mathbb{1}\left\{\begin{array}{ll}\displaystyle\max_{\pi\in\Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) & \\ \quad\leq \widehat{d}_h\left(s\left[\mathcal{J}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon, & \forall \mathcal{J} \in \mathscr{I}(\mathcal{I}^{(t+1,h)}), s[\mathcal{J}] \in \mathcal{S}[\mathcal{J}]\end{array}\right\},\end{aligned}$$

and recall that $\pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \in \Gamma^{(t)}[\mathcal{J}\cap\mathcal{I}]$ is the output from the optimization step at EndoPolicyOptimization. The analysis of $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ follow the recipe sketched in Appendix C. Most of our efforts are devoted to proving that the condition in Eq. (44) required by AbstractFactorSearch holds for EndoFactorSelection$^{\epsilon}_{t,h}$. In particular, we wish to prove the following claim: *If $\mathcal{I}$ satisfies the condition in Line 5 (sufficient_cover $=$ true for $\mathcal{I}$), then $\mathcal{I}_{\text{en}}$ satisfies the condition as well (sufficient_cover $=$ true for $\mathcal{I}_{\text{en}}$).* To show that the statement is true, we use a key structural result, Lemma B.4, which generalizes certain structural results used in the analysis of OSSR.Exact (Proposition 3.1). Let μ and ρ be endogenous policies, and consider a fixed state factor $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$. Lemma B.4 asserts that if an endogenous policy $\pi_{s[\mathcal{I}_{\text{en}}]}$ approximately maximizes the probability of reaching the endogenous part of $s[\mathcal{I}]$, which is given by

$$d_h\left(s[\mathcal{I}_{\text{en}}]\ ;\ \mu \circ_t \pi_{s[\mathcal{I}_{\text{en}}]} \circ_{t+1} \rho\right),$$

then the policy also approximately maximizes the probability of reaching $s[\mathcal{I}]$, which is given by

$$d_h\left(s[\mathcal{I}]\ ;\ \mu \circ_t \pi_{s[\mathcal{I}_{\text{en}}]} \circ_{t+1} \rho\right).$$

Hence, to approximately maximize the probability of reaching $s[\mathcal{I}]$, it suffices to execute a policy that approximately maximizes the probability of reaching the endogenous part of the state, $s[\mathcal{I}_{\text{en}}]$. We use this observation to show that exogenous factors are redundant in the sense that if sufficient_cover $=$ true for $\mathcal{I}$, then sufficient_cover $=$ true for $\mathcal{I}_{\text{en}}$; this proves the claim

Formal guarantee for EndoFactorSelection The following result is the main guarantee for EndoFactorSelection$^{\epsilon}_{t,h}$.

Theorem E.1 (Success of EndoFactorSelection$^{\epsilon}_{t,h}$). *Fix $h \in [H]$ and $t \in [h]$. Assume the following conditions hold:*

(A1) Endogeneity of arguments. *$\mu^{(t)} \in \Pi_{\text{mix}}[\mathcal{I}_\star]$ is endogenous, $\Psi^{(t+1,h)}$ contains only endogenous policies, and $\Gamma^{(t)}[\mathcal{I}]$ contains only endogenous policies for all $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$. In addition, $\mathcal{I}^{(t+1,h)} \subseteq \mathcal{I}_\star$.*

(A2) Quality of estimation. *$\widehat{\mathcal{D}}$ is a collection of $\epsilon/12k$-approximate state occupancy measures with respect to $\left(\mu^{(t)} \circ \Pi[\mathscr{I}_{\leq k}] \circ \Psi^{(t+1,h)}, \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h\right)$ (Definition A.1).*

(A3) Optimality for $\Gamma^{(t)}[\mathcal{I}]$. *For any factor set $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ and any $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$, the policy $\pi^{(t)}_{s[\mathcal{I}]} \in \Gamma^{(t)}\left[\mathcal{I}\right]$ satisfies the following optimality guarantee:*

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) \leq d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) + 4\epsilon.$$

Then EndoFactorSelection$^{\epsilon}_{t,h}$ does not output fail, and the tuple $(\widehat{\mathcal{I}}, \Gamma^{(t)}[\widehat{\mathcal{I}}])$ output by the algorithm satisfies the following guarantees:

1. *$\widehat{\mathcal{I}} \subseteq \mathcal{I}_\star$.*

2. *For all $s\left[\mathcal{I}_\star\right] \in \mathcal{S}\left[\mathcal{I}_\star\right]$, we have*

$$\max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) - d_h\left(s\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \pi^{(t)}_{s[\widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) \leq 16\epsilon,$$

   where we note that we can write $s\left[\mathcal{I}_\star\right] = \left(s\left[\widehat{\mathcal{I}}\right], s\left[\mathcal{I}_\star \setminus \widehat{\mathcal{I}}\right]\right) = \left(s\left[\mathcal{I}^{(t+1,h)}\right], s\left[\mathcal{I}_\star \setminus \mathcal{I}^{(t+1,h)}\right]\right)$ because $\mathcal{I}^{(t+1,h)}, \widehat{\mathcal{I}} \subseteq \mathcal{I}_\star$.

E.2. Proof of Theorem E.1

We use the three-step proof strategy described in Appendix C to prove correctness for $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$.

Step 1: $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ does not return fail. We show that given assumptions $(\mathrm{A1})$ – $(\mathrm{A3})$ $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ does not return fail. First, observe that $\mathcal{I}_\star \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$, since $\mathcal{I}^{(t+1,h)} \subseteq \mathcal{I}_\star$ by $(\mathrm{A1})$ and $|\mathcal{I}_\star| \leq k$ by assumption. We prove that $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ halts for $\mathcal{I} \leftarrow \mathcal{I}_\star$; meaning that $\mathcal{I}_\star$ satisfies the condition at Line 5 of $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$.

Fix $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ and $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$. Let $\mathcal{I}_{\mathrm{en}} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$[4] be the endogenous component of $\mathcal{I}$, so that $s[\mathcal{I}] = (s[\mathcal{I}_{\mathrm{en}}], s[\mathcal{I}_{\mathrm{ex}}])$. Consider the policy $\pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]} \in \Gamma^{(t)}[\mathcal{I}_{\mathrm{en}}]$. By assumption $(\mathrm{A3})$, $\pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]}$ is endogenous and satisfies

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s\left[\mathcal{I}_{\mathrm{en}}\right] \;;\; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) \leq d_h\left(s\left[\mathcal{I}_{\mathrm{en}}\right] \;;\; \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) + 4\epsilon. \tag{54}$$

Eq. (54) shows that $\pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]}$ has near-optimal probability for the endogenous component of $s[\mathcal{I}]$ near optimally (when the rollout policy $\psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}$ is fixed). Combined with the fact that both $\pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]}$ and $\psi^{(t+1,h)}_{s[\mathcal{I}_{\mathrm{en}}]}$ are endogenous (by $(\mathrm{A1})$), this allows us to apply Lemma B.4, which asserts that $\pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]}$ reaches the any state factor $s[\mathcal{I}]$ with $\mathcal{I}_{\mathrm{en}} \subseteq \mathcal{I}$ near-optimally as well. In particular,

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s\left[\mathcal{I}\right] \;;\; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) \leq d_h\left(s\left[\mathcal{I}\right] \;;\; \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) + 4\epsilon. \tag{55}$$

Now, observe that since $\widehat{\mathcal{D}}$ is $\epsilon/12k$-approximate with respect to $\left(\Pi\left[\mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)\right], \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h\right)$ (cf. $(\mathrm{A2})$), Eq. (55) and Lemma A.6 imply that

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s\left[\mathcal{I}\right] \;;\; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) \leq \widehat{d}_h\left(s\left[\mathcal{I}\right] \;;\; \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}_{\mathrm{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) + 5\epsilon. \tag{56}$$

Since $\mathcal{I}_{\mathrm{en}} = \mathcal{I} \cap \mathcal{I}_\star$, and since

$$5\epsilon \leq (1 + 1/k)^{k-k'+1}\, 5\epsilon := \epsilon_{k'}$$

for all $k' \in [k]$, this implies that the condition at Line 5 of $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ is satisfied by $\mathcal{I}_\star$.

Step 2: Proof of first claim ($\widehat{\mathcal{I}} \subseteq \mathcal{I}_\star$ is a set of endogenous factors). Since $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ does not return fail, it necessarily returns a pair $(\widehat{\mathcal{I}}, \Gamma^{(t)}[\widehat{\mathcal{I}}])$. We now show that $\widehat{\mathcal{I}}$ is endogenous. To do so, we prove the following claim.

4. $\mathcal{I}_{\mathrm{en}} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ since $\mathcal{I}_\star \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ and $\mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ is a π-system by Lemma A.10.

Lemma E.1. *If $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ satisfies the condition in Line 5 (sufficient_cover = true for $\mathcal{I}$), then $\mathcal{I}_{\text{en}}$ satisfies the condition as well (sufficient_cover = true for $\mathcal{I}_{\text{en}}$).*

Conditioned on Lemma E.1, the result quickly follows. Observe that for any $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$, we have $\mathcal{I}_{\text{en}} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$[4]. Furthermore, if $|\mathcal{I}_{\text{en}}| > |\mathcal{I}|$, then EndoFactorSelection will check whether $\mathcal{I}_{\text{en}}$ satisfies the condition in Line 5 prior to checking whether $\mathcal{I}$ satisfies it. Thus, EndoFactorSelection necessarily returns a set of endogenous factors; it remains to prove Lemma E.1.

Proof of Lemma E.1. Fix $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ with $\mathcal{I}_{\text{ex}} \neq \varnothing$. Assume that $\mathcal{I}$ satisfies the conditions in Line 5. That is, for $k_1 := |\mathcal{I}| \leq k$, it holds that for all $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$ and all $s\left[\mathcal{J}\right] = \left(s\left[\mathcal{I}\right], s\left[\mathcal{J} \setminus \mathcal{I}\right]\right) \in \mathcal{S}\left[\mathcal{J}\right]$,

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \leq \widehat{d}_h\left(s\left[\mathcal{J}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_1}, \tag{57}$$

where $\pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \in \Gamma^{(t)}[\mathcal{J} \cap \mathcal{I}]$. We will show that this implies that $\mathcal{I}_{\text{en}}$ also satisfies the conditions in Line 5.

