Provably Efficient Causal Model-Based Reinforcement Learning for Environment-Agnostic Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the sequential decision making setting, an agent aims to achieve *systematic generalization* over a large, possibly infinite, set of environments. Such environments are modeled as discrete Markov decision processes with both states and actions represented through a feature vector. The underlying structure of the environments allows the transition dynamics to be factored into two components: one that is environment-specific and another one that is shared. Consider a set of environments that share the laws of motion as an illustrative example. In this setting, the agent can take a finite amount of *reward-free* interactions from a subset of these environments. The agent then must be able to *approximately* solve any planning task defined over any environment in the original set, relying on the above interactions only. Can we design a provably efficient algorithm that achieves this ambitious goal of systematic generalization? In this paper, we give a partially positive answer to this question. First, we provide the first tractable formulation of systematic generalization by employing a *causal* viewpoint. Then, under specific structural assumptions, we provide a simple learning algorithm that allows us to guarantee any desired planning error up to an unavoidable sub-optimality term, while showcasing a polynomial sample complexity.

1 Introduction

Whereas recent breakthroughs have established Reinforcement Learning (RL) [37] as a powerful tool to address a wide range of sequential decision making problems, the curse of generalization [24] is still a main limitation of commonly used techniques. RL algorithms deployed on a given task are usually effective in discovering the correlation between an agent's behavior and the resulting performance from large amounts of labeled samples [19, 26]. However, those algorithms are usually unable to discover basic cause-effect relations between the agent's behavior and the environment dynamics. Crucially, the aforementioned correlations are oftentimes specific to the task at hand, and they are unlikely to be of any use for addressing different tasks or environments. Instead, some universal causal relations generalize over the environments, and once learned they can be exploited for solving any task. Let us consider as an illustrative example an agent interacting with a large set of physical environments. While each of these environments can have its specific dynamics, we expect the basic laws of motion to hold across the environments, as they encode general causal relations. Once they are learned, there is no need to discover them again from scratch when facing a new task, or an unseen environment. Even if the dynamics over these relations can change, such as moving underwater is different than moving in the air, or the gravity can change from planet to planet, the underlying causal structure still holds. This knowledge alone often allows the agent to solve new tasks in unseen environments by taking a few, or even zero, interactions.

Submitted to *A causal view on dynamical systems* workshop at NeurIPS 2022. Do not distribute.

Table 1: Sample complexity overview.

	Target	K (discrete MDP)	K (tabular MDP)
Causal Structure Estimation	$Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon) \leq \delta$	$O\big(n \log(d_S^2 d_A/\delta)/\epsilon^2\big)$	$O\big(\log(S^2 A/\delta)/\epsilon^2\big)$
Bayesian Network Estimation	$Pr(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$	$\widetilde{O}\big(d_S^3 n^{3Z+1}/\epsilon^2\big)$	$\widetilde{O}\big(S^2 2^{2Z}/\epsilon^2\big)$
Systematic Generalization	$Pr(V_1^* - V_1^\pi \geq \epsilon_\lambda + \epsilon) \leq \delta$	$\widetilde{O}(MH^6 d_S^3 Z^2 n^{5Z+3}/\epsilon^2)$	$\widetilde{O}(MH^6 S^4 A^2 Z^2/\epsilon^2)$

We argue that we should pursue this kind of generalization in RL, which we call *systematic generalization*, where learning universal causal relations from interactions with a few environments allows us to approximately solve any task in any other environment without further interactions. Although this problem setting might seem overly ambitious or even far-fetched, in this paper we provide the first tractable formulation of systematic generalization, thanks to a set of structural assumptions that are motivated by a causal viewpoint. The problem formulation is partially inspired by reward-free RL [20], in which the agent can take unlabelled interactions with an environment to learn a model that allows approximate planning for any reward function. Here, we extend this formulation to a large, potentially infinite, set of reward-free environments, or a *universe*, the agent can freely interact with. We consider discrete environments, such that both their states and actions can be described through vectors of discrete features. Crucially, these environments share a common causal structure that explains a significant portion, but not all, of their transition dynamics. Can we design a provably efficient algorithm that guarantees an arbitrarily small planning error for any possible task that can be defined over the set of environments, by taking reward-free interactions with a generative model?

In this paper, we provide a partially positive answer to this question by presenting a simple but principled causal model-based approach (see Figure 1). This algorithm interacts with a finite subset of the universe to learn the causal structure underlying the set of environments in the form of a causal dependency graph $\mathcal{G}$ [43]. Then, the causal transition model, which encodes the dynamics that is common across the environment, is obtained by estimating the Bayesian network $P_{\mathcal{G}}$ over $\mathcal{G}$ [9] from a mixture of the environments. Finally, the causal transition model is employed by a planning oracle to provide an approximately optimal policy for a latent environment and a given reward function. We can show that this simple recipe, with a sample complexity that is polynomial in all the relevant quantities of the problem, allows achieving any desired planning error up to an unavoidable error term. The latter is inherent to the setting, which demands generalization over an infinite set of environments, and cannot be overcome without additional samples from the test environment.

The contributions of this paper include (see Table 1 for a summary of the sample complexity rates):

- The first tractable formulation of the systematic generalization problem in RL, thanks to structural assumptions motivated by causal considerations (§ 3);
- A simple provably efficient algorithm to learn systematic generalization over an infinite set of environments with a polynomial sample complexity (§ 4.1);
- The sample complexity of estimating the causal structure underlying a discrete MDP (§ 4.2);
- The sample complexity of estimating the Bayesian network underlying a discrete MDP (§ 4.3).

With this work we aim to connect several active research areas on reward-free RL [20], model-based RL [37], factored MDPs [35], causal RL [49], experimental design [17], independence testing [5], in a general framework where individual progresses can be enhanced beyond the sum of their parts.

2 Preliminaries

We start with some basic notions about graphs, causality, and Markov decision processes that will be useful thereafter. We will denote a set of integers $\{1, \dots, a\}$ as $[a]$, and the probability simplex over the space $\mathcal{A}$ as $\Delta_{\mathcal{A}}$. For a factored space $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_a$ and a set of indices $Z \subseteq [a]$, which we call a *scope*, we denote the scope operator as $\mathcal{A}[Z] := \bigotimes_{i \in Z} \mathcal{A}_i$, in which $\bigotimes$ denotes a cardinal product. For any $A \in \mathcal{A}$, we denote with $A[Z]$ the vector $(A_i)_{i \in Z}$. For singletons we write $A[i]$ as a shorthand for $A[\{i\}]$. Given two probability measures P and Q over a discrete space $\mathcal{A}$, their L_1-distance is $\|P - Q\|_1 = \sum_{A \in \mathcal{A}} |P(A) - Q(A)|$, and their Kullback-Leibler (KL) divergence is $d_{KL}(P||Q) = \sum_{A \in \mathcal{A}} P(A) \log(P(A)/Q(A))$. We denote by $\mathcal{U}_{\mathcal{A}}$ the uniform distribution over $\mathcal{A}$.

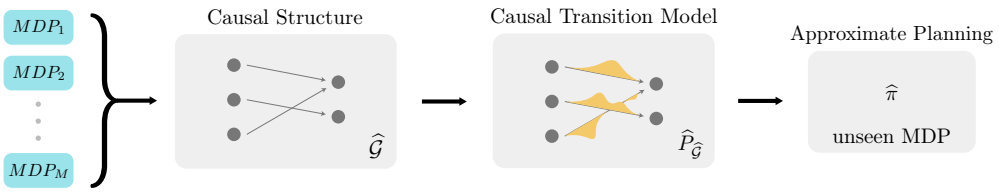


Figure 1: High-level illustration of the causal model-based approach to systematic generalization.

Graphs We define a graph $\mathcal{G}$ as a pair $\mathcal{G} := (\mathcal{V}, E)$, where $\mathcal{V}$ is a set of nodes and $E \subseteq N \times N$ is a set of edges between them. We call $\mathcal{G}$ a *directed graph* if all of its edges E are directed (i.e., ordered pairs of nodes). We also define the in-degree of a node to be its number of incoming edges: $\text{degree}_{\text{in}}(A) = |\{(B, A) : (B, A) \in E, \forall B\}|$. $\mathcal{G}$ is said to be a *Directed Acyclic Graph* (DAG) if it is a directed graph without cycles. We call $\mathcal{G}$ a *bipartite graph* if there exists a partition $X \cup Y = \mathcal{V}$ such that none of the nodes in X and Y are connected by an edge, i.e., $E \cap (X \times X) = E \cap (Y \times Y) = \emptyset$. For any subset of nodes $S \subset \mathcal{V}$, we define the *subgraph* induced by S as $\mathcal{G}[S] := (S, E[S])$, in which $E[S] = E \cap (S \times S)$. The *skeleton* of a graph $\mathcal{G}$ is the undirected graph that is obtained from $\mathcal{G}$ by replacing all the directed edges in E with undirected ones. Finally, the *graph edit distance* between two graphs is the minimum number of graph edits (addition or deletion of either a node or an edge) necessary to transform one graph into the other.

Causal Graphs and Bayesian Networks For a set $\mathcal{X}$ of random variables, we represent the causal structure over $\mathcal{X}$ with a DAG $\mathcal{G}_{\mathcal{X}} = (\mathcal{X}, E)$,[1] which we call the *causal graph* of $\mathcal{X}$. For each pair of variables $A, B \in \mathcal{X}$, a directed edge $(A, B) \in \mathcal{G}_{\mathcal{X}}$ denotes that B is conditionally dependent on A. For every variable $A \in \mathcal{X}$, we denote as $\text{Pa}(A)$ the *causal parents* of A, i.e., the set of all the variables $B \in \mathcal{X}$ on which A is conditionally dependent, $(B, A) \in \mathcal{G}_{\mathcal{X}}$. A Bayesian network [10] over the set $\mathcal{X}$ is defined as $\mathcal{N} := (\mathcal{G}_{\mathcal{X}}, P)$, where $\mathcal{G}_{\mathcal{X}}$ specifies the *structure* of the network, i.e., the dependencies between the variables in $\mathcal{X}$, and the distribution $P : \mathcal{X} \to \Delta_{\mathcal{X}}$ specifies the conditional probabilities of the variables in $\mathcal{X}$, such that $P(\mathcal{X}) = \prod_{X_i \in \mathcal{X}} P_i(X_i | \text{Pa}(X_i))$.

Markov Decision Processes A *tabular* episodic Markov Decision Process (MDP) [33] is defined as $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, H, r)$, where $\mathcal{S}$ is a set of $|\mathcal{S}| = S$ states, $\mathcal{A}$ is a set of $|\mathcal{A}| = A$ actions, P is a transition model such that $P(s'|s, a)$ gives the conditional probability of the next state s' having taken action a in state s, H is the episode horizon, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic reward function.

The strategy of an agent interacting with $\mathcal{M}$ is represented by a non-stationary, stochastic *policy*, a collection of functions $(\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}})_{h \in [H]}$ where $\pi_h(a|s)$ denotes the conditional probability of taking action a in state s at step h. The *value function* $V_h^{\pi} : \mathcal{S} \to \mathbb{R}$ associated to π is defined as the expected sum of the rewards that will be collected, under the policy π, starting from s at step h, i.e.,

$$V_h^{\pi}(s) := \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right].$$

For later convenience, we further define $PV_{h+1}^{\pi}(s, a) := \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_{h+1}^{\pi}(s')]$ and $V_1^{\pi} := \mathbb{E}_{s \sim P}[V_1^{\pi}(s)]$. We will write $V_{\mathcal{M},r}^{\pi}$ to denote V_1^{π} in the MDP $\mathcal{M}$ with reward function r (if not obvious from the context). For an MDP $\mathcal{M}$ with finite states, actions, and horizon, there always exists an *optimal policy* π^* that gives the value $V_h^*(s) = \sup_{\pi} V_h^{\pi}(s)$ for every s, a, h. The goal of the agent is to find a policy π that is ϵ-close to the optimal one, i.e., $V_1^* - V_1^{\pi} \leq \epsilon$.

Finally, we define a *discrete* Markov decision process as $\mathcal{M} := ((\mathcal{S}, d_S, n), (\mathcal{A}, d_A, n), P, H, r)$, where $\mathcal{S}, \mathcal{A}, P, H, r$ are specified as before, and where the states and actions spaces admit additional structure, such that every $s \in \mathcal{S}$ can be represented through a d_S-dimensional vector of discrete features taking value in $[n]$, and every $a \in \mathcal{A}$ can be represented through a d_A-dimensional vector of discrete features taking value in $[n]$. Note that any tabular MDP can be formulated under this alternative formalism through one-hot encoding by taking $n = 2$, $d_S = S$, and $d_A = A$.

[1] We will omit the subscript $\mathcal{X}$ whenever clear from the context.