$\mathcal{I}_{\text{en}}$ satisfies the conditions in Line 5. Since $\mathcal{I}$ satisfies Eq. (57) for all $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$, it must also satisfy the condition for all $\mathcal{J}_{\text{en}} \subseteq \mathcal{J}$. Fix $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$. Then for all $s[\mathcal{J}_{\text{en}}] \in \mathcal{S}[\mathcal{J}_{\text{en}}]$, we have

$$\begin{aligned}
&\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s[\mathcal{J}_{\text{en}}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\
&\leq \widehat{d}_h\left(s\left[\mathcal{J}_{\text{en}}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}_{\text{en}}\cap\mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_1} \\
&\overset{\text{(a)}}{\leq} \widehat{d}_h\left(s\left[\mathcal{J}_{\text{en}}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}_{\text{en}}\cap\mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_1},
\end{aligned} \tag{58}$$

where (a) follows because $\mathcal{J}_{\text{en}} \cap \mathcal{I} = \mathcal{J}_{\text{en}} \cap \mathcal{I}_{\text{en}}$.

Since (A2) asserts that $\widehat{\mathcal{D}}$ is $\epsilon/12k$-approximate with respect to $\left(\Pi\left[\mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)\right], \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h\right)$, we can relate the inequality above to the analogous inequality for the true occupancies using Lemma A.6. After multiplying both sides by $d_h\left(s[\mathcal{J}_{\text{ex}}]\right) \in [0,1]$, this yields

$$\begin{aligned}
&d_h\left(s[\mathcal{J}_{\text{ex}}]\right) \max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{J}_{\text{en}}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\
&\leq d_h\left(s[\mathcal{J}_{\text{ex}}]\right) d_h\left(s[\mathcal{J}_{\text{en}}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}_{\text{en}}\cap\mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_1} + \epsilon/6k.
\end{aligned} \tag{59}$$

We now manipulate both sides Eq. (58) to relate these quantities to the occupancy measure for $s[\mathcal{J}]$. This is done by appealing to the decoupling property for occupancy measures of endogenous policies (Appendix B.2). To begin, for the left-hand side of Eq. (59), we have

$$\begin{aligned}
&d_h\left(s[\mathcal{J}_{\text{ex}}]\right) \max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{J}_{\text{en}}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\
&\overset{\text{(a)}}{=} d_h\left(s[\mathcal{J}_{\text{ex}}]\right) \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s[\mathcal{J}_{\text{en}}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right)
\end{aligned}$$

$$\begin{aligned} &\overset{\text{(b)}}{=} \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\ &\overset{\text{(c)}}{=} \max_{\pi \in \Pi[\mathscr{I}_{\le k}]} d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right), \end{aligned} \tag{60}$$

where relations (a) and (c) hold by Lemma B.3 and relation (b) holds by Lemma B.1; note that the assumptions of these lemmas hold because $\mu^{(t)}$ and $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}$ are assumed to be endogenous, and because $\pi \in \Pi[\mathcal{I}_\star]$ is also endogenous. Moving on, we analyze the right-hand side of Eq. (59). We have

$$\begin{aligned} &d_h\left(s[\mathcal{J}_{\text{ex}}]\right) d_h\left(s[\mathcal{J}_{\text{en}}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}_{\text{en}} \cap \mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\ &\quad = d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}_{\text{en}} \cap \mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right), \end{aligned} \tag{61}$$

by Lemma B.1 (the assumptions of the lemma hold because $\mu^{(t)}, \pi^{(t)}_{s[\mathcal{J}_{\text{en}} \cap \mathcal{I}_{\text{en}}]}$ and $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}$ are endogenous). Plugging Eq. (61) and Eq. (60) back into Eq. (59), we have that

$$\begin{aligned} &\max_{\pi \in \Pi[\mathscr{I}_{\le k}]} d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\ &\le d_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}_{\text{en}} \cap \mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_1} + \epsilon/6k. \end{aligned} \tag{62}$$

It remains to relate this to the analogous inequality for the approximate occupancy measures. Since $\widehat{\mathcal{D}}$ is $\epsilon/12k$-approximate with respect to $\left(\Pi\left[\mathscr{I}_{\le k}\left(\mathcal{I}^{(t+1,h)}\right)\right], \mathscr{I}_{\le k}\left(\mathcal{I}^{(t+1,h)}\right), h\right)$ by (A2), Lemma A.6, and Eq. (62) imply that

$$\begin{aligned} &\max_{\pi \in \Pi[\mathscr{I}_{\le k}]} \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \\ &\le \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J} \cap \mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_1} + \epsilon/3k \\ &\overset{\text{(a)}}{\le} \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J} \cap \mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{k_2}, \end{aligned} \tag{63}$$

where (a) holds for all $k_1, k_2 \in [k]$ such that $k_2 \le k_1 - 1$, since

$$\epsilon_{k_1} + \epsilon/3k := (1 + 1/k)^{k-k_1} 5\epsilon + \epsilon/3k \le (1 + 1/k)^{k-k_2} 5\epsilon := \epsilon_{k_2}$$

by Lemma A.8 (with $c = 5$). Since $|\mathcal{I}_{\text{en}}| < |\mathcal{I}| := k_1$, we can set $k_2 = |\mathcal{I}_{\text{en}}|$ in Eq. (63), which implies that

$$\max_{\pi \in \Pi[\mathscr{I}_{\le k}]} \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) \le \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J} \cap \mathcal{I}_{\text{en}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + \epsilon_{|\mathcal{I}_{\text{en}}|}. \tag{64}$$

Since Eq. (64) holds for all $\mathcal{J} \in \mathscr{I}_{\le k}(\mathcal{I}^{(t+1,h)})$ and $s\left[\mathcal{J}\right] \in \mathcal{S}[\mathcal{J}]$, this yields the result. □

Step 3: Proof of second claim ($\Gamma^{(t)}[\widehat{\mathcal{I}}]$ is near-optimal). This claim is a direct consequence of the condition in Line 5. Let $\widehat{\mathcal{I}}$ be the output of $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$. Since $\mathsf{sufficient\_cover} = \mathsf{true}$, then the conditions at Line 5 are satisfied, and for all $\mathcal{J} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$, for all $s\left[\mathcal{J}\right] = \left(s\left[\mathcal{I}^{(t+1)}\right], s\left[\mathcal{J} \setminus \mathcal{I}^{(t+1,h)}\right]\right) \in \mathcal{S}\left[\mathcal{J}\right]$:

$$\begin{aligned}
&\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} \widehat{d}_h\left(s[\mathcal{J}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right)\\
&\qquad \leq \widehat{d}_h\left(s\left[\mathcal{J}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{J}\cap\widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + 15\epsilon, \qquad (65)
\end{aligned}$$

where $\pi^{(t)}_{s[\mathcal{J}\cap\mathcal{I}]} \in \Gamma^{(t)}[\mathcal{J}\cap\mathcal{I}]$; the upper bound holds because $\epsilon_{k'} := (1+1/k)^{k-k'}\, 5\epsilon \leq 15\epsilon$ for all $k' \in [k]$. Applying Eq. (65) with $\mathcal{J} \leftarrow \mathcal{I}_\star \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$, and using Lemma A.6 (which is admissible by assumption (A2)), we have that for all $s\left[\mathcal{I}_\star\right] \in \mathcal{S}\left[\mathcal{I}_\star\right]$,

$$\begin{aligned}
&\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{I}_\star]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right)\\
&\qquad \leq d_h\left(s\left[\mathcal{I}_\star\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}_\star\cap\widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + 16\epsilon\\
&\qquad \overset{\text{(a)}}{\leq} d_h\left(s\left[\mathcal{I}_\star\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\widehat{\mathcal{I}}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) + 16\epsilon,
\end{aligned}$$

where (a) holds because $\mathcal{I}_\star \cap \widehat{\mathcal{I}} = \widehat{\mathcal{I}}$, since $\widehat{\mathcal{I}} \subseteq \mathcal{I}_\star$ by the first claim. $\square$

Appendix F. PSDP with Exogenous Information: ExoPSDP

Algorithm 7 ExoPSDP: PSDP with Exogenous Information

require:
- Target precision $\epsilon \in (0,1)$ and failure probablitity $\delta \in (0,1)$.
- Collection $\{\Psi^{(h)}\}_{h=2}^{H}$ of endogenous $\eta/2$-approximate policy covers.

initialize:
- Let $N = C \cdot AS^{4k}H^2k^3 \log\left(\frac{dSAH}{\delta}\right)\epsilon^{-2}$ for sufficiently large constant $C > 0$ and $\epsilon_0 = \frac{\epsilon}{2S^kH}$.
- For all $t \in [H]$, define $\mu^{(t)} := \mathrm{Unf}\left(\Psi^{(t)}\right)$.
- Let $\widehat{\pi}^{(H,H)} = \varnothing$.

for $t = H-1, .., 1$ **do**

/* Estimate average value functions via importance weighting. */

Get dataset $\left\{\left(s_{t,n}, a_{t,n}, \{r_{t',n}\}_{t'=1}^{H}\right)\right\}_{n=1}^{N}$ by executing $\mu^{(t)} \circ_t \mathrm{Unf}(\mathcal{A}) \circ_{t+1} \widehat{\pi}^{(t+1,H)}$.

Estimate the $(t \to H)$ value for all $\pi \in \Pi[\mathscr{I}_{\leq k}]$ via importance weighting:

$$\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) = \frac{1}{N}\sum_{n=1}^{N} \frac{\mathbb{1}\left\{a_{t,n} = \pi\left(s_{t,n}\right)\right\}}{1/A}\left(\sum_{t'=t}^{H} r_{t',n}\right).$$

/* Apply policy optimization with estimated value functions. */

$\widehat{\pi}^{(t)} \leftarrow \mathsf{EndoPolicyOptimization}_{t,h}^{\epsilon_0}\left(\left\{\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)\right\}_{\pi \in \Pi\left[\mathscr{I}_{\leq k}\right]}\right)$.

$\widehat{\pi}^{(t,H)} = \widehat{\pi}^{(t)} \circ_{t+1} \widehat{\pi}^{(t+1,H)}$.

return: $\widehat{\pi}^{(1,H)}$.

In this section we present and analyze the ExoPSDP algorithm (Algorithm 7). ExoPSDP is based on the classical PSDP algorithm (Bagnell et al., 2004), but incorporates modifications to ensure that the policies produced are endogenous. In Appendix F.1, we motivate ExoPSDP and state the main guarantee concerning its performance (Theorem F.1). Then, in Appendix F.2, we prove this result.

F.1. Description of ExoPSDP

The ExoPSDP algorithm solves the following problem:

> *Given a collection of endogenous policy covers* $\{\Psi^{(t)}\}_{t=1}^{H}$ *for an ExoMDP* $\mathcal{M}$*, find a policy* $\widehat{\pi}$ *that is* ϵ*-optimal in the sense that* $J(\widehat{\pi}) \geq \max_\pi J(\pi) - \epsilon$.

To motivate the approach behind the algorithm, we first remind the reader of the classical PSDP algorithm.