3 Problem Formulation

In our setting, a learning agent aims to master a large, potentially infinite, set $\mathbb{U}$ of environments modeled as discrete MDPs without rewards,

$$\mathbb{U} := \left\{\mathcal{M}_i = ((\mathcal{S}, d_S, n), (\mathcal{A}, d_A, n), P_i, \mu)\right\}_{i=1}^{\infty},$$

which we call a *universe*. The agent can draw a finite amount of experience by interacting with the MDPs in $\mathbb{U}$. From these interactions alone, the agent aims to acquire sufficient knowledge to approximately solve any task that can be specified over the universe $\mathbb{U}$. Specifically, a *task* is defined as any pairing of an MDP $\mathcal{M} \in \mathbb{U}$ and a reward function r, whereas *solving it* refers to providing a slightly sub-optimal policy via planning, i.e., without taking additional interactions. We call this problem *systematic generalization*, which we can formalize as follows.

Definition 1 (Systematic Generalization)**.** *For any latent MDP $\mathcal{M} \in \mathbb{U}$ and any given reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, the systematic generalization problem requires the agent to provide a policy π, such that $V^*_{\mathcal{M},r} - V^{\pi}_{\mathcal{M},r} \leq \epsilon$ up to any desired sub-optimality $\epsilon > 0$.*

Since the set $\mathbb{U}$ is infinite, we clearly require additional structure to make the problem feasible. On the one hand, the state space $(\mathcal{S}, d_S, n)$, action space $(\mathcal{A}, d_A, n)$, and initial state distribution μ are shared across $\mathcal{M} \in \mathbb{U}$. The transition dynamics P_i is instead specific to each MDP $\mathcal{M}_i \in \mathbb{U}$. However, we assume the presence of a *common causal structure* that underlies the transition dynamics of the universe, and relates the single transition models P_i.

3.1 The Causal Structure of the Transition Dynamics

The transition dynamics of a discrete MDP gives the conditional probability of next state features s' given the current state-action features (s, a). To ease the notation, from now on we will denote the state-action features with a random vector $X = (X_i)_{i \in [d_S + d_A]}$, in which each X_i is supported in $[n]$, and the next state features with a random vector $Y = (Y_i)_{i \in [d_S]}$, in which each Y_i is supported in $[n]$.

For each environment $\mathcal{M}_i \in \mathbb{U}$, the conditional dependencies between the next state features Y and the current state-action features X are represented through a bipartite dependency graph $\mathcal{G}_i$, such that $(X[z], Y[j]) \in \mathcal{G}_i$ if and only if $Y[j]$ is conditionally dependent on $X[z]$. Clearly, each environment can display its own dependencies, but we assume there is a set of dependencies that represent general causal relationships between the features, and that appear in any $\mathcal{M}_i \in \mathbb{U}$. In particular, we call the intersection $\mathcal{G} := \cap_{i=0}^{\infty} \mathcal{G}_i$ the *causal structure* of $\mathbb{U}$, which is the set of conditional dependencies that are common across the universe. In Figure 2, we show an illustration of such a causal structure. Since it represents universal causal relationships, the causal structure $\mathcal{G}$ is time-consistent, i.e., $\mathcal{G}^{(h)} = \mathcal{G}^{(1)}$ for any step $h \in [H]$, and we further assume that $\mathcal{G}$ is sparse, which means that the number of features $X[z]$ on which a feature $Y[j]$ is dependent on is bounded from above.

Assumption 1 (Z-sparseness)**.** *The causal structure $\mathcal{G}$ is Z-*sparse *if* $\max_{j \in [d_S]} \text{degree}_{\text{in}}(Y[j]) \leq Z$.

Given a causal structure $\mathcal{G}$, and without losing generality,[2] we can express each transition model P_i as $P_i(Y|X) = P_{\mathcal{G}}(Y|X) F_i(Y|X)$, in which $P_{\mathcal{G}}$ is the Bayesian network over the causal structure $\mathcal{G}$, whereas F_i includes environment-specific factors affecting the conditional probabilities.[3] Since it represents the conditional probabilities due to universal causal relations in $\mathbb{U}$, we call $P_{\mathcal{G}}$ the *causal transition model* of $\mathbb{U}$. Thanks to the structure $\mathcal{G}$, this model can be further factored as

$$P_{\mathcal{G}}(Y|X) = \prod_{j=1}^{d_S} P_j(Y[j]|X[Z_j]), \tag{1}$$

where the scopes Z_j are the set of indices z such that $(X[z], Y[j]) \in \mathcal{G}$, i.e. the causal parents of $Y[j]$. In Figure 3, we show an illustration of the causal transition model and its factorization. Similarly to the underlying structure $\mathcal{G}$, the causal transition model $P_{\mathcal{G}}$ is also time-consistent, i.e., $P^{(h)}_{\mathcal{G}} = P^{(1)}_{\mathcal{G}}$ for any step $h \in [H]$. For the purpose of this work, we assume that the causal transition model is non-vacuous and that it explains a significant part of the transition dynamics of $\mathcal{M}_i \in \mathbb{U}$.

[2] Note that one can always take $P_{\mathcal{G}}(Y|Z) = 1, \forall (X, Y)$ to avoid shared structure on the transition dynamics.

[3] The parameters in F_i are numerical values such that P_i remains a well-defined probability measure.

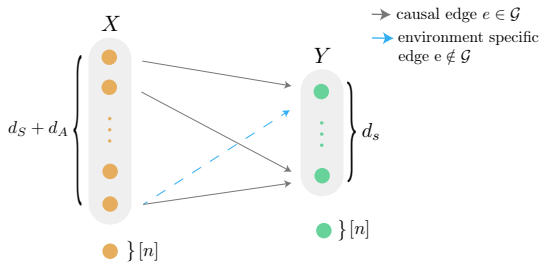


Figure 2: Causal structure $\mathcal{G}$ of $\mathbb{U}$.

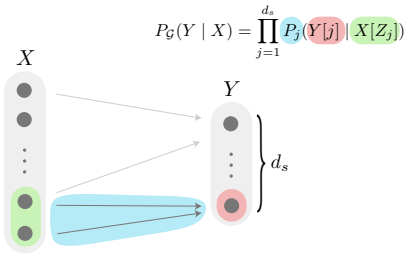


Figure 3: Causal transition model $P_{\mathcal{G}}$ of $\mathbb{U}$.

Assumption 2 (λ-sufficiency)**.** *Let $\lambda \in [0, 1]$ be a constant. The causal transition model $P_{\mathcal{G}}$ is* causally λ-sufficient *if* $\sup_X \|P_{\mathcal{G}}(\cdot|X) - P_i(\cdot|X)\|_1 \leq \lambda$, $\forall P_i \in \mathcal{M}_i \in \mathbb{U}$.

Notably, the parameter λ controls the amount of the transition dynamics that is due to the universal causal relations $\mathcal{G}$ ($\lambda = 0$ means that $P_{\mathcal{G}}$ is sufficient to explain the transition dynamics of any $\mathcal{M}_i \in \mathbb{U}$, whereas $\lambda = 1$ implies no shared structure between the transition dynamics of the $\mathcal{M}_i \in \mathbb{U}$). In this paper, we argue that learning the causal transition model $P_{\mathcal{G}}$ is a good target for systematic generalization and we provide theoretical support for this claim in § 4.

3.2 A Class of Training Environments

Even if the universe $\mathbb{U}$ admits the structure that we presented in the last section, it is still an infinite set. Instead, the agent can only interact with a finite subset of discrete MDPs

$$\mathbb{M} := \{\mathcal{M}_i = ((\mathcal{S}, d_S, n), (\mathcal{A}, d_A, n), P_i, \mu)\}_{i=1}^{M} \subset \mathbb{U},$$

which we call a *class* of size M. Crucially, the causal structure $\mathcal{G}$ is a property of the full set $\mathbb{U}$, and if we aim to infer it from interactions with a finite class $\mathbb{M}$, we have to assume that $\mathbb{M}$ is informative enough on the universal causal relations of $\mathbb{U}$.

Assumption 3 (Diversity)**.** *Let $\mathbb{M} \subset \mathbb{U}$ be class of size M. We say that $\mathbb{M}$ is* causally diverse *if* $\mathcal{G} = \cap_{i=1}^{M} \mathcal{G}_i = \cap_{i=1}^{\infty} \mathcal{G}_i$.[4]

Analogously, if we aim to infer the causal transition model $P_{\mathcal{G}}$ from interactions with the transition models P_i of the single MDPs $\mathcal{M}_i \in \mathbb{M}$, we have to assume that $\mathbb{M}$ is balanced in terms of the conditional probabilities displayed by its components, so that the factors that do not represent universal causal relations even out while learning.

Assumption 4 (Evenness)**.** *Let $\mathbb{M} \subset \mathbb{U}$ be class of size M. We say that $\mathbb{M}$ is* causally even *if* $\forall j \in [d_S]$, $\mathbb{E}_{i \sim \mathcal{U}_{[M]}} \left[F_i(Y[j]|X)\right] = 1$.

Whereas in this paper we assume that $\mathbb{M}$ is *diverse* and *even* by design, we leave as future work the interesting problem of selecting such a class from active interactions with $\mathbb{U}$, which would add to our problem formulation flavors of active learning and experimental design [18, 25, 17].

3.3 Learning Systematic Generalization

Before addressing the sample complexity of systematic generalization, it is worth considering the kind of interactions that we need in order to learn the causal transition model $P_{\mathcal{G}}$ and its underlying causal structure $\mathcal{G}$. Especially, thanks to the peculiar configuration of the causal structure $\mathcal{G}$, i.e., a bipartite graph in which the edges are necessarily directed from the state-action features X to the next state features Y, as a causation can only happen from the past to the future, learning the skeleton of $\mathcal{G}$ is equivalent to learning its full structure. Crucially, learning the skeleton of a causal graph does not need specific interventions, as it can be done from observational data alone [18].

Proposition 1. *The causal structure $\mathcal{G}$ of $\mathbb{U}$ can be identified from purely observational data.*

[4] W.l.o.g., we assume that the indices $i \in [M]$ refers to the $\mathcal{M}_i \in \mathbb{M}$, and $i \in (M, \infty)$ to the $\mathcal{M}_i \in \mathbb{U} \setminus \mathbb{M}$.

In this paper, we will consider the online learning setting with a *generative model* for estimating $\mathcal{G}$ and $P_{\mathcal{G}}$ from sampled interactions with a class $\mathbb{M}$ of size M. A generative model allows the agent to set the state of an MDP before sampling a transition, instead of drawing sequential interactions from the process. Finally, analogous results to what we obtain here can apply to the offline setting as well, in addition to convenient coverage assumptions on the dataset.

4 Sample Complexity Analysis

We provide a sample complexity analysis of the problem, which stands as a core contribution of this paper along with the problem formulation itself (§ 3). First, we consider the sample complexity of systematic generalization (§ 4.1). Then, we provide ancillary results on the estimation of the causal structure (§ 4.2) and the Bayesian network (§ 4.3) of an MDP, which can be of independent interest.

4.1 Sample Complexity of Systematic Generalization with a Generative Model

We have access to a class $\mathbb{M}$ of discrete MDPs within a universe $\mathbb{U}$, from which we draw interactions with a generative model $P(X)$. We aim to solve the systematic generalization problem as described in Definition 1. This problem requires to provide, for any combination of a (latent) MDP $\mathcal{M} \in \mathbb{U}$, and a given reward function r, a planning policy $\widehat{\pi}$ such that $V^*_{\mathcal{M},r} - V^{\widehat{\pi}}_{\mathcal{M},r} \leq \epsilon$. Especially, can we design an algorithm that guarantees this requirement with high probability by taking a number of samples K that is polynomial in ϵ and the relevant parameters of $\mathbb{M}$? Here we give a partially positive answer to this question, by providing a simple but provably efficient algorithm that guarantees systematic generalization over $\mathbb{U}$ up to an unavoidable sub-optimality term ϵ_λ that we will later specify.

The algorithm implements a model-based approach into two separated components. The first component, for which we report the pseudocode in Algorithm 1, is the procedure that actually interacts with the class $\mathbb{M}$ to obtain a principled estimation $\widehat{P}_{\widehat{\mathcal{G}}}$ of the causal transition model $P_{\mathcal{G}}$ of $\mathbb{U}$. The second, is a planning oracle that takes as input a reward function r and the estimated causal transition model, and returns an optimal policy $\widehat{\pi}$ operating on $\widehat{P}_{\widehat{\mathcal{G}}}$ as an approximation of the transition model P_i of the true MDP $\mathcal{M}_i$.[5]

Algorithm 1 Causal Transition Model Estimation

Input: class of MDPs $\mathbb{M}$, error ϵ, confidence δ
let $K' = C'\big(d_S^2 Z^2 n \log(2M d_S^2 d_A/\delta)\big/\epsilon^2\big)$
set the generative model $P(X) = \mathcal{U}_X$
for $i = 1, \dots, M$ **do**
  let $P_i(Y|X)$ the transition model of $\mathcal{M}_i \in \mathbb{M}$
  $\widehat{\mathcal{G}}_i \leftarrow$ *Causal Structure Est.* $(P_i, P(X), K')$
end for
let $\widehat{\mathcal{G}} = \cap_{i=1}^M \widehat{\mathcal{G}}_i$
let $K'' = C''\big(d_S^3 n^{3Z+1} \log(4 d_S n^Z/\delta)\big/\epsilon^2\big)$
let $P_{\mathbb{M}}(Y|X)$ be the mixture $\frac{1}{M}\sum_{i=1}^M P_i(Y|X)$
$\widehat{P}_{\widehat{\mathcal{G}}} \leftarrow$ *Bayesian Network Est.* $(P_{\mathbb{M}}, \widehat{\mathcal{G}}, K'')$
Output: causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$

First, we provide the sample complexity of the causal transition model estimation (Algorithm 1), which in turn is based on repeated causal structure estimations (Algorithm 3) to obtain $\widehat{\mathcal{G}}$, and an estimation procedure of the Bayesian network over $\widehat{\mathcal{G}}$ (Algorithm 2) to obtain $\widehat{P}_{\widehat{\mathcal{G}}}$.

Lemma 4.1. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^M$ be a class of M discrete MDPs, let $\delta \in (0,1)$, $\epsilon > 0$. The Alg. 1 returns a causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$ such that $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ with a sample complexity*

$$K = O\Big(M d_S^3 Z^2 n^{3Z+1} \log\big(\tfrac{4 M d_S^2 d_A n^Z}{\delta}\big)\Big/\epsilon^2\Big).$$

An analogous result can be derived for the tabular MDP setting, as stated by the following lemma.

Lemma 4.2. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^M$ be a class of M tabular MDPs. The result of Lemma 4.1 reduces to*

$$K = O\Big(M S^2 Z^2 2^{2Z} \log\big(\tfrac{4 M S^2 A 2^Z}{\delta}\big)\Big/\epsilon^2\Big).$$

Having established the sample complexity of the causal transition model estimation, we can now show how the learned model $\widehat{P}_{\widehat{\mathcal{G}}}$ allows us to approximately solve, via a planning oracle, any task defined by a combination of a latent MDP $\mathcal{M}_i \in \mathbb{U}$ and a given reward function r.

[5] Note that we do not lose generality here, as the planning oracle can be substituted with a principled approximate planning solver [see 20, Section 3.3].

Algorithm 3 MDP Causal Structure Estimation

Input: sampling model $P(Y|X)$, generative model $P(X)$, batch parameter K
draw $(x_k, y_k)_{k=1}^K \overset{\text{iid}}{\sim} P(Y|X)P(X)$
initialize $\widehat{\mathcal{G}} = \emptyset$
for each pair of nodes X_z, Y_j **do**
  compute the independence test $\mathbb{I}(X_z, Y_j)$
  if dependent add (X_z, Y_j) to $\widehat{\mathcal{G}}$
end for
Output: causal dependency graph $\widehat{\mathcal{G}}$

Algorithm 2 MDP Bayesian Network Estimation

Input: sampling model $P(Y|X)$, dependency graph $\mathcal{G}$, batch parameter K
let $K' = \lceil K/d_S n^Z \rceil$
for $j = 1, \dots, d_S$ **do**
  let Z_j the scopes $(X[Z_j], Y[j]) \subseteq \mathcal{G}$
  initialize the counts $N(X[Z_j], Y[j]) = 0$
  for each value $x \in [n]^{|Z_j|}$ **do**
    for $k = 1, \dots, K'$ **do**
      draw $y \sim P(Y[j]|X[Z_j] = x)$
      increment $N(X[Z_j] = x, Y[j] = y)$
    end for
  end for
  compute $\widehat{P}_j(Y[j]|X[Z_j]) = \frac{N(X[Z_j], Y[j])}{K'}$
end for
let $\widehat{P}_{\mathcal{G}}(Y|X) = \prod_{j=1}^{d_S} \widehat{P}_j(Y[j]|X[Z_j])$
Output: Bayesian network $\widehat{P}_{\mathcal{G}}$

To provide this result in the discrete MDP setting, we have to further assume that the transition dynamics P_i of the target MDP $\mathcal{M}_i$ admits factorization analogous to (1), such that we can write $P_i(Y|X) = \prod_{j=1}^{d_S} P_{i,j}(Y[j]|X[Z_j^{'}])$, where the scopes $Z_j^{'}$ are given by the environment-specific causal structure $\mathcal{G}_i$, which we assume to be $2Z$-sparse (Assumption 1).

Theorem 4.3. *Let $\delta \in (0,1)$ and $\epsilon > 0$. For a latent discrete MDP $\mathcal{M} \in \mathbb{U}$, and a given reward function r, a planning oracle operating on the causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$ as an approximation of $\mathcal{M}$ returns a policy $\widehat{\pi}$ such that $Pr\big(V^*_{\mathcal{M}_i,r} - V_{\mathcal{M}_i,r} \geq \epsilon_\lambda + \epsilon\big) \leq \delta$, where $\epsilon_\lambda = 2\lambda H^3 d_S n^{2Z+1}$, and $\widehat{P}_{\widehat{\mathcal{G}}}$ is obtained from Algorithm 1 with $\delta' = \delta$ and $\epsilon' = \frac{\epsilon}{2H^3 n^{Z+1}}$.*

Without the additional factorization of the environment-specific transition model, the result of Theorem 4.3 reduces to the analogous for the tabular MDP setting.

Corollary 4.4. *Let $\mathcal{M}$ a tabular MDP, the result of Th. 4.3 holds with $\epsilon_\lambda = 2\lambda SAH^3$, $\epsilon' = \frac{\epsilon}{2SAH^3}$.*

Theorem 4.3 and Corollary 4.4 establish the sample complexity of systematic generalization through Lemma 4.1 and Lemma 4.2 respectively. For the discrete MDP setting, we have that $\widetilde{O}(MH^6 d_S^3 Z^2 n^{5Z+3})$ are required, which reduces to $\widetilde{O}(MH^6 S^4 A^2 Z^2)$ in the tabular setting. Unfortunately, we are only able to obtain systematic generalization up to an unavoidable sub-optimality term ϵ_λ. This error term is related to the λ-sufficiency of the causal transition model (Assumption 2), and it accounts for the fact that $P_{\mathcal{G}}$ cannot fully explain the transition dynamics of each $\mathcal{M} \in \mathbb{U}$, even when it is estimated exactly. This is inherent to the ambitious problem setting, and can be only overcome with additional interactions with the test MDP $\mathcal{M}$.

4.2 Sample Complexity of Learning the Causal Structure of a Discrete MDP

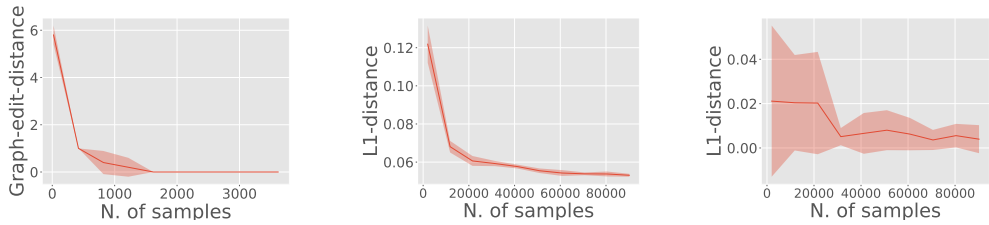
As a byproduct of the main result in Theorem 4.3, we can provide a sample complexity result for the problem of learning the causal structure $\mathcal{G}$ underlying a discrete MDP $\mathcal{M}$ with a generative model. We believe that this problem can be of independent interest, mainly in consideration of previous work on causal discovery of general stochastic processes [e.g., 43], for which we refine known results to account for the structure of an MDP, which allows for a tighter analysis of the sample complexity.

Instead of the exact dependency graph $\mathcal{G}$, which can include dependencies that are too weak to be detected with a finite number of samples, we only address the dependencies above a given threshold.

Definition 2. *We call $\mathcal{G}_\epsilon \subseteq \mathcal{G}$ the* ϵ-dependency subgraph *of $\mathcal{G}$ if it holds, for each pair $(A, B) \in \mathcal{G}$ distributed as $P_{A,B}$, $(A, B) \in \mathcal{G}_\epsilon$ if and only if* $\inf_{Q \in \{\Delta_A \times \Delta_B\}} \|P_{A,B} - Q\|_1 \geq \epsilon$.

Before presenting the result, we state the existence of a principled independence testing procedure.

Lemma 4.5 ([13])**.** *There exists an (ϵ, δ)-independence tester $\mathbb{I}(A, B)$ for distributions $P_{A,B}$ on $[n] \times [n]$, which returns* yes *if A, B are independent,* no *if* $\inf_{Q \in \{\Delta_A \times \Delta_B\}} \|P_{A,B} - Q\|_1 \geq \epsilon$*, both with probability at least $1 - \delta$ and sample complexity $O(n \log(1/\delta)/\epsilon^2)$.*



(a) Causal Structure (b) Causal Transition Model (c) Value Function

Figure 4: Estimation errors of key quantities as a function of the number of samples.

We can now provide an upper bound to the number of samples required by a simple estimation procedure to return an (ϵ, δ)-estimate $\widehat{\mathcal{G}}$ of the causal dependency graph $\mathcal{G}$.

Theorem 4.6. *Let $\mathcal{M}$ be a discrete MDP with an underlying causal structure $\mathcal{G}$, let $\delta \in (0, 1)$, and let $\epsilon > 0$. The Algorithm 3 returns a dependency graph $\widehat{\mathcal{G}}$ such that $Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon) \leq \delta$ with a sample complexity $K = O\big(n \log(d_S^2 d_A/\delta)/\epsilon^2\big)$.*

Corollary 4.7. *Let $\mathcal{M}$ a tabular MDP. The result of Th. 4.6 reduces to $K = O\big(\log(S^2 A/\delta)/\epsilon^2\big)$.*

4.3 Sample Complexity of Learning the Bayesian Network of a Discrete MDP

We present as a standalone result an upper bound to the sample complexity of learning the parameters of a Bayesian network $P_\mathcal{G}$ with a fixed structure $\mathcal{G}$. Especially, we refine known results [e.g., 9] by considering the specific structure $\mathcal{G}$ of an MDP. If the structure $\mathcal{G}$ is dense, the number of parameters of $P_\mathcal{G}$ grows exponentially, making the estimation problem mostly intractable. Thus, we consider a Z-sparse $\mathcal{G}$ (Assumption 1), as in previous works [9]. Then, we can provide a polynomial sample complexity for the problem of learning the Bayesian network $P_\mathcal{G}$ of a an MDP $\mathcal{M}$.

Theorem 4.8. *Let $\mathcal{M}$ be a discrete MDP, let $\mathcal{G}$ be its underlying causal structure, let $\delta \in (0, 1)$, and let $\epsilon > 0$. The Algorithm 2 returns a Bayesian network $\widehat{P}_\mathcal{G}$ such that $Pr(\|\widehat{P}_\mathcal{G} - P_\mathcal{G}\|_1 \geq \epsilon) \leq \delta$ with a sample complexity $K = O\big(d_S^3 n^{3Z+1} \log(d_S n^Z/\delta)/\epsilon^2\big)$.*

Corollary 4.9. *Let $\mathcal{M}$ a tabular MDP. The result of Th. 4.8 reduces to $K = O\big(S^2 2^{2Z} \log(\frac{S2^Z}{\delta})/\epsilon^2\big)$.*

5 Numerical Validation

We empirically validate the theoretical findings of this work by experimenting on a synthetic example where each environment is a person, and the MDP represents how a series of actions the person can take influences their weight (W) and academic performance (A). As actions we consider hours of physical training (P), hours of sleep (S), hours of study (St), amount of vegetables in the diet (D), and the amount of caffeine intake (C). The obvious use-case for such a model would be a tracking device that monitors how the actions of a person influence their weight and academic performance and provides personalized recommendations to reach the person's goals. While the physiological responses of different individuals can vary, there are some underlying mechanisms shared by all humans, and therefore deemed causal in our terminology. Examples of such causal links are the dependency of weight on the type of diet, and the dependency of academic performance on the number of hours of study. Other links, such as the dependency of weight on the amount of caffeine, are present in some individuals, but are generally not shared and therefore not causal. For simplicity, all variables are treated as discrete with possible values 0 (below average), 1 (average) or 2 (above average). See Appendix B for an illustration of the domain and details on how transition models of different environments are generated. A class $\mathbb{M}$ of 3 environments is used to estimate the causal transition model. All experiments are repeated 10 times and report the average and standard deviation.

Causal Structure Estimation We first empirically investigate the graph edit distance between estimated and ground-truth causal structures, i.e., $\text{GED}(\mathcal{G}, \widehat{\mathcal{G}})$, as a function of number of samples (K' in Algorithm 1). The causal structure is estimated by obtaining the causal graph for each training environment (using a series of independence tests), and taking the intersection of their edges. As expected, as we increase the number of samples, the distance converges to zero, and we can perfectly recover the causal graph (Figure 4a).

Causal Transition Model Estimation Figure 4b presents the L_1-distance between the estimate and ground truth causal transition model, as a function of the total number of samples ($K' + K''$ in Algorithm 1). As the number of samples grows, the L_1-distance shrinks, converging at around 0.05, an estimation error due to the environments not respecting the evenness assumption.

Value Function Estimation Finally, we investigate whether we can approximate the optimal value function for a new unseen environment. From Figure 4c, we observe that our algorithm is able to approximate the optimal value function up to a small error with a reasonable number of samples.