Background on PSDP. Suppose we have a set of mixture policies $\{\mu^{(h)}\}_{h=1}^{H}$ that ensure good coverage at every layer for an MDP $\mathcal{M}$, and our goal is to optimize the MDP's reward function. The PSDP algorithm (Bagnell et al., 2004) addresses this problem by using the dynamic programming principle to learn a near-optimal policy through a series of backward steps $t = H, \ldots, 1$. Assume access to a policy class Π. At each step t, assuming that step $t+1$ has already produced a near-optimal $(t+1) \to H$ policy $\widehat{\pi}^{(t+1,H)}$, the algorithm estimates the value function $V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ for all $\pi \in \Pi$ where (see also Eq. (38))

$$V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) := \mathbb{E}_{\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}}\left[\sum_{t'=t}^{H} r_{t'}\right].$$

The estimates are calculated via importance-weighting by

$$\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) = \frac{1}{N}\sum_{n=1}^{N} \frac{\mathbb{1}\left\{a_{t,n} = \pi\left(s_{t,n}\right)\right\}}{1/A}\left(\sum_{t'=t}^{H} r_{t',n}\right)$$

where the data is generated by rolling in with $\mu^{(t)}$, taking random action on the t^{th} time-step and rolling out with $\widehat{\pi}_{t+1:H}$ using N trajectories. Then, PSDP computes

$$\pi^{(t)} \in \underset{\pi \in \Pi}{\operatorname{argmax}}\, \widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right), \tag{66}$$

and sets $\widehat{\pi}^{(t,h)} = \pi^{(t)} \circ_t \widehat{\pi}^{(t+1,H)}$. The final policy $\widehat{\pi} := \widehat{\pi}^{(1,H)}$ is guaranteed to be near-optimal as long as $\{\mu^{(h)}\}_{h=1}^{H}$ have good coverage.

Insufficiency of vanilla PSDP. The first issue with applying PSDP to the ExoMDP model is that, if we want the policy class Π to contain all possible policies, we will have $|\Pi| = \Theta(A^{S^d})$, which leads to sample complexity scaling with $\log|\Pi| = \Omega(\text{poly}(S^d))$; this is prohibitively large. An alternative policy class one my hope can address this issue is $\Pi[\mathscr{I}_{\leq k}]$. Indeed, this class has much smaller cardinality: $|\Pi[\mathscr{I}_{\leq k}]| = \Theta(d^k A^{S^k})$. However, for an ExoMDP, naively optimizing over this class via Eq. (66) may lead to roll-out policies $\widehat{\pi}_{t+1:H}$ that depend on the exogenous state factors, since there is no mechanism in place to ensure endogeneity. This in turn may invalidate the *realizability assumption* needed to apply standard PSDP (see Misra et al. (2020), Assumption 2). In particular, PSDP requires that the policy class Π contains the optimal policy in the sense that

$$\max_{\pi \in \Pi_{\text{NS}}} V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) = \max_{\pi \in \Pi} V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right). \tag{67}$$

If the roll-out policy $\widehat{\pi}_{t+1:H}$ depends on the exogenous state factors, then the optimal policy that maximizes $V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ may depend on exogenous state factors as well. Then, Eq. (67) may be violated when instantiating PSDP with the policy class $\Pi[\mathscr{I}_{\leq k}]$.

A solution: ExoPSDP. To address the issues above, ExoPSDP applies an alternative to the optimization step in (66). In particular, ExoPSDP uses the sub-routine EndoPolicyOptimization (see Line 6), which finds an *endogenous* near-optimal policy. In particular, as long as $\widehat{\pi}^{(t+1,H)}$ is endogenous, which can be guaranteed inductively, EndoPolicyOptimization, will succeed in finding an endogenous policy at step t. Importantly, since (i) the reward in a ExoMDP depends only on the endogenous factors, and (ii) the policy $\widehat{\pi}^{(t+1,H)}$ is endogenous (by the guarantees of EndoPolicyOptimization), $\widehat{\pi}^{(t)}$ can be shown to be near-optimal with respect to the entire policy class Π. Hence, in spite of optimizing over the restricted policy class $\mathscr{I}_{\leq k}$, we are able to find a near-optimal policy with respect set of all policies. Using this argument inductively allows us to prove that $\widehat{\pi}^{(1,H)}$ is near-optimal and endogenous.

Theorem F.1 (Main guarantee for ExoPSDP). *Suppose that the sets $\{\Psi^{(t)}\}_{t=1}^{H}$ passed into ExoPSDP are endogenous $\eta/2$-approximate policy covers for all t. Then, for any $\epsilon, \delta > 0$, with probability at least $1-\delta$,*

1. $\widehat{\pi}^{(1,H)}$ *is endogenous.*
2. $\widehat{\pi}^{(1,H)}$ *is ϵ-optimal in the sense that*

$$\max_{\pi\in\Pi_{\mathrm{NS}}} J(\pi) \leq J(\widehat{\pi}^{(1,H)}) + \epsilon.$$

Furthermore, the algorithm uses at most $N = O\left(\frac{AH^4k^3S^{3k}\log\left(\frac{dAH}{\delta}\right)}{\epsilon^2}\right)$ trajectories.

F.2. Proof of Theorem F.1

Fix a pair of endogenous policies $\pi, \widehat{\pi} \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$. Further, let $\mathcal{M}_{\mathrm{en}} = \left(\mathcal{S}, \mathcal{A}, T_{\mathrm{en}}, R_{s[\mathcal{I}_\star]}, H, d_{1,\mathrm{en}}\right)$ denote the restriction of the ExoMDP to its endogenous component, and let $Q^{\pi}_{t,\mathrm{en}}(s[\mathcal{I}_\star], a)$ denote the associated state-action value function for $\mathcal{M}_{\mathrm{en}}$.

We decompose the difference in performance as follows.

$$\begin{aligned}
&J(\pi) - J(\widehat{\pi}) \\
&\overset{(a)}{=} \sum_{t=1}^{h} \mathbb{E}_{\pi}\left[Q^{\widehat{\pi}}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \pi_t(s_t\left[\mathcal{I}_\star\right]) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \widehat{\pi}_t(s_t\left[\mathcal{I}_\star\right]))\right] \\
&\leq \sum_{t=1}^{h} \mathbb{E}_{s[\mathcal{I}_\star]\sim d_t(\cdot\,;\,\pi)}\left[\max_a Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \widehat{\pi}_t(s\left[\mathcal{I}_\star\right]))\right] \\
&\leq \sum_{t=1}^{h} \sum_{s[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} \max_{\pi\in\Pi[\mathcal{I}_\star]} d_t(s\left[\mathcal{I}_\star\right]\,;\,\pi)\left(\max_a Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \widehat{\pi}_t(s\left[\mathcal{I}_\star\right]))\right). \\
&\overset{(b)}{\leq} 2S^k \sum_{t=1}^{h} \sum_{s[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} d_t(s\left[\mathcal{I}_\star\right]\,;\,\mu^{(t)})\left(\max_a Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \widehat{\pi}_t(s\left[\mathcal{I}_\star\right]))\right). \\
&= 2S^k \sum_{t=1}^{h} \mathbb{E}_{\mu^{(t)}}\left[\max_a Q^{\widehat{\pi}}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], a) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \widehat{\pi}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right]))\right]. \\
&\overset{(c)}{=} 2S^k \sum_{t=1}^{h} \max_{\pi'\in\Pi[\mathcal{I}_\star]} \mathbb{E}_{\mu^{(t)}}\left[Q^{\widehat{\pi}}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \pi'\left(s_t\left[\mathcal{I}_\star\right]\right)) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \widehat{\pi}_t(s_t\left[\mathcal{I}_\star\right]))\right]. \\
&= 2S^k \sum_{t=1}^{h} \max_{\pi'\in\Pi[\mathcal{I}_\star]} V_{t,H}\left(\mu^{(t)} \circ_t \pi' \circ_{t+1} \widehat{\pi}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \widehat{\pi}^{(t)} \circ_{t+1} \widehat{\pi}\right).
\end{aligned} \tag{68}$$

The key steps above are justified as follows:

- Relation (a) holds by the performance difference lemma for endogenous policies (Lemma B.6), since both $\pi, \widehat{\pi} \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$ by assumption.

- Relation (b) holds because

$$\frac{\max_{\pi\in\Pi[\mathcal{I}_\star]} d_t\left(\cdot\,;\,\pi\right)}{d_t\left(\cdot\,;\,\mu^{(t)}\right)} \leq 2S^k,$$

which is a consequence of Lemma A.2. In particular, we use that (i) $\{\Psi^{(t)}\}_{t=1}^{h-1}$ are endogenous $\eta/2$-approximate policy covers, (ii) for all states, either $\max_{\pi\in\Pi[\mathcal{I}_\star]} d_t\left(s[\mathcal{I}_\star]\,;\,\pi\right) \geq \eta$ or $\max_{\pi\in\Pi[\mathcal{I}_\star]} d_t\left(s[\mathcal{I}_\star]\,;\,\pi\right) = 0$ (by the reachability assumption), and (iii)

$$\max_a Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\widehat{\pi}}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \widehat{\pi}_t(s\left[\mathcal{I}_\star\right])) \geq 0.$$

- Relation (c) holds by the skolemization principle (Lemma A.9).

Let $\mathcal{G}_{\mathsf{ExoPSDP}}$ denote the success event for Lemma F.1 (stated and proven in the sequel), which is the event in which for all $t\in[H]$, $\mathsf{EndoPolicyOptimization}^{\epsilon_0}_{t,h}$ returns a policy $\widehat{\pi}^{(t)}$ such that

1. $\widehat{\pi}^{(t)}$ is endogenous.
2. $\widehat{\pi}^{(t)}$ is near-optimal in the following sense:

$$\max_{\pi'\in\Pi[\mathcal{I}_\star]} V_{t,H}\left(\mu^{(t)} \circ_t \pi' \circ_{t+1} \widehat{\pi}^{(t+1,H)}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \widehat{\pi}^{(t)} \circ_{t+1} \widehat{\pi}^{(t+1,H)}\right) \leq \epsilon_0. \tag{69}$$

Lemma F.1 asserts that $\mathcal{G}_{\mathsf{ExoPSDP}}$ holds with probability at least $1-\delta$ whenever $N = \Omega\left(\frac{AH^2k^3S^k\log\left(\frac{dAH}{\delta}\right)}{\epsilon_0^2}\right)$. Conditioning on $\mathcal{G}_{\mathsf{ExoPSDP}}$, it follows immediately that $\widehat{\pi}^{(1,H)}$ is endogenous. To show that the policy is near-optimal, we apply Eq. (68) with $\widehat{\pi} = \widehat{\pi}^{(1,H)}$ and bound each term in the sum using Eq. (69). Maximizing over $\pi\in\Pi_{\mathrm{NS}}[\mathcal{I}_\star]$ yields

$$\max_{\pi\in\Pi[\mathcal{I}_\star]} J(\pi) - J(\widehat{\pi}) \leq 2S^kH\epsilon_0 = \epsilon,$$

by the choice $\epsilon_0 := \epsilon/2S^kH$. Finally, by the fact that $\max_{\pi\in\Pi[\mathcal{I}_\star]} J(\pi) = \max_{\pi\in\Pi_{\mathrm{NS}}} J(\pi)$, which holds because the reward is endogenous (Efroni et al. (2021b), Proposition 5), we conclude the proof. □