6 Related Work

Causal Discovery and Bayesian Networks On a technical level, our work is related to previous efforts on the sample complexity of causal discovery [43] and Bayesian network estimation [15, 9, 3]. None of these works is related to the specific MDP setting. In this paper, we account for the peculiar structure of the MDP to get sharper rates w.r.t. a blind application of previous results.

Reward-Free RL The reward-free RL formulation [20] is essentially akin to a particular case of our systematic generalization framework (Definition 1) in which the set of MDPs is a singleton $\mathbb{U} = \{\mathcal{M}\}$ instead of an infinite set of MDPs sharing a causal structure. Recent works have proposed provably efficient algorithms for reward-free RL, both in tabular [20, 22, 29, 50] and continuous settings with structure [44, 48, 34]. It is worth investigating how our sample complexity result compare to an approach that performs reward-free exploration independently for each MDP over a large set $\mathbb{U}$. Let $|\mathbb{U}| = U$, from [20, Theorem 4.1] we know that the agnostic reward-free approach would require at least $\Omega(UH^3S^2A/\epsilon^2)$ samples to obtain systematic generalization up to an ϵ threshold over a set of tabular MDPs $\mathbb{U}$. This compares favorably with our $\widetilde{O}(MH^6S^4A^2/\epsilon^2)$ complexity (see Corollary 4.4) whenever U is small, but leveraging the inner structure of $\mathbb{U}$ becomes crucial as U grows to infinity, while M remains constant. However, our approach pays this further generality with the additional error term ϵ_λ, which is unavoidable. Finally, it is an interesting direction for future work to establish whether the additional factors in S, A, H w.r.t. reward-free RL are also unavoidable.

Hidden Structures in RL Previous works have considered learning an hidden structure of the MDP for sample efficient RL, such as [14, 30, 31, 1]. Their focus is on learning latent representations of states assuming a linear structure in the MDP. This is orthogonal to our work, which instead targets the causal structure shared by infinitely many MDPs, while assuming access to the state features. Other works [e.g., 21, 4, 47] study the impact of structural properties of the MDP assuming access to the features. Our structural assumption is strictly more general than the linear structures they consider, but their work could provide useful inspiration to extend our results beyond discrete settings.

Model-Based RL The model-based RL [37] methodology prescribes learning an approximate model of transition dynamics in order to learn an optimal policy. Theoretical works [e.g., 19, 2] generally concern with the estimation of the approximate value functions obtained through the learned model, rather than the estimation of the model itself. A notable exception is the work in [39], which targets point-wise high probability guarantees on the model estimation as we do in Lemma 4.1, 4.2. However, they address the model estimation of a single tabular MDP $\mathcal{M}$, instead of the shared transition dynamics of an infinite set of MDPs $\mathbb{U}$ that we target in this paper.

Factored MDPs The factored MDP formalism [23] allows encoding transition dynamics that are the product of multiple independent factors. This is closely related to how we define the causal transition model as a product of independent factors in (1), which can be seen as a factored MDP. Previous works have considered learning in factored MDPs, either assuming full knowledge of the factorization [11, 46, 38, 40], or by estimating its structure from data [36, 42, 6, 32, 35]. To the best of our knowledge, none of the existing works have considered the factored MDP framework in combination with a reward-free setting and systematic generalization, which bring unique challenges to the identification of the underlying factorization and the estimation of the transition factors.

Causal RL Previous works [49, 41, 16] have also addressed model-based RL from a causal perspective. The motivations behind [49] are especially similar to ours. However, they have come to different structural assumptions, which lead to non-overlapping results. To the best of our knowledge, we are the first to prove a polynomial sample complexity for causal model-based RL in systematic generalization. Finally, [27] shows how to exploit a known causal representation of the MDP for sample efficient RL, which can complement our work on how to learn such a representation.

References

[1] 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*, 33:20095–20107, 2020.

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

[3] Arnab Bhattacharyya, Clément L Canonne, and Joy Qiping Yang. Independence testing for bounded degree bayesian network. *arXiv preprint arXiv:2204.08690*, 2022.

[4] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[5] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–57. IEEE, 2018.

[6] Doran Chakraborty and Peter Stone. Structure learning in ergodic factored mdps without knowledge of the transition function's in-degree. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 737–744. Citeseer, 2011.

[7] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[8] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning, 2018.

[9] Sanjoy Dasgupta. The sample complexity of learning fixed-structure bayesian networks. *Machine Learning*, 29(2):165–180, 1997.

[10] Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150, 1989.

[11] Karina Valdivia Delgado, Scott Sanner, and Leliane Nunes De Barros. Efficient solutions to factored mdps with imprecise transition probabilities. *Artificial Intelligence*, 175(9-10):1498–1527, 2011.

[12] Amir Dembo and Ofer Zeitouni. Ldp for finite dimensional spaces. In *Large deviations techniques and applications*, pages 11–70. Springer, 2009.

[13] Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 542–555, 2021.

[14] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[15] Nir Friedman and Zohar Yakhini. On the sample complexity of learning bayesian networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 274–282, 1996.

[16] Maxime Gasse, Damien Grasset, Guillaume Gaudron, and Pierre-Yves Oudeyer. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421*, 2021.

[17] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning, 2018.

[18] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, Jun 2014.

[19] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[20] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

[21] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[22] 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.

[23] Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999.

[24] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*, 2021.

[25] Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[26] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[27] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Efficient reinforcement learning with prior causal knowledge. In *First Conference on Causal Learning and Reasoning*, 2021.

[28] Jay Mardia, Jiantao Jiao, Ervin Tánczos, Robert D Nowak, and Tsachy Weissman. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9(4):813–850, 2020.

[29] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

[30] 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.

[31] Dipendra Misra, Qinghua Liu, Chi Jin, and John Langford. Provable rich observation reinforcement learning with combinatorial latent states. In *International Conference on Learning Representations*, 2020.

[32] Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. *Advances in Neural Information Processing Systems*, 27:604–612, 2014.

[33] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[34] Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pages 8737–8747. PMLR, 2021.

[35] Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored mdps with unknown structure. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[36] Alexander L Strehl, Carlos Diuk, and Michael L Littman. Efficient structure learning in factored-state mdps. In *AAAI*, volume 7, pages 645–650, 2007.

[37] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[38] Mohammad Sadegh Talebi, Anders Jonsson, and Odalric-Ambrym Maillard. Improved exploration in factored average-reward mdps. *arXiv preprint arXiv:2009.04575*, 2020.

[39] Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, pages 1019–1028. PMLR, 2020.

[40] Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

[41] Manan Tomar, Amy Zhang, Roberto Calandra, Matthew E Taylor, and Joelle Pineau. Model-invariant state abstractions for model-based reinforcement learning. *arXiv preprint arXiv:2102.09850*, 2021.

[42] Christopher M Vigorito and Andrew G Barto. Incremental structure learning in factored mdps with continuous states and actions. *University of Massachusetts Amherst-Department of Computer Science, Tech. Rep*, 2009.

[43] Samir Wadhwa and Roy Dong. On the sample complexity of causal discovery and the value of domain expertise. *arXiv preprint arXiv:2102.03274*, 2021.

[44] Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

[45] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[46] Ziping Xu and Ambuj Tewari. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. *arXiv preprint arXiv:2002.02302*, 2020.

[47] Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory*, pages 1165–1192, 2022.

[48] Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *arXiv preprint arXiv:2008.07737*, 2020.

[49] Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pages 11214–11224. PMLR, 2020.

[50] Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pages 12402–12412. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See lines 61-67 for a list of the main contributions with pointers.

(b) Did you describe the limitations of your work? [Yes] We do not include a specific section for the limitations, but we provided the assumptions under which our results apply (see § 3), and we discussed the sub-optimality term we cannot avoid with our approach (see Theorems 4.3, 4.4 and comments).

(c) Did you discuss any potential negative societal impacts of your work? [No] We believe our work should be categorised as fundamental research. Future works willing to employ the reported theoretical results in practical methodologies should consider potential negative impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] See § 3 and especially Assumptions 1, 2, 3, 4.

(b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Our paper is mainly theoretical. While we reported a brief numerical evaluation, the full set of experiments are straightforward to reproduce with the information provided in § 5 and Appendix B.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See § 5 and Appendix B.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All the experiments are repeated 10 times and we report mean and standard deviation.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The needed computation is negligible.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

Proofs of Section 3.3

Proposition 1. *The causal structure $\mathcal{G}$ of $\mathbb{U}$ can be identified from purely observational data.*

Proof. First, recall that with observational data alone, a causal graph can be identified up to its Markov equivalence class [18]. This means that its skeleton and v-structure are properly identified, meanwhile determining the edge orientations requires interventional data in the general case. Since in the considered causal graph $\mathcal{G}$ the edges orientations are determined a priori (as they follow the direction of time), the causal graph can be entirely determined by using only observational data. □

Proofs of Section 4.1: Causal Transition Model Estimation

Before reporting the proof of the main result in Theorem 4.3, it is worth considering a set of lemmas that will be instrumental to the main proof.

First, we provide an upper bound to the L1-norm between the Bayesian network $P_{\mathcal{G}}$ over a given structure $\mathcal{G}$ and the Bayesian network $P_{\mathcal{G}_\epsilon}$ over the structure $\mathcal{G}_\epsilon$, which is the ϵ-dependency subgraph of $\mathcal{G}$ as defined in Definition 2.

Lemma A.1. *Let $\mathcal{G}$ a Z-sparse dependency graph, and let $\mathcal{G}_\epsilon$ its corresponding ϵ-dependence subgraph for a threshold $\epsilon > 0$. The L1-norm between the Bayesian network $P_{\mathcal{G}}$ over $\mathcal{G}$ and the Bayesian network $P_{\mathcal{G}_\epsilon}$ over $\mathcal{G}_\epsilon$ can be upper bounded as*

$$\|P_{\mathcal{G}} - P_{\mathcal{G}_\epsilon}\|_1 \leq d_S Z \epsilon.$$

Proof. The proof is based on the fact that every edge (X_i, Y_j) such that $(X_i, Y_j) \in \mathcal{G}$ and $(X_i, Y_j) \notin \mathcal{G}_\epsilon$ corresponds to a weak conditional dependence (see Definition 2), which means that $\|P_{Y_j|X_i} - P_{Y_j}\|_1 \leq \epsilon$.

We denote with Z_j the scopes of the parents of the node $Y[j]$ in $\mathcal{G}$, i.e., $\mathrm{Pa}_{\mathcal{G}}(Y[j]) = X[Z_j]$, and with $Z_{j,\epsilon}$ the scopes of the parents of the node $Y[j]$ in $\mathcal{G}_\epsilon$, i.e., $\mathrm{Pa}_{\mathcal{G}_\epsilon}(Y[j]) = X[Z_{j,\epsilon}]$. As a direct consequence of Definition 2, we have $Z_{j,\epsilon} \subseteq Z_j$ for any $j \in d_S$, and we can write

$$P_{\mathcal{G}}(Y|X) = \prod_{j=1}^{d_S} P_j(Y[j] \mid X[Z_j]) = \prod_{j=1}^{d_S} P_j(Y[j] \mid X[Z_{j,\epsilon}], X[Z_j \setminus Z_{j,\epsilon}]),$$

$$P_{\mathcal{G}_\epsilon}(Y|X) = \prod_{j=1}^{d_S} P_j(Y[j] \mid X[Z_{j,\epsilon}]).$$

Then, we let $Z_j \setminus Z_{j,\epsilon} = [I]$ overwriting the actual indices for the sake of clarity, and we derive

$$\begin{aligned}
&\|P_{\mathcal{G}} - P_{\mathcal{G}_\epsilon}\|_1 \\
&\leq \sum_{j=1}^{d_S} \left\| P_j(Y[j] \mid X[Z_{j,\epsilon}], \cup_{i=1}^{I} X[i]) - P_j(Y[j] \mid X[Z_{j,\epsilon}]) \right\|_1 && (2)\\
&\leq \sum_{j=1}^{d_S} \sum_{i'=1}^{I} \left\| P_j(Y[j] \mid X[Z_{j,\epsilon}], \cup_{i=i'}^{I} X[i]) - P_j(Y[j] \mid X[Z_{j,\epsilon}], \cup_{i=i'+1}^{I} X[i]) \right\|_1 && (3)\\
&\leq \sum_{j=1}^{d_S} \sum_{i'=1}^{I} \epsilon \leq d_S Z \epsilon, && (4)
\end{aligned}$$

in which we employed the property $\|\mu - \nu\|_1 \leq \|\prod_i \mu_i - \prod_i \nu_i\|_1 \leq \sum_i \|\mu_i - \nu_i\|_1$ for the L1-norm between product distributions $\mu = \prod_i \mu_i, \nu = \prod_i \nu_i$ to write (2), we repeatedly applied the triangle inequality $\|\mu - \nu\|_1 \leq \|\mu - \rho\|_1 + \|\rho - \nu\|_1$ to get (3) from (2), we upper bounded each term of the sum in (3) with ϵ thanks to Definition 2, and we finally employed the Z-sparseness Assumption 1 to upper bound I with Z in (4). □

Next, we provide a crucial sample complexity result for a provably efficient estimation of a Bayesian network $\widehat{P}_{\widehat{\mathcal{G}}}$ over an estimated ϵ-dependency subgraph $\widehat{\mathcal{G}}$, which relies on both the causal structure estimation result of Theorem 4.6 and the Bayesian network estimation result of Theorem 4.8.

Lemma A.2. *Let $\mathcal{M}$ be a discrete MDP, let $\mathbb{M} = \{\mathcal{M}\}$ be a singleton class, let $\delta \in (0,1)$, and let $\epsilon > 0$. The Algorithm 1 returns a Bayesian network $\widehat{P}_{\widehat{\mathcal{G}}}$ such that $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ with a sample complexity*

$$K = O\left(\frac{d_S^3 Z^2 n^{3Z+1} \log\left(\frac{4d_S^2 d_A n^Z}{\delta}\right)}{\epsilon^2}\right).$$

Proof. We aim to obtain the number of samples $K = K' + K''$ for which Algorithm 1 is guaranteed to return a Bayesian network estimate $\widehat{P}_{\widehat{\mathcal{G}}}$ over a causal structure estimate $\widehat{\mathcal{G}}$ such that $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ in a setting with a singleton class of discrete MDPs. First, we derive the following decomposition of the error

$$\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \leq \|\widehat{P}_{\widehat{\mathcal{G}}} \pm P_{\widehat{\mathcal{G}}} \pm P_{\mathcal{G}_{\epsilon'}} - P_{\mathcal{G}}\|_1 \leq \|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\widehat{\mathcal{G}}}\|_1 + \|P_{\widehat{\mathcal{G}}} - P_{\mathcal{G}_{\epsilon'}}\|_1 + \|P_{\mathcal{G}_{\epsilon'}} - P_{\mathcal{G}}\|_1 \quad (5)$$

in which we employed the triangle inequality $\|\mu - \nu\|_1 \leq \|\mu - \rho\|_1 + \|\rho - \nu\|_1$. Then, we can write

$$Pr\big(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon\big) \leq \underbrace{Pr\Big(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\widehat{\mathcal{G}}}\|_1 \geq \frac{\epsilon}{3}\Big)}_{\text{Bayesian network estimation } (\star)} + \underbrace{Pr\Big(\|P_{\widehat{\mathcal{G}}} - P_{\mathcal{G}_{\epsilon'}}\|_1 \geq \frac{\epsilon}{3}\Big)}_{\text{causal structure estimation } (\bullet)} + \underbrace{Pr\Big(\|P_{\mathcal{G}_{\epsilon'}} - P_{\mathcal{G}}\|_1 \geq \frac{\epsilon}{3}\Big)}_{\text{Bayesian network subgraph } (\diamond)}$$

through the decomposition (5) and a union bound to isolate the three independent sources of error $(\star), (\bullet), (\diamond)$. To upper bound the latter term $(\diamond)$ with 0, we invoke Lemma A.1 to have $d_s Z \epsilon' \leq \frac{\epsilon}{3}$, which gives $\epsilon' \leq \frac{\epsilon}{3d_S Z}$. Then, we consider the middle term $(\bullet)$, for which we can write

$$Pr\left(\|P_{\widehat{\mathcal{G}}_{\epsilon'}} - P_{\mathcal{G}_{\epsilon'}}\|_1 \geq \frac{\epsilon}{3}\right) \leq Pr\big(\widehat{\mathcal{G}} \neq \mathcal{G}_{\epsilon}\big). \quad (6)$$

We can now upper bound $(\bullet) \leq \delta/2$ through (6) by invoking Theorem 4.6 with threshold $\epsilon' = \frac{\epsilon}{3d_S Z}$ and confidence $\delta' = \frac{\delta}{2}$, which gives

$$K' = C'\left(\frac{d_S^{4/3} Z^{4/3} n \log^{1/3}(2d_S^2 d_A/\delta)}{\epsilon^{4/3}} + \frac{d_S^2 Z^2 n \log^{1/2}(2d_S^2 d_A/\delta) + \log(2d_S^2 d_A/\delta)}{\epsilon^2}\right). \quad (7)$$

Next, we can upper bound $(\star) \leq \delta/2$ by invoking Theorem 4.8 with threshold $\epsilon' = \frac{\epsilon}{3}$ and confidence $\delta' = \frac{\delta}{2}$, which gives

$$K'' = C''\left(\frac{d_S^3 n^{3Z+1} \log(4d_S n^Z/\delta)}{\epsilon^2}\right). \quad (8)$$

Finally, through the combination of (7) and (8), we can derive the sample complexity that guarantees $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ under the assumption $\epsilon^{4/3} \ll \epsilon^2$, i.e.,

$$K = K' + K'' \leq \frac{d_S^3 Z^2 n^{3Z+1} \log\left(\frac{4d_S^2 d_A n^Z}{\delta}\right)}{\epsilon^2},$$

which concludes the proof. □

Whereas Lemma A.2 is concerned with the sample complexity of learning the Bayesian network of a singleton class, we can now extend the result to account for a class $\mathbb{M}$ composed of M discrete MDPs.

Lemma 4.1. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^M$ be a class of M discrete MDPs, let $\delta \in (0,1)$, $\epsilon > 0$. The Alg. 1 returns a causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$ such that $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ with a sample complexity*

$$K = O\Big(M d_S^3 Z^2 n^{3Z+1} \log\big(\tfrac{4M d_S^2 d_A n^Z}{\delta}\big) \Big/ \epsilon^2\Big).$$

Proof. We aim to obtain the number of samples $K = MK' + K''$ for which Algorithm 1 is guaranteed to return a Bayesian network estimate $\widehat{P}_{\widehat{\mathcal{G}}}$ over a causal structure estimate $\widehat{\mathcal{G}}$ such that $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ in a setting with a class of M discrete MDPs. First, we can derive an analogous decomposition as in (5), such that we have

$$Pr\big(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon\big) \leq \underbrace{Pr\Big(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\widehat{\mathcal{G}}}\|_1 \geq \frac{\epsilon}{3}\Big)}_{\text{Bayesian network estimation } (\star)} + \underbrace{Pr\Big(\|P_{\widehat{\mathcal{G}}} - P_{\mathcal{G}_{\epsilon'}}\|_1 \geq \frac{\epsilon}{3}\Big)}_{\text{causal structure estimation } (\bullet)} + \underbrace{Pr\Big(\|P_{\mathcal{G}_{\epsilon'}} - P_{\mathcal{G}}\|_1 \geq \frac{\epsilon}{3}\Big)}_{\text{Bayesian network subgraph } (\diamond)}$$

through a union bound. Crucially, the terms $(\star), (\diamond)$ are unaffected by the class size, which leads to $K'' =$ (8) by upper bounding $(\star)$, and $\epsilon' \leq \frac{\epsilon}{3d_S Z}$ by upper bounding $(\diamond)$, exactly as in the proof of Lemma A.2. Instead, the number of samples K' has to guarantee that $(\bullet) = Pr(\|P_{\widehat{\mathcal{G}}} - P_{\mathcal{G}_{\epsilon'}}\|_1 \geq \epsilon/3) \leq \delta/2$, where the causal structure $\mathcal{G}_{\epsilon'}$ is now the intersection of the causal structures of the single class components $\mathcal{M}_i$, i.e., $\mathcal{G}_{\epsilon'} = \cap_{i=1}^{M} \mathcal{G}_{\epsilon',i}$. Especially, we can write

$$(\bullet) = Pr\Big(\|P_{\widehat{\mathcal{G}}} - P_{\mathcal{G}_{\epsilon'}}\|_1 \geq \frac{\epsilon}{3}\Big) \leq Pr\Big(\widehat{\mathcal{G}} \neq \mathcal{G}_{\epsilon'}\Big) \leq Pr\Big(\bigcup_{i=1}^{M} \widehat{\mathcal{G}}_i \neq \mathcal{G}_{\epsilon',i}\Big) \leq \sum_{i=0}^{M} Pr\Big(\widehat{\mathcal{G}}_i \neq \mathcal{G}_{\epsilon',i}\Big), \tag{9}$$

through a union bound on the estimation of the single causal structures $\widehat{\mathcal{G}}_i$. Then, we can upper bound $(\bullet) \leq \delta/2$ through (9) by invoking Theorem 4.6 with threshold $\epsilon' = \frac{\epsilon}{3d_S Z}$ and confidence $\delta' = \frac{\delta}{2M}$, which gives

$$K' = C'\Big(\frac{d_S^{4/3} Z^{4/3} n \log^{1/3}(2Md_S^2 d_A/\delta)}{\epsilon^{4/3}} + \frac{d_S^2 Z^2 n \log^{1/2}(2Md_S^2 d_A/\delta) + \log(2Md_S^2 d_A/\delta)}{\epsilon^2}\Big). \tag{10}$$

Finally, through the combination of (10) and (8), we can derive the sample complexity that guarantees $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ under the assumption $\epsilon^{4/3} \ll \epsilon^2$, i.e.,

$$K = MK' + K'' \leq \frac{Md_S^3 Z^2 n^{3Z+1} \log\left(\frac{4Md_S^2 d_A n^Z}{\delta}\right)}{\epsilon^2},$$

which concludes the proof. □

It is now straightforward to extend Lemma 4.1 for a class $\mathbb{M}$ composed of M tabular MDPs.

Lemma 4.2. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^{M}$ be a class of M tabular MDPs. The result of Lemma 4.1 reduces to*

$$K = O\Big(MS^2 Z^2 2^{2Z} \log\big(\tfrac{4MS^2 A 2^Z}{\delta}\big)\Big/\epsilon^2\Big).$$

Proof. To obtain $K = MK' + K''$, we follows similar steps as in the proof of Lemma 4.1, to have the usual decomposition of the event $Pr(\|\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}\|_1 \geq \epsilon)$ in the $(\star), (\bullet), (\diamond)$ terms. We can deal with $(\diamond)$ as in Lemma 4.1 to get $\epsilon' \leq \frac{\epsilon}{3SZ}$. Then, we upper bound $(\bullet) \leq \delta/2$ by invoking Corollary 4.7 (instead of Theorem 4.6) with threshold $\epsilon' = \frac{\epsilon}{3SZ}$ and confidence $\delta' = \frac{\delta}{2M}$, which gives

$$K' = C'\Big(\frac{S^{4/3} Z^{4/3} \log^{1/3}(2MS^2 A/\delta)}{\epsilon^{4/3}} + \frac{S^2 Z^2 \log^{1/2}(2MS^2 A/\delta) + \log(2MS^2 A/\delta)}{\epsilon^2}\Big). \tag{11}$$

Similarly, we upper bound $(\star) \leq \delta/2$ by invoking Corollary 4.9 (instead of Theorem 4.5) with threshold $\epsilon' = \frac{\epsilon}{3}$ and confidence $\delta' = \frac{\delta}{2}$, which gives

$$K'' = \frac{18S^2 2^{2Z} \log(4S2^Z/\delta)}{\epsilon^2}. \tag{12}$$

Finally, we combine 11 with 12 to obtain

$$K = MK' + K'' \leq \frac{MS^2 Z^2 2^{2Z} \log\big(\frac{4MS^2 A 2^Z}{\delta}\big)}{\epsilon^2}$$

□

Proofs of Section 4.1: Planning

Theorem 4.3. *Let $\delta \in (0,1)$ and $\epsilon > 0$. For a latent discrete MDP $\mathcal{M} \in \mathbb{U}$, and a given reward function r, a planning oracle operating on the causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$ as an approximation of $\mathcal{M}$ returns a policy $\widehat{\pi}$ such that $Pr\big(V^*_{\mathcal{M}_i,r} - V_{\mathcal{M}_i,r} \geq \epsilon_\lambda + \epsilon\big) \leq \delta$, where $\epsilon_\lambda = 2\lambda H^3 d_S n^{2Z+1}$, and $\widehat{P}_{\widehat{\mathcal{G}}}$ is obtained from Algorithm 1 with $\delta' = \delta$ and $\epsilon' = \frac{\epsilon}{2H^3 n^{Z+1}}$.*