F.3. Computational Complexity of ExoPSDP

We now show that ExoPSDP can be implemented with computational complexity of

$$O\left(d^kNS^kAH\right),$$

where N is the number of trajectories. The main computational bottleneck of ExoPSDP occurs at Line 5 of $\mathsf{EndoPolicyOptimization}^{\epsilon_0}_{t,h}$. There, we need to optimize over $\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ estimated by the empirical averages (Line 5) for all $\mathcal{I}\in\mathscr{I}_k$. Meaning,

$$\max_{\pi\in\Pi[\mathcal{I}]} \widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$$

To sketch how to do this efficiently, we first show how to optimize over the set $\Pi[\mathcal{I}]$ when a factor set $\mathcal{I}$ is fixed. We show that instead of enumerating over all policies, one can optimize $\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ as follows. Observe that

$$\widehat{V}_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) = \sum_{s[\mathcal{I}]\in\mathcal{S}[\mathcal{I}]} \widehat{Q}_{t,h}^{\mu^{(t)}\circ_t\pi\circ_{t+1}\widehat{\pi}_{t+1:H}}\left(s[\mathcal{I}], \pi\left(s[\mathcal{I}]\right)\right),$$

where we note that $|\mathcal{S}[\mathcal{I}]| \le S^k$, and where

$$\widehat{Q}_{t,h}^{\mu^{(t)}\circ_t\pi\circ_{t+1}\widehat{\pi}_{t+1:H}}\left(s[\mathcal{I}], a\right) := \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\{s_t[\mathcal{I}] = s[\mathcal{I}], a_t = a\}\left(\sum_{t'=t}^{h} r_{n,t'}\right).$$

To maximize $\widehat{V}_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ it suffices to maximize each individual function $\widehat{Q}_{t}^{\mu^{(t)}\circ_t\pi\circ_{t+1}\widehat{\pi}_{t+1:H}}\left(s[\mathcal{I}], a\right)$. Letting

$$\widehat{\pi}_{\mathcal{I}}\left(s[\mathcal{I}]\right) \in \operatorname*{argmax}_a \widehat{Q}_{t}^{\mu^{(t)}\circ_t\pi\circ_{t+1}\widehat{\pi}_{t+1:H}}\left(s[\mathcal{I}], a\right),$$

we have that

$$\max_{\pi\in\Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) = \widehat{V}_{t,h}\left(\mu \circ_t \widehat{\pi}_{\mathcal{I}} \circ_{t+1} \psi\right).$$

Furthermore, observe that $\widehat{\pi}_{\mathcal{I}}\left(s[\mathcal{I}]\right) \in \Pi[\mathcal{I}]$.

This shows that it is possible to solve $\max_{\pi\in\Pi[\mathcal{I}]} \widehat{V}_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ with computational complexity $O\left(NS^kA\right)$. Since $\mathsf{EndoPolicyOptimization}_{t,h}^{\epsilon}$ optimizes over all possible factor sets $\mathcal{I} \in \mathscr{I}_{\le k}$ where $|\mathscr{I}_{\le k}| = O\left(d^k\right)$ for H times the total computational complexity is $O\left(d^kNS^kAH\right)$.

F.4. Application of EndoPolicyOptimization within ExoPSDP

In this section we state and prove Lemma F.1, which shows that the application of EndoPolicyOptimization within ExoPSDP (Line 6) is admissible, in the sense that the preconditions required by the algorithm are satisfied.

Lemma F.1 (Guarantees of EndoPolicyOptimization for ExoPSDP)**.** *Let precision parameter $\epsilon \in (0,1)$ and failure probability $\delta \in (0,1)$ be given. Assume that the mixture policies $\mu^{(t)} \in \Pi_{\text{mix}}$ used in Algorithm 7 are endogenous for all t. Then, if $N = \Omega\left(\frac{AH^2k^3S^k\log\left(\frac{dAH}{\delta}\right)}{\epsilon^2}\right)$ trajectories are used for each layer, we have that with probability at least $1-\delta$, for all t:*

1. *$\widehat{\pi}^{(t)}$ is an endogenous policy.*
2. *$\widehat{\pi}^{(t)}$ is near-optimal in the sense that*

$$\max_{\pi'\in\Pi[\mathcal{I}_\star]} V_{t,H}\left(\mu^{(t)} \circ_t \pi' \circ_{t+1} \widehat{\pi}^{(t+1,H)}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \widehat{\pi}^{(t)} \circ_{t+1} \widehat{\pi}^{(t+1,H)}\right) \le 4\epsilon.$$

Proof of Lemma F.1. Let $\mathcal{G}^{(t)}$ denote the event in which

1. $\widehat{\pi}^{(t)}$ is an endogenous policy.

2. $\widehat{\pi}^{(t)}$ is near optimal:

$$\max_{\pi' \in \Pi[\mathcal{I}_\star]} V_{t,H}\left(\mu^{(t)} \circ_t \pi' \circ_{t+1} \widehat{\pi}^{(t+1,H)}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \widehat{\pi}^{(t)} \circ_{t+1} \widehat{\pi}^{(t+1,H)}\right) \leq 4\epsilon.$$

We will prove that for any $\delta > 0$,

$$\mathbb{P}\left(\mathcal{G}^{(t)} \mid \cap_{t'=t+1}^{H} \mathcal{G}^{(t')}\right) \geq 1 - \delta, . \tag{70}$$

as long at least $\Omega\left(\frac{AH^2k^3S^k \log\left(\frac{dAH}{\delta}\right)}{\epsilon^2}\right)$ trajectories are used at layer t. Whenever Eq. (70) holds, Lemma A.4 implies that

$$\mathbb{P}\left(\cap_{t=1}^{H} \mathcal{G}^{(t')}\right) \geq 1 - H\delta, \tag{71}$$

and scaling $\delta \leftarrow \delta/H$ concludes the proof.

We now prove that Eq. (70) holds. To do so, we apply Theorem D.1 and verify that assumptions (A1) and (A2) required by it hold.

(A1) Conditioning on the event $\cap_{t'=t+1}^{H} \mathcal{G}^{(t')}$, we have that $\widehat{\pi}^{(t+1,H)}$ is an endogenous policy. In addition $\mu^{(t)}$ is an endogenous policy and the reward function is endogenous by assumption. Thus, the conditions of Lemma B.7 are satisfied, and the restriction property holds:

$$\max_{\pi \in \Pi[\mathcal{I}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right) = \max_{\pi \in \Pi[\mathcal{I}_{\text{en}}]} V_{t,h}\left(\mu \circ_t \pi \circ_{t+1} \psi\right).$$

(A2) The proof of this result uses similar arguments to Lemma A.5 . Fix $\pi \in \Pi[\mathscr{I}_{\leq k}]$ and observe that $\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$ is an unbiased estimator for $V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)$, and is bounded by AH. Using Lemma A.3 and following the same steps as in the proof of Lemma A.5, we have that with probability at least $1 - \delta$,

$$\begin{aligned}&\left|\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)\right| \\ &\leq O\left(\sqrt{\frac{AH^2 \log\left(\frac{1}{\delta}\right)}{N}} + \frac{AH \log\left(\frac{1}{\delta}\right)}{N}\right).\end{aligned}$$

Taking a union bound over all $\pi \in \Pi[\mathscr{I}_{\leq k}]$ and using that $|\Pi[\mathscr{I}_{\leq k}]| \leq O\left(d^{k+1} A^{S^k}\right)$, we have that with probability at least $1 - \delta$,

$$\begin{aligned}&\left|\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)\right| \\ &\leq O\left(\sqrt{\frac{AH^2 k S^k \log\left(\frac{dA}{\delta}\right)}{N}} + \frac{AHkS^k \log\left(\frac{dA}{\delta}\right)}{N}\right).\end{aligned}$$

Hence, setting $N = \Omega\left(\frac{AH^2k^3S^k \log\left(\frac{dA}{\delta}\right)}{\epsilon^2}\right)$ and using that $\epsilon^2 \leq \epsilon$ for $\epsilon \in (0, 1)$, we have that with probability at least $1 - \delta$, for all $\pi \in \Pi[\mathscr{I}_{\leq k}]$,

$$\left|\widehat{V}_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right) - V_{t,H}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H}\right)\right| \leq \frac{\epsilon}{12k}.$$

□

Part III
Additional Details and Proofs for Main Results

Appendix G. OSSR Description and Proof of Theorem 3.1

In this section we present and analyze the full $\mathsf{OSSR}_h^{\epsilon,\delta}$ algorithm (Algorithm 8). The algorithm may be thought of as a sample-based version of the OSSR.Exact algorithm described in Section 3.2. While OSSR.Exact assumes exact access to state occupancy measures, $\mathsf{OSSR}_h^{\epsilon,\delta}$ estimates the occupancy measures in a data-driven fashion, which introduces the need to account for statistical errors.

This section is organized as follows. First, in Appendix G.1 we give a high-level overview of the algorithm design principles behind $\mathsf{OSSR}_h^{\epsilon,\delta}$. Then, in Appendix G.2, we prove the main result concerning its performance, Theorem 3.1. Appendices G.3 and G.4 contain proofs for supporting results used in the proof of Theorem 3.1.

G.1. OSSR: Algorithm Overview

The $\mathsf{OSSR}_h^{\epsilon,\delta}$ algorithm follows the same template as OSSR.Exact: For each $h \in [H]$, given policy covers $\Psi^{(1)}, \dots, \Psi^{(h-1)}$, the algorithm builds a policy cover $\Psi^{(h)}$ for layer h in a backwards fashion using dynamic programming. There are two differences from the exact algorithm. First, we only have sample access to the underlying ExoMDP, the algorithm estimates the relevant occupancy measures for each backward step using Monte Carlo rollouts. Second, the optimization and selection phases from OSSR.Exact are replaced by error-tolerant variants given by the subroutines EndoPolicyOptimization and EndoFactorSelection (Algorithm 5 in Appendix D and Algorithm 6 in Appendix E, respectively).