Proof. Consider the MDPs with transition model P and $\widehat{P}_{\widehat{\mathcal{G}}}$. We refer to the respective optimal policies as π^* and $\widehat{\pi}^*$. Moreover, since the reward r is fixed, we remove it from the expressions for the sake of clarity, and refer with $\widehat{V}$ to the value function of the MDP with transition model $\widehat{P}_{\widehat{\mathcal{G}}}$. As done in [20, Theorem 3.5], we can write the following decomposition with $V^* := V^{\pi^*}$,

$$\begin{aligned}
\mathbb{E}_{s_1\sim P}\left[V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1)\right] &\leq \underbrace{\left|\mathbb{E}_{s_1\sim P}\left[V_1^*(s_1) - \widehat{V}_1^{\widehat{\pi}^*}(s_1)\right]\right|}_{\text{evaluation error}} + \underbrace{\mathbb{E}_{s_1\sim P}\left[\widehat{V}_1^*(s_1) - \widehat{V}_1^{\widehat{\pi}^*}(s_1)\right]}_{\leq 0 \text{ by def.}} \\
&\quad + \underbrace{\mathbb{E}_{s_1\sim P}\left[\widehat{V}_1^{\widehat{\pi}^*}(s_1) - \widehat{V}_1^{\widehat{\pi}}(s_1)\right]}_{\text{optimization error}} + \underbrace{\left|\mathbb{E}_{s_1\sim P}\left[\widehat{V}_1^{\widehat{\pi}}(s_1) - V_1^{\widehat{\pi}}(s_1)\right]\right|}_{\text{evaluation error}} \\
&\leq \underbrace{2n^{Z+1}H^3\epsilon'}_{\epsilon} + \underbrace{2n^{2Z+1}d_S H^3\lambda}_{\epsilon_\lambda}
\end{aligned}$$

where in the last step we have set to 0 the approximation due to the planning oracle assumption, and we have bounded the evaluation errors according to Lemma A.3. In order to get $2n^{Z+1}H^3\epsilon' = \epsilon$ we have to set $\epsilon' = \frac{\epsilon}{2n^{Z+1}H^3}$. Considering the sample complexity result in Lemma 4.1 the final sample complexity will be

$$K = O\left(\frac{M\ d_S^3\ Z^2\ n^{3Z+1}\ \log\left(\frac{4Md_S^2 d_A n^Z}{\delta}\right)}{(\epsilon')^2}\right) = O\left(\frac{4\ M\ d_S^3\ Z^2\ n^{5Z+3}\ H^6\ \log\left(\frac{4Md_S^2 d_A n^Z}{\delta}\right)}{\epsilon^2}\right).$$

□

Lemma A.3. *Under the preconditions of Theorem 4.3, with probability $1-\delta$, for any reward function r and policy π, we can bound the value function estimation error as follows.*

$$\left|\mathbb{E}_{s\sim P}\left[\widehat{V}_{1,r}^{\pi}(s) - V_{1,r}^{\pi}(s)\right]\right| \leq \underbrace{n^{Z+1}H^3\epsilon'}_{\epsilon} + \underbrace{n^{2Z+1}d_S H^3\lambda}_{\epsilon_\lambda} \tag{13}$$

where $\widehat{V}$ is the value function of the MDP with transition model $\widehat{P}_{\widehat{\mathcal{G}}}$, ϵ' is the approximation error between $\widehat{P}_{\widehat{\mathcal{G}}}$ and P_G studied in Lemma 4.1, and λ stands for the λ-sufficiency parameter of $P_{\mathcal{G}}$.

Proof. The proof will be along the lines of that of Lemma 3.6 in [20]. We first recall [8, Lemma E.15], which we restate in Lemma A.5. In this proof, we consider an environment specific true MDP $\mathcal{M}$ with transition model P, and an MDP $\widehat{\mathcal{M}}$ that has as transition model the estimated causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$. In the following, the expectations will be w.r.t. P. Moreover, since the reward r is fixed, we remove it from the expressions for the sake of clarity. We can start deriving

$$\begin{aligned}
\left|\mathbb{E}_{s\sim P}\left[\widehat{V}_1^{\pi}(s) - V_1^{\pi}(s)\right]\right| &\leq \left|\mathbb{E}_X\left[\sum_{h=1}^{H}(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}_{h+1}^{\pi}(X)\right]\right| \\
&\leq \mathbb{E}_X\left[\sum_{h=1}^{H}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}_{h+1}^{\pi}(X)\right|\right] \\
&= \sum_{h=1}^{H}\mathbb{E}_X\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}_{h+1}^{\pi}(X)\right|.
\end{aligned} \tag{14}$$

We now bound a single term within the sum above as follows

$$\begin{aligned}
\mathbb{E}_X \left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}_{h+1}(X)\right| &= \mathbb{E}_X \left|(\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}} + P_{\mathcal{G}} - P)\widehat{V}^{\pi}_{h+1}(X)\right| \\
&= \mathbb{E}_X \left|(\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^{\pi}_{h+1}(X) + (P_{\mathcal{G}} - P)\widehat{V}^{\pi}_{h+1}(X)\right| \\
&\leq \mathbb{E}_X \left[\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^{\pi}_{h+1}(X)\right| + \left|(P_{\mathcal{G}} - P)\widehat{V}^{\pi}_{h+1}(X)\right|\right] \\
&= \mathbb{E}_X \left|(\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^{\pi}_{h+1}(X)\right| + \mathbb{E}_X \left|(P_{\mathcal{G}} - P)\widehat{V}^{\pi}_{h+1}(X)\right|. \qquad (15)
\end{aligned}$$

We can now bound each term. Let us start considering the first one

$$\begin{aligned}
&\mathbb{E}_X \left|(\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^{\pi}_{h+1}(X)\right| \\
&\quad= \mathbb{E}_X \left|\widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^{\pi}_{h+1}(X) - P_{\mathcal{G}}\widehat{V}^{\pi}_{h+1}(X)\right| \\
&\quad= \mathbb{E}_X \Big|\sum_Y \widehat{P}_{\widehat{\mathcal{G}}}(Y|X)\widehat{V}^{\pi}_{h+1}(Y) - \sum_Y P_{\mathcal{G}}(Y|X)\widehat{V}^{\pi}_{h+1}(Y)\Big| \\
&\quad= \mathbb{E}_X \Big|\sum_Y \widehat{P}_{\widehat{\mathcal{G}}}(Y|X)\,\mathbb{E}_{X'\sim\pi}\left[r(X') + \widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^{\pi}_{h+2}(X')\right] \\
&\qquad - \sum_Y P_{\mathcal{G}}(Y|X)\,\mathbb{E}_{X'\sim\pi}\left[r(X') + P_{\mathcal{G}}\widehat{V}^{\pi}_{h+2}(X')\right]\Big| \\
&\quad= \mathbb{E}_X \Big|\sum_Y \left(\widehat{P}_{\widehat{\mathcal{G}}}(Y|X) - P_{\mathcal{G}}(Y|X)\right)\mathbb{E}_{X'\sim\pi}\left[r(X')\right] \\
&\qquad + \sum_Y \widehat{P}_{\widehat{\mathcal{G}}}(Y|X)\,\mathbb{E}_{X'\sim\pi}\left[\widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^{\pi}_{h+2}(X')\right] - \sum_Y P_{\mathcal{G}}(Y|X)\,\mathbb{E}_{X'\sim\pi}\left[P_{\mathcal{G}}\widehat{V}^{\pi}_{h+2}(X')\right]\Big| \\
&\quad\leq \mathbb{E}_X \Big|\sum_Y \left(\widehat{P}_{\widehat{\mathcal{G}}}(Y|X) - P_{\mathcal{G}}(Y|X)\right)\Big| \qquad (16) \\
&\qquad + \mathbb{E}_X \Big|\sum_Y \widehat{P}_{\widehat{\mathcal{G}}}(Y|X)\,\mathbb{E}_{X'\sim\pi}\left[\widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^{\pi}_{h+2}(X')\right] - \sum_Y P_{\mathcal{G}}(Y|X)\,\mathbb{E}_{X'\sim\pi}\left[P_{\mathcal{G}}\widehat{V}^{\pi}_{h+2}(X')\right]\Big|.
\end{aligned}$$

We can now bound the first term of (16)

$$\begin{aligned}
&\mathbb{E}_X \Big|\sum_Y \left(\widehat{P}_{\widehat{\mathcal{G}}}(Y|X) - P_{\mathcal{G}}(Y|X)\right)\Big| \\
&\qquad= \mathbb{E}_X \left|\sum_Y \Big(\prod_{j=1}^{d_S} \widehat{P}_j(Y[j]|X[Z_j]) - \prod_{j=1}^{d_S} P_j(Y[j]|X[Z_j])\Big)\right| \\
&\qquad\leq \mathbb{E}_X \left[\sum_Y \sum_{j=1}^{d_S} \left|\widehat{P}_j(Y[j]|X[Z_j]) - P_j(Y[j]|X[Z_j])\right|\right] \\
&\qquad= \sum_X P^{\pi}_{\mathcal{G}}(X) \left[\sum_Y \sum_{j=1}^{d_S} \left|\widehat{P}_j(Y[j]|X[Z_j]) - P_j(Y[j]|X[Z_j])\right|\right] \\
&\qquad= \sum_Y \sum_{j=1}^{d_S} \sum_{X[Z_j]} P^{\pi}_{\mathcal{G}}(X[Z_j]) \left|\widehat{P}_j(Y[j]|X[Z_j]) - P_j(Y[j]|X[Z_j])\right|. \qquad (17)
\end{aligned}$$

Due to the uniform sampling and Z-sparseness assumptions, we have $P_{\mathcal{G}}(X[Z_j]) = \frac{1}{n^Z}$, hence

$$\max_{\pi^\dagger} \frac{P^{\pi^\dagger}_{\mathcal{G}}(X[Z_j])}{P_{\mathcal{G}}(X[Z_j])} \leq \frac{1}{P_{\mathcal{G}}(X[Z_j])} = n^Z.$$

Therefore, we have $P_{\mathcal{G}}^{\pi^\dagger}(X[Z_j]) \leq n^Z \cdot P_{\mathcal{G}}(X[Z_j])$. Replacing this in (17) and marginalizing over $Y \setminus Y[j]$ we obtain

$$\begin{aligned}
&\mathbb{E}_X \left| \sum_Y \left(\widehat{P}_{\widehat{\mathcal{G}}}(Y|X) - P_{\mathcal{G}}(Y|X) \right) \right| \\
&\qquad = n^Z \sum_{j=1}^{d_s} \sum_{Y[j]} \sum_{X[Z_j]} \left| \widehat{P}_j(Y[j]|X[Z_j]) - P_j(Y[j]|X[Z_j]) \right| P_{\mathcal{G}}(X[Z_j]) \\
&\qquad \leq n^Z \sum_{j=1}^{d_S} \sum_{Y[j]} \frac{\epsilon'}{d_S} \sum_{X[Z_j]} P_{\mathcal{G}}(X[Z_j]) \\
&\qquad = n^{Z+1}\epsilon',
\end{aligned}$$

where $\frac{\epsilon'}{d_S}$ is the approximation term of each component. By plugging this bound into (16) we get

$$\begin{aligned}
&\mathbb{E}_X \left| (\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}}) \widehat{V}_{h+1}^{\pi}(X) \right| \\
&\qquad \leq n^{Z+1}\epsilon' + \mathbb{E}_X \left| \sum_Y \widehat{P}_{\widehat{\mathcal{G}}}(Y|X)\, \mathbb{E}_{X' \sim \pi} \left[\widehat{P}_{\widehat{\mathcal{G}}} \widehat{V}_{h+2}^{\pi}(X') \right] - \sum_Y P_{\mathcal{G}}(Y|X)\, \mathbb{E}_{X' \sim \pi} \left[P_{\mathcal{G}} \widehat{V}_{h+2}^{\pi}(X') \right] \right| \\
&\qquad \leq \sum_{i=h+1}^{H} i \cdot n^{Z+1}\epsilon' \\
&\qquad \leq H^2 n^{Z+1}\epsilon'
\end{aligned}$$

where in the last step we have recursively bounded the right terms as in (24). By considering $2Z$-sparseness, λ-sufficiency, and that the transition model P factorizes, we can apply the same procedure to bound the second term of equation (15) as

$$\mathbb{E}_X \left| (P_{\mathcal{G}} - P) \widehat{V}_{h+1}^{\pi}(X) \right| \leq H^2 n^{Z+1} d_S \lambda.$$

Therefore, the initial expression in (14) becomes

$$\begin{aligned}
\left| \mathbb{E}_{s \sim P} \left[\widehat{V}_1^{\pi}(s) - V_1^{\pi}(s) \right] \right| &\leq \sum_{h=1}^{H} \mathbb{E}_X \left| (\widehat{P}_{\widehat{\mathcal{G}}} - P) \widehat{V}_{h+1}^{\pi}(X) \right| && (18) \\
&\leq \sum_{h=1}^{H} [n^{Z+1} H^2 \epsilon' + n^{2Z+1} d_S H^2 \lambda] && (19) \\
&\leq \underbrace{n^{Z+1} H^3 \epsilon'}_{\epsilon} + \underbrace{n^{2Z+1} d_S H^3 \lambda}_{\epsilon_\lambda}. && (20)
\end{aligned}$$

□

Corollary 4.4. *Let $\mathcal{M}$ a tabular MDP, the result of Th. 4.3 holds with $\epsilon_\lambda = 2\lambda SAH^3$, $\epsilon' = \frac{\epsilon}{2SAH^3}$.*

Proof. Consider the MDPs with transition model P and $\widehat{P}_{\widehat{\mathcal{G}}}$. We refer to the respective optimal policies as π^* and $\widehat{\pi}^*$. Moreover, since the reward r is fixed, we remove it from the expressions for the sake of clarity, and refer with $\widehat{V}$ to the value function of the MDP with transition model $\widehat{P}_{\widehat{\mathcal{G}}}$. As done in [20, Theorem 3.5], we can write the following decomposition with $V^* := V^{\pi^*}$

$$\begin{aligned}
\mathbb{E}_{s_1 \sim P} \left[V_1^*(s_1) - V_1^{\widehat{\pi}}(s_1) \right] &\leq \underbrace{\left| \mathbb{E}_{s_1 \sim P} \left[V_1^*(s_1) - \widehat{V}_1^{\widehat{\pi}^*}(s_1) \right] \right|}_{\text{evaluation error}} + \underbrace{\mathbb{E}_{s_1 \sim P} \left[\widehat{V}_1^*(s_1) - \widehat{V}_1^{\widehat{\pi}^*}(s_1) \right]}_{\leq 0 \text{ by def.}} \\
&\quad + \underbrace{\mathbb{E}_{s_1 \sim P} \left[\widehat{V}_1^{\widehat{\pi}^*}(s_1) - \widehat{V}_1^{\widehat{\pi}}(s_1) \right]}_{\text{optimization error}} + \underbrace{\left| \mathbb{E}_{s_1 \sim P} \left[\widehat{V}_1^{\widehat{\pi}}(s_1) - V_1^{\widehat{\pi}}(s_1) \right] \right|}_{\text{evaluation error}} \\
&\leq \underbrace{2SAH^3\epsilon'}_{\epsilon} + \underbrace{2SAH^3\lambda}_{\epsilon_\lambda}
\end{aligned}$$

where in the last step we have set to 0 the approximation due to the planning oracle assumption, and we have bounded the evaluation errors according to Lemma A.4. In order to get $2SAH^3\epsilon' = \epsilon$ we have to set $\epsilon' = \frac{\epsilon}{2SAH^3}$. Considering the sample complexity result in Lemma 4.2 the final sample complexity will be

$$K = O\left(\frac{M\ S^2\ Z^2\ 2^{2Z}\ \log\left(\frac{4MS^2A2^Z}{\delta}\right)}{(\epsilon')^2}\right) = O\left(\frac{4M\ S^4\ A^2\ H^6\ Z^2\ 2^{2Z}\ \log\left(\frac{4MS^2A2^Z}{\delta}\right)}{\epsilon^2}\right).$$

□

Lemma A.4. *Under the preconditions of Corollary 4.4, with probability $1-\delta$, for any reward function r and policy π, we can bound the value function estimation error as follows.*

$$\left|\mathbb{E}_{s\sim P}\left[\hat{V}^{\pi}_{1,r}(s) - V^{\pi}_{1,r}(s)\right]\right| \leq \underbrace{SAH^3\epsilon'}_{\epsilon} + \underbrace{SAH^3\lambda}_{\epsilon_\lambda} \tag{21}$$

where $\widehat{V}$ is the value function of the MDP with transition model $\widehat{P}_{\widehat{\mathcal{G}}}$, ϵ' is the approximation error between $\widehat{P}_{\widehat{\mathcal{G}}}$ and P_G studied in Lemma 4.1, and λ stands for the λ-sufficiency parameter of $P_{\mathcal{G}}$.

Proof. The proof will be along the lines of that of Lemma 3.6 in [20]. We first recall [8, Lemma E.15], which we restate in Lemma A.5. In this proof, we consider an environment specific true MDP $\mathcal{M}$ with transition model P, and an MDP $\widehat{\mathcal{M}}$ that has as transition model the estimated causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$. In the following, the expectations will be w.r.t. P. Moreover, since the reward r is fixed, we remove it from the expressions for the sake of clarity. We can start deriving

$$\begin{aligned}
\left|\mathbb{E}_{s\sim P}\left[\hat{V}^{\pi}_{1}(s) - V^{\pi}_{1}(s)\right]\right| &\leq \left|\mathbb{E}_{\pi}\left[\sum_{h=1}^{H}(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}_{h+1}(s_h, a_h)\right]\right| \\
&\leq \mathbb{E}_{\pi}\left[\sum_{h=1}^{H}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}_{h+1}(s_h, a_h)\right|\right] \\
&= \sum_{h=1}^{H}\mathbb{E}_{\pi}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}_{h+1}(s_h, a_h)\right|.
\end{aligned}$$

We now bound a single term within the sum above as follows

$$\begin{aligned}
\mathbb{E}_{\pi}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}_{h+1}(s_h, a_h)\right| &\leq \sum_{s,a}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}(s,a)\right|P^{\pi}(s,a) \\
&= \sum_{s,a}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}(s,a)\right|P^{\pi}(s)\pi(a|s) \\
&\leq \max_{\pi'}\sum_{s,a}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}(s,a)\right|P^{\pi}(s)\pi'(a|s) \\
&= \max_{\nu:\mathcal{S}\to\mathcal{A}}\sum_{s,a}\left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^{\pi}(s,a)\right|P^{\pi}(s)\mathbb{1}\{a = \nu(s)\},
\end{aligned}$$

where in the last step we have used the fact that there must exist an optimal deterministic policy. Due to the uniform sampling assumption, we have $P(s,a) = \frac{1}{SA}$, hence

$$\max_{\pi^\dagger}\frac{P^{\pi^\dagger}(s,a)}{P(s,a)} \leq \frac{1}{P(s,a)} = SA.$$

Therefore, $P^{\pi_\dagger}(s,a) \leq SA \cdot P(s,a)$. Moreover, notice that, since π' is deterministic we have $P^{\pi}(s) = P^{\pi'}(s) = P^{\pi'}(s,a) \leq SA \cdot P(s,a)$. Replacing it in the expression above we get

$$\begin{aligned}
\mathbb{E}_\pi \left| (\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^\pi_{h+1}(s_h, a_h) \right| &\leq SA \cdot \sum_{s,a} \left| (\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^\pi_{h+1}(s,a) \right| P(s)\mathbb{1}\{a = \nu(s)\} \\
&\leq SA \cdot \left| (\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}^\pi_{h+1}(s,a) \right| \\
&\leq SA \cdot \left| (\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^\pi_{h+1}(s,a) \right| + SA \cdot \left| (P_{\mathcal{G}} - P)\widehat{V}^\pi_{h+1}(s,a) \right| \quad (22) \\
&\leq SA \cdot \sum_{i=h+1}^{H} i \cdot \epsilon' + SA \cdot \sum_{i=h+1}^{H} i \cdot \lambda \\
&\leq SAH^2\epsilon' + SAH^2\lambda \quad (23)
\end{aligned}$$

where ϵ' is the approximation error between $\widehat{P}_{\widehat{\mathcal{G}}}$ and P_G studied in Lemma 4.1, and in the second-to-last step we have used the following derivation

$$\begin{aligned}
&\left| (\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^\pi_{h+1}(s,a) \right| \\
&= \left| \widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^\pi_{h+1}(s,a) - P_{\mathcal{G}}\widehat{V}^\pi_{h+1}(s,a) \right| \quad (24) \\
&= \Big| \sum_{s'} \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a)\widehat{V}^\pi_{h+1}(s') - \sum_{s'} P_{\mathcal{G}}(s'|s,a)\widehat{V}^\pi_{h+1}(s') \Big| \\
&= \Big| \sum_{s'} \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \left[r(s',a') + \widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^\pi_{h+2}(s',a') \right] \\
&\quad - \sum_{s'} P_{\mathcal{G}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \left[r(s',a') + P_{\mathcal{G}}\widehat{V}^\pi_{h+2}(s',a') \right] \Big| \\
&= \Big| \sum_{s'} \left(\widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a) - P_{\mathcal{G}}(s'|s,a) \right) \mathbb{E}_{a'\sim\pi} \left[r(s',a') \right] \\
&\quad + \sum_{s'} \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \left[\widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^\pi_{h+2}(s',a') \right] - \sum_{s'} P_{\mathcal{G}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \left[P_{\mathcal{G}}\widehat{V}^\pi_{h+2}(s',a') \right] \Big| \\
&\leq \epsilon' + \Big| \sum_{s'} \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \left[\widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^\pi_{h+2}(s',a') \right] - \sum_{s'} P_{\mathcal{G}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \left[P_{\mathcal{G}}\widehat{V}^\pi_{h+2}(s',a') \right] \Big| \\
&= \epsilon' + \Big| \sum_{s'} \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \Big[\sum_{s''} \widehat{P}_{\widehat{\mathcal{G}}}(s''|s',a')\, \mathbb{E}_{a''\sim\pi} \left[r(s'',a'') + \widehat{P}_{\widehat{\mathcal{G}}}\widehat{V}^\pi_{h+3}(s'',a'') \right] \Big] \\
&\quad - \sum_{s'} P_{\mathcal{G}}(s'|s,a)\, \mathbb{E}_{a'\sim\pi} \Big[\sum_{s''} P_{\mathcal{G}}(s''|s',a')\, \mathbb{E}_{a''\sim\pi} \left[r(s'',a'') + P_{\mathcal{G}}\widehat{V}^\pi_{h+3}(s'',a'') \right] \Big] \Big| \\
&\leq \epsilon' + \sum_{s',s'',a'} \left| \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a)\widehat{P}_{\widehat{\mathcal{G}}}(s''|s',a') - P_{\mathcal{G}}(s'|s,a)P_{\mathcal{G}}(s''|s',a') \right|_1 + \dots \\
&\leq \epsilon' + \sum_{s',s'',a'} \left[\left| \widehat{P}_{\widehat{\mathcal{G}}}(s'|s,a) - P_{\mathcal{G}}(s'|s,a) \right|_1 + \left| \widehat{P}_{\widehat{\mathcal{G}}}(s''|s',a') - P_{\mathcal{G}}(s''|s',a') \right|_1 \right] + \dots \\
&\leq \epsilon' + 2\epsilon' + \dots
\end{aligned}$$

Hence, due to this recursive unrolling, we have

$$\left| (\widehat{P}_{\widehat{\mathcal{G}}} - P_{\mathcal{G}})\widehat{V}^\pi_{h+1}(s,a) \right| \leq \sum_{i=h+1}^{H} i\epsilon' \leq H^2\epsilon.$$

Notice that the same argument holds also for the second term of (22), replacing ϵ' with λ.

By plugging the result in equation (23) into the initial expression we get

$$\begin{aligned}\left|\mathbb{E}_{s\sim P}\left[\widehat{V}_1^\pi(s) - V_1^\pi(s)\right]\right| &\leq \sum_{h=1}^{H} \mathbb{E}_\pi \left|(\widehat{P}_{\widehat{\mathcal{G}}} - P)\widehat{V}_{h+1}^\pi(s_h, a_h)\right| \\ &\leq \sum_{h=1}^{H} SAH^2\epsilon' + SAH^2\lambda \\ &= SAH^3\epsilon' + SAH^3\lambda.\end{aligned}$$

□

In the following we restate [8, Lemma E.15] for the case of a stationary transition model.

Lemma A.5. *For any two MDPs $\mathcal{M}'$ and $\mathcal{M}''$ with rewards r' and r'' and transition models P' and P'', the difference in value functions V', V'' w.r.t. the same policy π can be written as:*

$$V_h'(s) - V_h''(s) = \mathbb{E}_{\mathcal{M}'',\pi}\left[\sum_{i=h}^{H}[r'(s_i,a_i) - r''(s_i,a_i) + (P' - P'')V_{i+1}'(s_i,a_i)] \mid s_h = s\right]. \quad (25)$$

Proofs of Section 4.2

We provide the proof of the sample complexity result for learning the causal structure of a discrete MDP with a generative model.

Theorem 4.6. *Let $\mathcal{M}$ be a discrete MDP with an underlying causal structure $\mathcal{G}$, let $\delta \in (0,1)$, and let $\epsilon > 0$. The Algorithm 3 returns a dependency graph $\widehat{\mathcal{G}}$ such that $Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon) \leq \delta$ with a sample complexity $K = O\big(n\log(d_S^2 d_A/\delta)/\epsilon^2\big)$.*

Proof. We aim to obtain the number of samples K for which Algorithm 3 is guaranteed to return a causal structure estimate $\widehat{\mathcal{G}}$ such that $Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon) \leq \delta$ in a discrete MDP setting. First, we can upper bound the probability of the bad event $Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon)$ in terms of the probability of a failure in the independence testing procedure $\mathbb{I}(X_z, Y_j)$ for a single pair of nodes $X_z \in \mathcal{G}_\epsilon, Y_z \in \mathcal{G}_\epsilon$, i.e.,

$$Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon) \leq Pr\Bigg(\bigcup_{z=1}^{d_S+d_A}\bigcup_{j=1}^{d_S} \text{test } \mathbb{I}(X_z, Y_j) \text{ fails}\Bigg) \leq \sum_{z=1}^{d_S+d_A}\sum_{j=1}^{d_S} Pr\Big(\text{test } \mathbb{I}(X_z, Y_j) \text{ fails}\Big), \quad (26)$$

where we applied an union bound to obtain the last inequality. Now we can look at the probability of a single independence test failure. Especially, for a provably efficient independence test (the existence of such a test is stated by Lemma 4.5, whereas the Algorithm 2 in [13] reports an actual testing procedure), we have $Pr(\text{test } \mathbb{I}(X_z, Y_j) \text{ fails}) \leq \delta'$, for any choice of $\delta' \in (0,1)$, $\epsilon' > 0$, with a number of samples

$$K' = C\left(\frac{n\ \log^{1/3}(1/\delta')}{(\epsilon')^{4/3}} + \frac{n\ \log^{1/2}(1/\delta') + \log(1/\delta')}{(\epsilon')^2}\right), \quad (27)$$

where C is a sufficiently large universal constant [13, Theorem 1.3]. Finally, by letting $\epsilon' = \epsilon$, $\delta' = \frac{\delta}{d_S^2 d_A}$ and combining (26) with (27), we obtain $Pr(\widehat{\mathcal{G}} \neq \mathcal{G}_\epsilon)$ with a sample complexity

$$K = O\left(\frac{n\log(d_S^2 d_A/\delta)}{\epsilon^2}\right),$$

under the assumption $\epsilon^2 \ll \epsilon^{4/3}$, which concludes the proof. □

The proof of the analogous sample complexity result for a tabular MDP setting (Corollary 4.7) is a direct consequence of Theorem 4.6 by letting $n = 2, d_S = S, d_A = A$.