State occupancy estimation. In order to apply dynamic programming in the same fashion as OSSR.Exact, each backward step $1 \le t \le h-1$ of $\mathsf{OSSR}_h^{\epsilon,\delta}$ proceeds by building estimates for the layer-h occupancies $d_h(s[\mathcal{I}] \ ; \ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)})$ for all $\mathcal{I} \in \mathscr{I}_{\le k}$, $\pi \in \Pi[\mathscr{I}_{\le k}]$ and $\psi^{(t+1,h)} \in \Psi^{(t+1,h)}$. This is accomplished through Monte Carlo: We gather trajectories by running $\mu^{(t)}$ up to layer t, sampling $a_t \sim \mathrm{Unf}(\mathcal{A})$ uniformly, then sampling $\psi^{(t+1,h)} \sim \mathrm{Unf}(\Psi^{(t+1,h)})$ and using it to roll out from layer $t+1$ to h. We then build estimates by importance weighting the empirical frequencies. We appeal to uniform convergence to ensure that the estimated occupancies are uniformly close for all $\mathcal{I} \in \mathscr{I}_{\le k}$ and $\pi \in \Pi[\mathscr{I}_{\le k}]$; this argument critically uses that $|\Psi^{(t+1,h)}| \le S^k$ and $\log|\Pi[\mathscr{I}_{\le k}]| \le O\left(kS^k \log(dA)\right)$, as well as the fact that we only require convergence for factors of size at most k.

Error-tolerant backward state refinement. Given the estimated state occupancy measures above, each backward step $1 \le t \le h-1$ of $\mathsf{OSSR}_h^{\epsilon,\delta}$ follows the general optimization-selection template used in OSSR.Exact. For the optimization step (Line 7), it applies the subroutine $\mathsf{EndoPolicyOptimization}_{t,h}^{\epsilon}$ (Algorithm 5 in Appendix D), which finds a collection of endogenous "one-step" policy covers $(\Gamma^{(t)}[\mathcal{I}])_{\mathcal{I} \in \mathscr{I}_{\le k}\left(\mathcal{I}^{(t+1,h)}\right)}$, which have the property that for all $\mathcal{I} \in \mathscr{I}_{\le k}\left(\mathcal{I}^{(t+1,h)}\right)$ and $s \in \mathcal{S}$, the $t \to h$ policy $\pi_{s[\mathcal{I}]}^{(t)} \circ \psi_{s\left[\mathcal{I}^{(t+1,h)}\right]}^{(t+1,h)}$ (approximately) maximizes the probability that $s_h[\mathcal{I}] = s[\mathcal{I}]$. Then, at selection step (Line 9), $\mathsf{OSSR}_h^{\epsilon,\delta}$ applies the subroutine $\mathsf{EndoFactorSelection}_{t,h}^{\epsilon}$ (Algorithm 6 in Appendix E), which selects a single factor set $\mathcal{I}^{(t,h)} \subseteq \mathcal{I}_\star$ such that—by choosing $\Psi^{(t,h)}$ to be the composition of $\Gamma^{(t)}[\mathcal{I}^{(t,h)}]$ and $\Psi^{(t+1,h)}$—we obtain an (approximate) $t \to h$ policy cover.

Algorithm 8 $\mathsf{OSSR}_h^{\epsilon,\delta}$: Optimization-Selection State Refinement

require:

- Timestep h, precision parameter $\epsilon > 0$, failure probability $\delta \in (0,1)$.
- Policy covers $\{\Psi^{(t)}\}_{t=1}^{h-1}$ for steps $1, \dots, h-1$.
- Upper bound $k \geq 0$ on the cardinality of $\mathcal{I}_\star$.

initialize:

- Let $\mathcal{I}^{(h,h)} \leftarrow \varnothing$ and $\Psi^{(h,h)} \leftarrow \varnothing$.
- Define $N = CAS^{4k}H^2k^3 \log\left(\frac{dSAH}{\delta}\right)\epsilon^{-2}$ for sufficiently large constant $C > 0$, and let $\epsilon_0 := \frac{\epsilon}{2S^kH}$.

for $t = h-1, h-2, .., 1$ **do**

Estimate occupancy measures

Collect dataset $\left\{(s_{t,n}, a_{t,n}, \psi_n^{(t+1,h)}, s_{h,n})\right\}_{n=1}^N$ by drawing N trajectories from the process:

- Execute $\mu^{(t)} := \mathrm{Unf}(\Psi^{(t)})$ up to layer t (resulting in state $s_{t,n}$).
- Sample action $a_{t,n} \sim \mathrm{Unf}(\mathcal{A})$ and play it, transitioning to $s_{t+1,n}$ in the process.
- Sample $\psi_n^{(t+1,h)} \sim \mathrm{Unf}(\Psi^{(t+1,h)})$ and execute it from layers $t+1$ to h (resulting in $s_{h,n}$).

For each $\mathcal{I} \in \mathscr{I}_{\leq k}$, $\pi \in \Pi[\mathscr{I}_{\leq k}]$, and $\psi^{(t+1,h)} \in \Psi^{(t+1,h)}$, define

$$\widehat{d}_h\big(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}\big) = \frac{1}{N}\sum_{n=1}^N \frac{\mathbb{1}\left\{a_{t,n} = \pi(s_{t,n}), \psi_n^{(t+1,h)} = \psi^{(t+1,h)}, s_{h,n}[\mathcal{I}] = s[\mathcal{I}]\right\}}{(1/|\mathcal{A}|)\cdot(1/|\Psi^{(t+1,h)}|)}.$$

Let $\widehat{\mathcal{D}}^{(t,h)} := \left\{\widehat{d}_h(\cdot\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}) \mid \pi \in \Pi(\mathscr{I}_{\leq k}), \psi^{(t+1,h)} \in \Psi^{(t+1,h)})\right\}$.

Phase I: Optimization (Algorithm 5 in Appendix D)

// Beginning from any state at layer t, $\pi_{s[\mathcal{I}]}^{(t)} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}$ maximizes probability that $s_h[\mathcal{I}] = s[\mathcal{I}]$.

For each $\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})$ and $s\,[\mathcal{I}] \in \mathcal{S}\,[\mathcal{I}]$, let

$$\pi_{s[\mathcal{I}]}^{(t)} \leftarrow \mathsf{EndoPolicyOptimization}_{t,h}^{\epsilon_0}\left(\left\{\widehat{d}_h\left(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{t+1,h}]}\right)\right\}_{\pi \in \Pi[\mathscr{I}_{\leq k}]}\right).$$

Let $\Gamma^{(t)}[\mathcal{I}] := \left\{\pi_{s[\mathcal{I}]}^{(t)} \mid s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]\right\}$.

Phase II: Selection (Algorithm 6 in Appendix E)

// Find factor set $\mathcal{I}^{(t,h)} \subseteq \mathcal{I}_\star$ such that $\Gamma^{(t)}\left[\mathcal{I}^{(t,h)}\right]$ has good coverage for all factors in $\mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$.

$(\mathcal{I}^{(t,h)}, \Gamma^{(t)}[\mathcal{I}^{(t,h)}]) \leftarrow \mathsf{EndoFactorSelection}_{t,h}^{\epsilon_0}\left(\{\Gamma^{(t)}\,[\mathcal{I}]\}_{\mathcal{I} \in \mathscr{I}_{\leq k}(\mathcal{I}^{(t+1,h)})}\ ;\ \mathcal{I}^{(t+1,h)}, \Psi^{(t+1,h)}, \widehat{\mathcal{D}}^{(t,h)}\right)$.

Policy composition // Recall that $\pi^{(t)}_{s[\mathcal{I}^{(t,h)}]} \in \Gamma^{(t)}\left[\mathcal{I}^{(t,h)}\right]$ and $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]} \in \Psi^{(t+1,h)}$.

Let $\mathcal{I}^{(t,h)} \leftarrow \widehat{\mathcal{I}}$, then for each $s[\mathcal{I}^{(t,h)}] \in \mathcal{S}[\mathcal{I}^{(t,h)}]$ define $\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} := \pi^{(t)}_{s[\mathcal{I}^{(t,h)}]} \circ_t \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}$.

Let $\Psi^{(t,h)} \leftarrow \left\{\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} : s[\mathcal{I}^{(t,h)}] \in \mathcal{S}[\mathcal{I}^{(t,h)}]\right\}$.

return $\Psi^{(h)} := \Psi^{(1,h)}$. // Policy cover for timestep h.

Full descriptions and proofs of correctness for $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ and $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ are given in Appendix D and Appendix E. Briefly, both subroutines are based on approximate versions of the constraints used in the optimization and selection phase for OSSR.Exact (Line 5 and Line 7 of Algorithm 2), but ensuring endogeneity of the resulting factors is more challenging due to approximation errors, and it no longer suffices to simply search for the factor set with minimum cardinality. Instead, we search for factor sets that satisfy approximate versions of Line 5 and Line 7 with an *additive regularization term* based on cardinality. We show that as long as this penalty is carefully chosen as a function of the statistical error in the occupancy estimates, the resulting factor sets will be endogenous while inducing sufficient amount of exploration (with high probability).

In Appendix C, we provide a general template for designing error-tolerant algorithms that search for endogenous factors using the approach described; both $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ and $\mathsf{EndoFactorSelection}^{\epsilon}_{t,h}$ are special cases of this template.

G.2. Proof of Theorem 3.1

We now restate and prove Theorem 3.1, which shows that $\mathsf{OSSR}^{\epsilon,\delta}_h$ learns an endogenous ϵ-optimal policy cover with sample complexity depending only logarithmically on the number of factors d.

Theorem 3.1 (Sample complexity of OSSR)**.** *Suppose that $\mathsf{OSSR}^{\epsilon,\delta}_h$ is invoked with $\{\Psi^{(t)}\}_{t=1}^{h-1}$, where each $\Psi^{(t)}$ is an endogenous, $\eta/2$-approximate policy cover for layer t. Then with probability at least $1-\delta$, the set $\Psi^{(h)}$ returned by $\mathsf{OSSR}^{\epsilon,\delta}_h$ is an endogenous ϵ-approximate policy cover for layer h, and has $|\Psi^{(h)}| \le S^k$. The algorithm uses at most $O\left(AS^{4k}H^2k^3\log\left(\frac{dSAH}{\delta}\right)\cdot\epsilon^{-2}\right)$ episodes.*

Proof of Theorem 3.1. We begin by defining a success event for ExoRL.

Definition G.1 (Success of OSSR at the layer h)**.** *$\mathcal{G}^{(h)}$ is defined as the event in which the following properties hold:*

1. *$\Psi^{(h)}$ is an endogenous $\eta/2$-approximate policy cover for layer h.*
2. *$\mathcal{I}^{(h)}$ contains only endogenous factors.*

In addition, we define $\mathcal{G}^{(<h)} = \cap_{h'=1}^{h-1}\mathcal{G}^{(h')}$. The following intermediate result—proven in the sequel (Appendix G.3)—serves as our starting point.

Theorem G.1 (Success of State Refinement)**.** *Fix $h\in[H]$ and condition on $\mathcal{G}^{(<h)}$. Then, for any $\epsilon>0$ (recalling that $\epsilon_0 := \frac{\epsilon}{2S^kH}$), by setting*

$$N = \Theta\Big(AS^{2k}k^3\log\left(\frac{dSAH}{\delta}\right)\cdot\epsilon_0^{-2}\Big),$$

$\mathsf{OSSR}^{\epsilon,\delta}_h$ guarantees that with probability at least $1-\delta$, for all $t\le h$,

1. *$\mathcal{I}^{(t,h)}\subseteq\mathcal{I}_\star$, and $\Psi^{(t,h)}$ contains only endogenous policies.*
2. *For all $s\in\mathcal{S}$,*

$$\max_{\pi\in\Pi[\mathcal{I}_\star]} d_h\left(s[\mathcal{I}_\star];\mu^{(t)}\circ_t\pi\circ_{t+1}\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) - d_h\left(s[\mathcal{I}_\star];\mu^{(t)}\circ_t\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]}\right)\le\epsilon_0, \tag{72}$$

where we recall that $\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]}\in\Psi^{(t,h)}$ and $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\in\Psi^{(t+1,h)}$.

We now show that conditioned on the event in Theorem G.1, the set $\Psi^{(h)}$ is an endogenous, ϵ-approximate policy cover (as long as ϵ_0 is chosen to be sufficiently small). In particular, we will show that for all $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$ there exists a policy $\psi \in \Psi^{(h)}$ such that

$$\max_{\pi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]} d_h(s[\mathcal{I}_\star]; \pi) \leq d_h(s[\mathcal{I}_\star]; \psi) + \epsilon. \tag{73}$$

Fix $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$. From first part of Theorem G.1, we have that $\mathcal{I}^{(1,h)} \subseteq \mathcal{I}_\star$, so we can write $s[\mathcal{I}_\star] = (s[\mathcal{I}^{(1,h)}], s[\mathcal{I}_\star \setminus \mathcal{I}^{(1,h)}])$. We will show that the policy $\psi^{(1,h)}_{s[\mathcal{I}^{(1,h)}]} \in \Psi^{(h)} = \Psi^{(1,h)}$ maximizes the probability of reaching $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$ in the sense of Eq. (73).

Define a endogenous "reward function" $R_{s[\mathcal{I}_\star]}$, with

$$R_{s[\mathcal{I}_\star],h}(s_h\left[\mathcal{I}_\star\right]) := \mathbb{1}\left\{s_h\left[\mathcal{I}_\star\right] := s\left[\mathcal{I}_\star\right]\right\}$$

and $R_{s[\mathcal{I}_\star],t}(\cdot) := 0$ for $t \neq h$. Letting $r_{s[\mathcal{I}_\star],t} := R_{s[\mathcal{I}_\star],t}(s_t\left[\mathcal{I}_\star\right])$, we can write

$$d_h(s\left[\mathcal{I}_\star\right]; \pi) := \mathbb{E}_\pi\left[\sum_{t=1}^{h} r_{s[\mathcal{I}_\star],t}\right]. \tag{74}$$

That is, we can view the state occupancy $d_h(s\left[\mathcal{I}_\star\right]; \pi)$ as the state value function for the ExoMDP $\mathcal{M} := \left(\mathcal{S}, \mathcal{A}, T, R_{s[\mathcal{I}_\star]}, H, d_1\right)$. Let $\pi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$ be an endogenous policy. We let $\mathcal{M}_{\mathrm{en}} = \left(\mathcal{S}, \mathcal{A}, T_{\mathrm{en}}, R_{s[\mathcal{I}_\star]}, H, d_{1,\mathrm{en}}\right)$ denote the endogenous component of this MDP, and let $Q^{\pi}_{t,\mathrm{en}}(s[\mathcal{I}_\star], a)$ denote the associated state-action value function for $\mathcal{M}_{\mathrm{en}}$.

To proceed, we use the representation above within the performance difference lemma (Lemma B.6) to bound the suboptimality of $\psi^{(1,h)}_{s[\mathcal{I}^{(1,h)}]}$ by a sum of "per-step" errors for each of the backward steps. In particular for any pair of endogenous policies $\pi, \psi \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]$, Lemma B.6 implies that

$$\begin{aligned}
&d_h(s\left[\mathcal{I}_\star\right]; \pi) - d_h(s\left[\mathcal{I}_\star\right]; \psi) \\
&\overset{\text{(a)}}{=} \sum_{t=1}^{h} \mathbb{E}_\pi\left[Q^{\psi}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \pi_t(s_t\left[\mathcal{I}_\star\right]) - Q^{\psi}_{t,\mathrm{en}}(s_t\left[\mathcal{I}_\star\right], \psi_t(s_t\left[\mathcal{I}_\star\right]))\right] \\
&\leq \sum_{t=1}^{h} \mathbb{E}_{s[\mathcal{I}_\star]\sim d_t(\cdot\,;\,\pi)}\left[\max_a Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \psi_t(s\left[\mathcal{I}_\star\right]))\right] \\
&\leq \sum_{t=1}^{h} \sum_{s[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} \max_{\pi' \in \Pi_{\mathrm{NS}}[\mathcal{I}_\star]} d_t(s\left[\mathcal{I}_\star\right]\;;\;\pi')\left(\max_a Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \psi_t(s\left[\mathcal{I}_\star\right]))\right). \\
&\overset{\text{(b)}}{\leq} 2S^k \sum_{t=1}^{h} \sum_{s[\mathcal{I}_\star]\in\mathcal{S}[\mathcal{I}_\star]} d_t(s\left[\mathcal{I}_\star\right]\;;\;\mu^{(t)})\left(\max_a Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \psi_t(s\left[\mathcal{I}_\star\right]))\right) \\
&= 2S^k \sum_{t=1}^{h} \mathbb{E}_{s[\mathcal{I}_\star]\sim d_t(\cdot\,;\,\mu^{(t)})}\left[\max_a Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \psi_t(s\left[\mathcal{I}_\star\right]))\right] \\
&\overset{\text{(c)}}{=} 2S^k \sum_{t=1}^{h} \left(\max_{\pi' \in \Pi[\mathcal{I}_\star]} d_h(s_h\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \pi' \circ_{t+1} \psi) - d_h(s_h\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \psi)\right).
\end{aligned} \tag{75}$$

We justify the steps above as follows:

- The equality (a) follows from Lemma B.6).
- Relation (b) holds because

$$\frac{\max_{\pi\in\Pi[\mathcal{I}_\star]} d_t\left(\cdot\ ;\ \pi\right)}{d_t\left(\cdot\ ;\ \mu^{(t)}\right)} \leq 2S^k$$

which is a consequence of Lemma A.2. In particular, we use that (i) $\{\Psi^{(t)}\}_{t=1}^{h-1}$ are endogenous $\eta/2$-approximate policy covers, (ii) either $\max_{\pi\in\Pi[\mathcal{I}_\star]} d_t\left(s[\mathcal{I}_\star]\ ;\ \pi\right) \geq \eta$ or $\max_{\pi\in\Pi[\mathcal{I}_\star]} d_t\left(s[\mathcal{I}_\star]\ ;\ \pi\right) = 0$ for all $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$ by the reachability assumption, and (iii) nonnegativity:

$$\max_a Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], a) - Q^{\psi}_{t,\mathrm{en}}(s\left[\mathcal{I}_\star\right], \psi_t(s\left[\mathcal{I}_\star\right])) \geq 0.$$

- Relation (c) holds by the skolemization principle (Lemma A.9) and the tower rule for conditional probabilities.

Recall that the event defined in Theorem G.1 (Eq. (72)) implies that for all $t \leq h$,

$$\max_{\pi'\in\Pi[\mathcal{I}_\star]} d_h(s_h\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \pi' \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]} - d_h(s_h\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \psi^{(t,h)}_{s\left[\mathcal{I}^{(t,h)}\right]}) \leq \epsilon_0.$$

Plugging this bound into Eq. (75) with $\psi \leftarrow \psi^{(1,h)}_{s\left[\mathcal{I}^{(1,h)}\right]}$, we have that for all endogenous policies π,

$$d_h(s\left[\mathcal{I}_\star\right]; \pi) - d_h(s\left[\mathcal{I}_\star\right]; \psi^{(1,h)}_{s\left[\mathcal{I}^{(1,h)}\right]}) \leq 2S^k H\epsilon_0.$$

By using that $\epsilon_0 := \epsilon/2S^k H$ and taking the maximum with respect to $\pi \in \Pi[\mathcal{I}_\star]$, we conclude that for all $s\left[\mathcal{I}_\star\right] = (s\left[\mathcal{I}^{(1,h)}\right], s\left[\mathcal{I}_\star \setminus \mathcal{I}^{(1,h)}\right])$, the policy $\psi^{(1,h)}_{s\left[\mathcal{I}^{(1,h)}\right]}$ satisfies

$$\max_{\pi\in\Pi[\mathcal{I}_\star]} d_h(s\left[\mathcal{I}_\star\right]; \pi) - d_h(s\left[\mathcal{I}_\star\right]; \psi^{(1,h)}_{s\left[\mathcal{I}^{(1,h)}\right]}) \leq \epsilon. \tag{76}$$

This establishes that the set $\Psi^{(h)}$ is an endogenous ϵ-approximate policy cover. With this choice for ϵ_0, the total sample complexity is $O\left(AS^{4k}H^2k^3\log\left(\frac{dSAH}{\delta}\right)\cdot\epsilon^{-2}\right)$. Finally, we note that as a consequence of Theorem G.1, we have $\mathcal{I}^{(1,h)} \subseteq \mathcal{I}_\star$ as desired. We have $|\Psi^{(h)}| \leq S^k$ by construction. □

G.3. Proof of Theorem G.1 (Success of State Refinement Step)

In this section we prove Theorem G.1, a supporting result used in the proof of Theorem 3.1. The result shows for each step t, the optimization and selection phases in $\mathsf{OSSR}^{\epsilon,\delta}_h$ lead to a set of endogenous $t \to h$ policies $\Psi^{(t,h)}$, as long as certain preconditions are satisfied.

Theorem G.1 (Success of State Refinement)**.** *Fix $h \in [H]$ and condition on $\mathcal{G}^{(<h)}$. Then, for any $\epsilon > 0$ (recalling that $\epsilon_0 := \frac{\epsilon}{2S^k H}$), by setting*

$$N = \Theta\Big(AS^{2k}k^3\log\left(\frac{dSAH}{\delta}\right)\cdot\epsilon_0^{-2}\Big),$$

$\mathsf{OSSR}^{\epsilon,\delta}_h$ guarantees that with probability at least $1-\delta$, for all $t \leq h$,

1. *$\mathcal{I}^{(t,h)} \subseteq \mathcal{I}_\star$, and $\Psi^{(t,h)}$ contains only endogenous policies.*
2. *For all $s \in \mathcal{S}$,*

$$\max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s[\mathcal{I}_\star]; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}\right) - d_h\left(s[\mathcal{I}_\star]; \mu^{(t)} \circ_t \psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]}\right) \leq \epsilon_0, \tag{72}$$

where we recall that $\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} \in \Psi^{(t,h)}$ and $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]} \in \Psi^{(t+1,h)}$.