Proofs of Section 4.3

We first report a useful concentration inequality for the L1-norm between the empirical distribution computed over K samples and the true distribution [45, Theorem 2.1].

Lemma A.6 ([45])**.** *Let $X_1, \dots, X_K$ be i.i.d. random variables over $[n]$ having probabilities $Pr(X_k = i) = P_i$, and let $\widehat{P}_K(i) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}(X_k = i)$. Then, for every threshold $\epsilon > 0$, it holds*

$$Pr\Big(\|\widehat{P}_K - P\|_1 \geq \epsilon\Big) \leq 2\exp(-K\epsilon^2/2n).$$

We can now provide the proof of the sample complexity result for learning the Bayesian network of a discrete MDP with a given causal structure.

Theorem 4.8. *Let $\mathcal{M}$ be a discrete MDP, let $\mathcal{G}$ be its underlying causal structure, let $\delta \in (0,1)$, and let $\epsilon > 0$. The Algorithm 2 returns a Bayesian network $\widehat{P}_{\mathcal{G}}$ such that $Pr(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ with a sample complexity $K = O\big(d_S^3 n^{3Z+1}\log(d_S n^Z/\delta)/\epsilon^2\big)$.*

Proof. We aim to obtain the number of samples K for which Algorithm 2 is guaranteed to return a Bayesian network estimate $\widehat{P}_{\mathcal{G}}$ such that $Pr(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ in a discrete MDP setting. First, we note that

$$Pr\Big(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \epsilon\Big) \leq Pr\Big(\sum_{j=1}^{d_S}\|\widehat{P}_j - P_j\|_1 \geq \epsilon\Big) \tag{28}$$

$$\leq Pr\Big(\frac{1}{d_S}\sum_{j=1}^{d_S}\|\widehat{P}_j - P_j\|_1 \geq \frac{\epsilon}{d_S}\Big) \tag{29}$$

$$\leq Pr\Big(\bigcup_{j=1}^{d_S}\|\widehat{P}_j - P_j\|_1 \geq \frac{\epsilon}{d_S}\Big) \tag{30}$$

$$\leq \sum_{j=1}^{d_S} Pr\Big(\|\widehat{P}_j - P_j\|_1 \geq \frac{\epsilon}{d_S}\Big), \tag{31}$$

in which we employed the property $\|\mu - \nu\|_1 \leq \|\prod_i \mu_i - \prod_i \nu_i\|_1 \leq \sum_i \|\mu_i - \nu_i\|_1$ for the L1-norm between product distributions $\mu = \prod_i \mu_i, \nu = \prod_i \nu_i$ to write (28), and we applied a union bound to derive (31) from (30). Similarly, we can write

$$Pr\Big(\|\widehat{P}_j - P_j\|_1 \geq \frac{\epsilon}{d_S}\Big) \leq Pr\Big(\bigcup_{x\in[n]^{|Z_j|}}\|\widehat{P}_j(\cdot|x) - P_j(\cdot|x)\|_1 \geq \frac{\epsilon}{d_S n^{|Z_j|}}\Big) \tag{32}$$

$$\leq \sum_{x\in[n]^{|Z_j|}} Pr\Big(\|\widehat{P}_j(\cdot|x) - P_j(\cdot|x)\|_1 \geq \frac{\epsilon}{d_S n^{|Z_j|}}\Big) \tag{33}$$

$$\leq \sum_{x\in[n]^{|Z_j|}} Pr\Big(\|\widehat{P}_j(\cdot|x) - P_j(\cdot|x)\|_1 \geq \frac{\epsilon}{d_S n^{Z}}\Big) \tag{34}$$

by applying a union bound to derive (33) from (32), and by employing Assumption 1 to bound $|Z_j|$ with Z in (34). We can now invoke Lemma A.6 to obtain the sample complexity K' that guarantees $Pr(\|\widehat{P}_j(\cdot|x) - P_j(\cdot|x)\|_1 \geq \epsilon') \leq \delta'$, i.e.,

$$K' = \frac{2n\log(2/\delta')}{(\epsilon')^2} = \frac{2\, d_S^2\, n^{2Z+1}\, \log(2d_S n^Z/\delta)}{\epsilon^2},$$

where we let $\epsilon' = \frac{\epsilon}{d_S n^Z}$, $\delta' = \frac{\delta}{d_S n^Z}$. Finally, by summing K' for any $x \in [nm]^{|Z_j|}$ and any $j \in [d_S]$, we obtain

$$K = \sum_{j\in[d_S]}\sum_{x\in[n]^{|Z_j|}} K' \leq \frac{2\, d_S^3\, n^{3Z+1}\, \log(2d_S n^Z/\delta)}{\epsilon^2},$$

which proves the theorem. □

To prove the analogous sample complexity result for a tabular MDP we can exploit a slightly tighter concentration on the KL divergence between the empirical distribution and the true distribution in the case of binary variables [12, Theorem 2.2.3][6], which we report for convenience in the following lemma.

Lemma A.7 ([12])**.** *Let $X_1, \dots, X_K$ be i.i.d. random variables over $[2]$ having probabilities $Pr(X_k = i) = P_i$, and let $\widehat{P}_K(i) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}(X_k = i)$. Then, for every threshold $\epsilon > 0$, it holds*

$$Pr\Big(d_{KL}\big(\widehat{P}_K||P\big) \geq \epsilon\Big) \leq 2\exp(-K\epsilon).$$

We can now provide the proof of Corollary 4.9.

Corollary 4.9. *Let $\mathcal{M}$ a tabular MDP. The result of Th. 4.8 reduces to $K = O\big(S^2 2^{2Z}\log(\frac{S2^Z}{\delta})/\epsilon^2\big)$.*

Proof. We aim to obtain the number of samples K for which Algorithm 2 is guaranteed to return a Bayesian network estimate $\widehat{P}_{\mathcal{G}}$ such that $Pr(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ in a tabular MDP setting. We start by considering the KL divergence $d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}})$. Especially, we note

$$\begin{aligned}
d_{KL}\big(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}}\big) &= \sum_{X,Y} \widehat{P}_{\mathcal{G}}(X,Y)\log\frac{\widehat{P}_{\mathcal{G}}(X,Y)}{P_{\mathcal{G}}(X,Y)} \\
&= \sum_{X,Y} \widehat{P}_{\mathcal{G}}(X,Y)\log\frac{\prod_{j=1}^{S}\widehat{P}_j(Y[j]|X[Z_j])}{\prod_{j=1}^{S}P_j(Y[j]|X[Z_j])} \\
&= \sum_{X,Y} \widehat{P}_{\mathcal{G}}(X,Y)\sum_{j=1}^{S}\log\frac{\widehat{P}_j(Y[j]|X[Z_j])}{P_j(Y[j]|X[Z_j])} = \sum_{j=1}^{S} d_{KL}\big(\widehat{P}_j||P_j\big).
\end{aligned}$$

Then, for any $\epsilon' > 0$ we can write

$$\begin{aligned}
Pr\Big(d_{KL}\big(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}}\big) \geq \epsilon'\Big) &\leq Pr\Big(\bigcup_{j=1}^{S} d_{KL}\big(\widehat{P}_j||P_j\big) \geq \frac{\epsilon'}{S}\Big) && (35)\\
&\leq \sum_{j=1}^{S} Pr\Big(d_{KL}\big(\widehat{P}_j||P_j\big) \geq \frac{\epsilon'}{S}\Big) && (36)\\
&\leq \sum_{j=1}^{S} Pr\Big(\bigcup_{x\in[2]^{|Z_j|}} d_{KL}\big(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)\big) \geq \frac{\epsilon'}{S2^{|Z_j|}}\Big) && (37)\\
&\leq \sum_{j=1}^{S}\sum_{x\in[2]^{|Z_j|}} Pr\Big(d_{KL}\big(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)\big) \geq \frac{\epsilon'}{S2^{|Z_j|}}\Big) && (38)\\
&\leq \sum_{j=1}^{S}\sum_{x\in[2]^{|Z_j|}} Pr\Big(d_{KL}\big(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)\big) \geq \frac{\epsilon'}{S2^{Z}}\Big), && (39)
\end{aligned}$$

in which we applied a first union bound to get (36) from (35), a second union bound to get (38) from (37), and Assumption 1 to bound $|Z_j|$ with Z in (39). We can now invoke Lemma A.7 to obtain the sample complexity K'' that guarantees $Pr(d_{KL}(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)) \geq \epsilon'') \leq \delta''$, i.e.,

$$K'' = \frac{\log(2/\delta'')}{\epsilon''} = \frac{S2^Z\log(2S2^Z/\delta')}{\epsilon'},$$

where we let $\epsilon'' = \frac{\epsilon'}{S2^Z}$, and $\delta'' = \frac{\delta'}{S2^Z}$ for any choice of $\delta' \in (0,1)$. By summing K'' for any $x \in [2]^{|Z_j|}$ and and $j \in [S]$, we obtain the sample complexity K' that guarantees $Pr(d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}}) \geq$

[6] Also reported in [28, Example 1].

$\epsilon') \leq \delta'$, i.e.,

$$K' = \sum_{j=1}^{S} \sum_{x \in [2]^{|Z_j|}} K'' \leq \frac{S^2 2^{2Z} \log(2S2^Z/\delta')}{\epsilon'}. \tag{40}$$

Finally, we employ the Pinsker's inequality $\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \leq \sqrt{2d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}})}$ [7] to write

$$Pr\Big(d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}}) \geq \epsilon'\Big) = Pr\Big(\sqrt{2d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}})} \geq \sqrt{2\epsilon'}\Big) \geq Pr\Big(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \sqrt{2\epsilon'}\Big),$$

which gives the sample complexity K that guarantees $Pr(\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \geq \epsilon) \leq \delta$ by letting $\epsilon' = \frac{\epsilon^2}{2}$ and $\delta' = \delta$ in (40), i.e.,

$$K = \frac{2S^2 2^{2Z} \log(2S2^Z/\delta)}{\epsilon^2}.$$

□

B Numerical Validation

Let A, W, P, S, D, C and St represent academic performance, weight, physical activity, sleep, diet and study respectively. We start by defining the causal transition model as follows:

$$\begin{aligned}
A &\sim \mathcal{N}(\mu_A, \sigma) \\
W &\sim \mathcal{N}(\mu_W, \sigma) \\
\mu_A &= A + 0.2D + 0.5S + 0.8St - 0.8 \\
\mu_W &= W - 0.5D - 0.5P + 1 \\
\sigma &= 0.1
\end{aligned}$$

The transition model for a specific environment is then generated by adding independent white noise to each coefficient in the above equations for μ_A and μ_W, including the ones not shown because set to 0:

$$\begin{aligned}
\mu_A &= (1+\epsilon_1)A + \epsilon_2 W + \epsilon_3 P + (0.2+\epsilon_4)D + (0.5+\epsilon_5)S + \epsilon_6 C + (0.8+\epsilon_7)St - 0.8 \\
\mu_W &= \epsilon_8 A + (1+\epsilon_9)W + (-0.5+\epsilon_{10})P + (-0.5+\epsilon_{11})D + \epsilon_{12} S + \epsilon_{13} C + \epsilon_{14} St + 0.8 \\
\epsilon_i &\sim \mathcal{N}(0, 0.1) \quad \forall i \in \{1 \dots 14\}
\end{aligned}$$

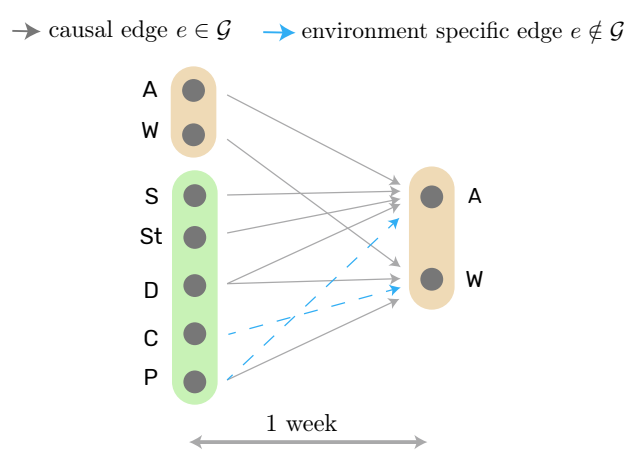


Figure 5: Bayesian network for a specific environment (person) of our synthetic example. Grey edges are causal and therefore shared by other environments in the same universe $\mathbb{U}$, while blue dashed edges are environment-specific dependencies.