Proof of Theorem G.1. The event $\mathcal{G}^{(<h)}$ (Definition G.1) holds by assumption, which implies that the policy sets $\Psi^{(t)}$ for $t \in [h-1]$ contain only endogenous policies. As a result,

$$\mu^{(t)} := \mathrm{Unf}\left(\Psi^{(t)}\right). \tag{77}$$

is an endogenous mixture policy. To proceed, we define some intermediate success events which will be used throughout the proof. First, for $t \leq h$ define

$$\mathcal{G}_1^{(t,h)} := \left\{\Psi^{(t,h)} \text{ contains only endogenous policies, and } \mathcal{I}^{(t,h)} \subseteq \mathcal{I}_\star\right\}.$$

Observe when $\mathcal{G}_1^{(t,h)}$ holds, we can express all states $s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star]$ as

$$s[\mathcal{I}_\star] = \left(s[\mathcal{I}^{(t,h)}], s[\mathcal{I}_\star \setminus \mathcal{I}^{(t,h)}]\right) = \left(s[\mathcal{I}^{(t+1,h)}], s[\mathcal{I}_\star \setminus \mathcal{I}^{(t+1,h)}]\right),$$

since $\mathcal{I}^{(t+1,h)} \subseteq \mathcal{I}^{(t,h)} \subseteq \mathcal{I}_\star$. Next, we define an event $\mathcal{G}_2^{(t,h)}$ via

$$\mathcal{G}_2^{(t,h)} := \left\{\forall s[\mathcal{I}_\star] \in \mathcal{S}[\mathcal{I}_\star] : \max_{\pi \in \Pi[\mathcal{I}_\star]} d_h(s[\mathcal{I}_\star]; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]}) - d_h(s[\mathcal{I}_\star]; \mu^{(t)} \circ_t \psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]}) \leq \epsilon_0\right\},$$

where we recall that $\psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]} \in \Psi^{(t,h)}$ and $\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} \in \Psi^{(t,h)}$. Finally, let $\mathcal{G}^{(t,h)} := \mathcal{G}_1^{(t,h)} \cap \mathcal{G}_2^{(t,h)}$. We will prove that for all $t \leq h$,

$$\mathbb{P}\left(\mathcal{G}^{(t,h)} \mid \cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}, \mathcal{G}^{(<h)}\right) \geq 1 - \delta/H. \tag{78}$$

Taking a union bound (Lemma A.4), this implies that $\mathbb{P}\left(\cap_{t'=1}^{h} \mathcal{G}^{(t',h)} \mid \mathcal{G}^{(<h)}\right) \geq 1 - \delta$, which establishes Theorem G.1.

Proving Eq. (78). Let $t < h$ be fixed, and condition on $\cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}$ and $\mathcal{G}^{(<h)}$. We will show that whenever these events hold and the estimated occupancy measures have sufficiently high accuracy, $\mathcal{G}^{(t,h)}$ holds. Formally, recalling Definition A.1, define an event

$$\mathcal{G}_{\text{stat}}^{(t,h)} = \left\{\widehat{\mathcal{D}} \text{ is } \frac{\epsilon_0}{12k}\text{-approximate with respect to } \left(\mu^{(t)} \circ_t \Pi[\mathscr{I}_{\leq k}] \circ_{t+1} \Psi^{(t+1,h)}, \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h\right)\right\}. \tag{79}$$

Our goal is to show that conditioned on $\cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}$ and $\mathcal{G}^{(<h)}$, $\mathcal{G}_{\text{stat}}^{(t,h)} \implies \mathcal{G}^{(t,h)}$, so that

$$\mathbb{P}\left(\mathcal{G}^{(t,h)} \mid \cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}, \mathcal{G}^{(<h)}\right) \geq \mathbb{P}\left(\mathcal{G}_{\text{stat}}^{(t,h)} \mid \cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}, \mathcal{G}^{(<h)}\right) \overset{\text{(a)}}{\geq} 1 - \delta.$$

Here (a) is a consequence of Lemma A.5, which asserts that by setting

$$N = \Omega\Big(AS^{2k}k^3 \log\left(\frac{dSA}{\delta}\right) \cdot \epsilon_0^{-2}\Big), \tag{80}$$

the estimated state occupancies $\widehat{\mathcal{D}}$ produced in Line 5 of $\mathsf{OSSR}_h^{\epsilon,\delta}$ are $\epsilon_0/12k$-approximate with respect to $(\mu^{(t)} \circ_t \Pi[\mathscr{I}_{\leq k}] \circ_{t+1} \Psi^{(t+1,h)}, \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h)$, in the sense of Definition A.1. We formally verify that the preconditions required to apply Lemma A.5 are satisfied at the end of the proof for completeness.

We now prove that conditioned on $\cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}$ and $\mathcal{G}^{(<h)}$, $\mathcal{G}_{\mathrm{stat}}^{(t,h)} \implies \mathcal{G}^{(t,h)}$. This relies on two claims: Success of EndoPolicyOptimization and success of EndoFactorSelection.

Success of EndoPolicyOptimization$_{t,h}^{\epsilon_0}$. We appeal to Lemma G.1, verifying that the assumptions it requires, (A1) and (A2), are satisfied (conditioned on $\cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}$ and $\mathcal{G}^{(<h)}$).

(A1) $\mu^{(t)}$ is an endogenous policy when $\mathcal{G}^{(<h)}$ holds (see Eq. (77)) and $\Psi^{(t+1,h)}$ contains only endogenous policies whenever $\mathcal{G}^{(t+1,h)}$ holds.

(A2) $\widehat{\mathcal{D}}$ is $\epsilon_0/12k$-approximate with respect to $(\Pi[\mathscr{I}_{\leq k}], \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h)$ whenever $\mathcal{G}_{\mathrm{stat}}^{(t,h)}$ holds.

Thus, Lemma G.1 implies that for all $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t,h)}\right)$ and $s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]$, the respective invocation of the sub-routine $\mathsf{EndoPolicyOptimization}_{t,h}^{\epsilon_0}$ outputs a policy $\pi_{s[\mathcal{I}]}^{(t)} \in \Gamma^{(t)}[\mathcal{I}]$ that is (i) endogenous, and (ii) near-optimal in the following one-step sense:

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi_{s[\mathcal{I}^{t+1,h}]}^{(t+1,h)}\right) \leq d_h\left(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi_{s[\mathcal{I}]}^{(t)} \circ_{t+1} \psi_{s[\mathcal{I}^{t+1,h}]}^{(t+1,h)}\right) + 4\epsilon_0. \tag{81}$$

Success of EndoFactorSelection$_{t,h}^{\epsilon_0}$. We appeal to Theorem E.1, verifying that the assumptions (A1)-(A3) required by it are satisfied.

(A1) $\mu^{(t)}$ is endogenous whenever $\mathcal{G}^{(<h)}$ holds. Whenever $\mathcal{G}^{(t+1,h)}$ holds, we are guaranteed that $\Psi^{(t+1,h)}$ contains only endogenous policies, so that $\psi_{s[\mathcal{I}^{t+1,h}]}^{(t+1,h)} \in \Psi^{(t+1,h)}$ is endogenous in particular.

(A2) $\widehat{\mathcal{D}}$ is $\epsilon_0/12k$-approximate with respect to $(\Pi[\mathscr{I}_{\leq k}], \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h)$ by $\mathcal{G}_{\mathrm{stat}}^{(t,h)}$.

(A3) Due to the success of $\mathsf{EndoPolicyOptimization}_{t,h}^{\epsilon_0}$ (verified above), the condition in Eq. (81) is satisfied.

Hence, by Theorem E.1, $\mathsf{EndoFactorSelection}_{t,h}^{\epsilon_0}$ returns a tuple $(\mathcal{I}^{(t,h)}, \Psi^{(t,h)}[\mathcal{I}^{(t,h)}])$ such that

1. $\mathcal{I}^{(t,h)} \subseteq \mathcal{I}_\star$.
2. For all $s \in \mathcal{S}$,

$$\max_{\pi \in \Pi[\mathcal{I}_\star]} d_h\left(s\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \pi \circ_{t+1} \psi_{s\left[\mathcal{I}^{(t+1,h)}\right]}^{(t+1,h)}\right) - d_h\left(s\left[\mathcal{I}_\star\right]; \mu^{(t)} \circ_t \pi_{s[\mathcal{I}^{(t,h)}]}^{(t)} \circ_{t+1} \psi_{s\left[\mathcal{I}^{(t+1,h)}\right]}^{(t+1,h)}\right) \leq 16\epsilon_0,$$

where we recall that $\psi_{s\left[\mathcal{I}^{(t+1,h)}\right]}^{(t+1,h)} \in \Psi^{(t+1,h)}$ and $\pi_{s[\mathcal{I}^{(t,h)}]}^{(t)} \in \Gamma^{(t)}[\mathcal{I}^{(t,h)}]$.

Wrapping up. Scaling $\epsilon_0 \leftarrow \epsilon_0/16$ and $\delta \leftarrow \delta/H$, and recalling that $\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} \in \Psi^{(t,h)}$ is given by

$$\psi^{(t,h)}_{s[\mathcal{I}^{(t,h)}]} := \pi^{(t)}_{s[\mathcal{I}^{(t,h)}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{(t+1,h)}]},$$

we have that for all $t < h$,

$$\mathbb{P}\left(\mathcal{G}^{(t,h)} \mid \cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}, \cap_{h'=1}^{h-1} \mathcal{G}^{(h')}\right) \geq 1 - \delta/H,$$

proving the result.

Verifying conditions of Lemma A.5. We conclude by verifying that the four conditions required by Lemma A.5 hold, conditioned on $\cap_{t'=t+1}^{h} \mathcal{G}^{(t',h)}$ and $\mathcal{G}^{(<h)}$; this justifies the application in the prequel.

1. By construction, $\Psi^{(t+1,h)} = \left\{\psi^{(t+1,h)}_{s[\mathcal{I}^{(t,h)}]} \mid s\left[\mathcal{I}^{(t+1,h)}\right] \in \mathcal{S}\left[\mathcal{I}^{(t+1,h)}\right]\right\}$. Thus, $|\Psi^{(t+1,h)}| = |\mathcal{S}\left[\mathcal{I}^{(t,h)}\right]| \leq S^k$, since $|\mathcal{I}^{(t+1,h)}| \leq k$.
2. We have $|\Pi[\mathscr{I}_{\leq k}]| \leq O\left(d^k A^{S^k}\right)$, since the number of factor sets of size at most k is

$$\sum_{k'=0}^{k} \binom{d}{k'} \leq \left(\frac{ed}{k}\right)^k \leq O\left(d^k\right), \tag{82}$$

   and for any factor set $\mathcal{I}$ with $|\mathcal{I}| \leq k$ we have $|\Pi[\mathcal{I}]| \leq A^{S^k}$.
3. $|\mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)| \leq |\mathscr{I}_{\leq k}| \leq O\left(d^k\right)$ by Eq. (82),
4. For any fixed set $\mathcal{I}$ with $|\mathcal{I}| \leq k$, we have $|\mathcal{S}\left[\mathcal{I}\right]| \leq S^k$.

□

G.4. Application of EndoPolicyOptimization in OSSR

The main guarantee for the $\mathsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ subroutine (Theorem D.1) implies that the policy $\pi^{(t)}_{s[\mathcal{I}]}$ returned in Line 7 of OSSR is endogenous, as well as near-optimal in the following this sense:

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{t+1,h}]}\right) \leq d_h\left(s[\mathcal{I}]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s[\mathcal{I}^{t+1,h}]}\right) + O(\epsilon).$$

In this subsection we state and prove Lemma G.1, which shows that the preconditions (A1) and (A2) required to apply Theorem D.1 are satisfied, so that the claim above indeed holds.

Lemma G.1. *Fix $h \in [H]$ and $t \leq h$. Suppose that the following conditions hold:*

(C1) $\mu^{(t)} \in \Pi_{\text{mix}}[\mathcal{I}_\star]$ *is endogenous and* $\Psi^{(t+1,h)}$ *contains only endogenous policies.*

(C2) *The collection $\widehat{\mathcal{D}}$ of occupancy measures is $\epsilon/12k$-approximate with respect to* $\left(\mu^{(t)} \circ \Pi[\mathscr{I}_{\leq k}] \circ \Psi^{(t+1,h)}, \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h\right)$.

Then assumptions $(\mathrm{A1})$ *and* $(\mathrm{A2})$ *of Theorem D.1 are satisfied when* $\textsf{EndoPolicyOptimization}^{\epsilon}_{t,h}$ *is invoked within* OSSR, *and for all* $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$*:*

1. *The set* $\Gamma^{(t)}\left[\mathcal{I}\right] = \left\{\pi^{(t)}_{s[\mathcal{I}]} \mid s[\mathcal{I}] \in \mathcal{S}[\mathcal{I}]\right\}$ *contains only endogenous policies.*
2. *For all* $s\left[\mathcal{I}\right] \in \mathcal{S}[\mathcal{I}]$*, the policy* $\pi^{(t)}_{s[\mathcal{I}]} \in \Gamma\left[\mathcal{I}\right]$ *satisfies*

$$\max_{\pi \in \Pi[\mathscr{I}_{\leq k}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right)$$
$$\leq d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi^{(t)}_{s[\mathcal{I}]} \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) + 4\epsilon.$$

Proof of Lemma G.1. Toward proving the result, we begin with a basic observation. Fix $\mathcal{I} \in \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right)$ and $s\left[\mathcal{I}\right] \in \mathcal{S}\left[\mathcal{I}\right]$. Define an MDP $\left(\mathcal{S}, \mathcal{A}, T, R_{s[\mathcal{I}]}, h\right)$ where $R_{s[\mathcal{I}],h} = \mathbb{1}\left\{s_h[\mathcal{I}] = s[\mathcal{I}]\right\}$ and $R_{s[\mathcal{I}],h'} = 0$ for all $h' \neq h$. Observe that the occupancy measure for $s[\mathcal{I}]$ at layer h is equivalent to the (t, h) value function in this MDP:

$$V_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) = d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right). \tag{83}$$

We now show that assumptions $(\mathrm{A1})$ and $(\mathrm{A2})$ of Theorem D.1 hold when the theorem is invoked with this value function, from which the result will follow.

Verifying assumption $(\mathrm{A1})$ *of Theorem D.1.* The policies $\mu^{(t)}$ and $\psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]} \in \Psi^{(t+1,h)}$ are endogenous by condition $(\mathrm{C1})$. Hence, the assumptions of the restriction lemma (Lemma B.2) are satisfied, which gives

$$\max_{\pi \in \Pi[\mathcal{I}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) = \max_{\pi \in \Pi[\mathcal{I}_{\mathrm{en}}]} d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right)$$
$$\iff \max_{\pi \in \Pi[\mathcal{I}]} V_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) = \max_{\pi \in \Pi[\mathcal{I}_{\mathrm{en}}]} V_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right).$$

Verifying assumption $(\mathrm{A2})$ *of Theorem D.1.* By condition $(\mathrm{C2})$, we have that $\widehat{\mathcal{D}}$ is $\epsilon/12k$-approximate with respect to $(\mu^{(t)} \circ \Pi[\mathscr{I}_{\leq k}] \circ \Psi^{(t+1,h)}, \mathscr{I}_{\leq k}\left(\mathcal{I}^{(t+1,h)}\right), h)$, and hence

$$\left|\widehat{d}_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) - d_h\left(s\left[\mathcal{I}\right]\ ;\ \mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right)\right| \leq \epsilon/12k$$
$$\iff \left|\widehat{V}_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right) - V_{t,h}\left(\mu^{(t)} \circ_t \pi \circ_{t+1} \psi^{(t+1,h)}_{s\left[\mathcal{I}^{(t+1,h)}\right]}\right)\right| \leq \epsilon/12k.$$

□

Appendix H. Proof of Theorem 4.1 (Correctness of ExoRL)

In this section we formally prove Theorem 4.1, which shows that ExoRL (Algorithm 3) learns an ϵ-optimal policy for a general ExoMDP. The correctness of ExoRL is essentially a direct corollary of the results derived for OSSR and PSDP in Appendix G and Appendix F. The high probability guarantee for OSSR (Theorem 3.1) implies that iteratively applying $\mathsf{OSSR}_h^{\eta/2,\delta}$ results in an endogenous $\eta/2$-approximate policy covers for every layer $h \in [H]$. Conditioning on this event, ExoPSDP is guaranteed to find an ϵ-optimal policy with high probability (Theorem F.1).

Theorem 4.1 (Sample complexity of ExoRL)**.** *ExoRL, when invoked with parameter, $\epsilon \in (0,1)$ and $\delta \in (0,1)$, returns an ϵ-optimal policy with probability at least $1-\delta$, and does so using at most $O\left(AS^{3k}H^2(S^k+H^2)k^3\log\left(\frac{dSAH}{\delta}\right)\cdot\left(\epsilon^{-2}+\eta^{-2}\right)\right)$ episodes.*

Proof of Theorem 4.1. We first show that OSSR results in a near-optimal (endogenous) policy cover, then show that the application of ExoPSDP is successful.

Application of OSSR. Let $\mathcal{G}^{(h)}$ denote the event in which $\mathsf{OSSR}_h^{\eta/2,\delta}(\{\Psi^{(t)}\}_{t=1}^{h-1})$ returns an endogenous $\eta/2$-approximate policy cover $\Psi^{(h)}$ with $|\Psi^{(h)}| \leq S^k$, and let $\mathcal{G}^{(<h)} := \cap_{h'=1}^{h-1}\mathcal{G}^{(h)}$. Theorem 3.1 states that for all $h \geq 2$, if we condition on $\mathcal{G}^{(<h)}$, then given $N = O\left(\frac{AS^{4k}H^2k^3\log\left(\frac{dSAH}{\delta}\right)}{\eta^2}\right)$ samples, $\mathsf{OSSR}_h^{\eta/2,\delta}$ ensures that $\mathcal{G}^{(h)}$ holds probability at least $1-\delta$. Furthermore, $\mathcal{G}^{(1)}$ holds trivially for $h=1$. By Lemma A.4, this implies that $\mathbb{P}\left(\cap_{h=1}^{H}\mathcal{G}^{(h)}\right) \geq 1-H\delta$. Scaling $\delta \leftarrow \delta/2H$, we conclude that given

$$N_{\mathsf{OSSR}} = O\left(\frac{AS^{4k}H^2k^3\log\left(\frac{dSAH}{\delta}\right)}{\eta^2}\right)$$

samples across all applications of $\mathsf{OSSR}_h^{\eta/2,\delta}$, the collection $\{\Psi^{(h)}\}_{h=1}^{H}$ is a set of endogenous $\eta/2$-approximate policy covers with probability at least $1-\delta/2$. We denote this event by $\mathcal{G}_{\mathsf{OSSR}}$, so that $\mathbb{P}\left(\mathcal{G}_{\mathsf{OSSR}}\right) \geq 1-\delta/2$.

Application of PSDP. Conditioned on the event $\mathcal{G}_{\mathsf{OSSR}}$, the conditions of Theorem F.1 hold, so that the application of ExoPSDP is admissible. As a result, given

$$N_{\mathsf{ExoPSDP}} = O\left(\frac{AS^{3k}H^4k^3\log\left(\frac{dSAH}{\delta}\right)}{\epsilon^2}\right)$$

samples, ExoPSDP finds an endogenous ϵ-optimal policy. We denote this event by $\mathcal{G}_{\mathsf{ExoPSDP}}$, so that $\mathbb{P}\left(\mathcal{G}_{\mathsf{ExoPSDP}} \mid \mathcal{G}_{\mathsf{OSSR}}\right) \geq 1-\delta/2$.

Concluding the proof. ExoRL returns an endogenous ϵ-optimal policy when $\mathcal{G}_{\mathsf{OSSR}}$ and $\mathcal{G}_{\mathsf{ExoPSDP}}$ hold, and by the union bound $\mathbb{P}\left(\mathcal{G}_{\mathsf{OSSR}} \cap \mathcal{G}_{\mathsf{ExoPSDP}}\right) \geq 1-\delta$. The total number of samples is

$$N = N_{\mathsf{OSSR}} + N_{\mathsf{ExoPSDP}} \leq O\left(\frac{AS^{4k}H^2k^3\log\left(\frac{dSAH}{\delta}\right)}{\eta^2} + \frac{AS^{3k}H^4k^3\log\left(\frac{dSAH}{\delta}\right)}{\epsilon^2}\right).$$

□

H.1. Computational Complexity of ExoRL

The ExoRL procedure can be implemented with $O(d^k N S^k A H)$ runtime. In Appendix F.3, we show that ExoPSDP can be implemented in runtime $O(d^k N S^k A H)$. Similarly, $\mathsf{OSSR}_h^{\epsilon,\delta}$ can be implemented with runtime $O(d^k N S^k A)$. The most computationally demanding aspect of OSSR is optimizing the function $\widehat{V}_{t,H}\,(\mu^{(t)} \circ_t \pi \circ_{t+1} \widehat{\pi}_{t+1:H})$ over the policy class $\Pi[\mathscr{I}_{\leq k}]$. As shown in Appendix F.3, this procedure can be implemented with runtime $O(d^k N S^k A)$, which is repeated for H times in ExoRL.