Invariance Discovery for Systematic Generalization in Reinforcement Learning

Mirco Mutti [*1 2] **Riccardo De Santi** [*3] **Emanuele Rossi** [4 5] **Juan Felipe Calderon** [1]
Michael Bronstein [4 6] **Marcello Restelli** [1]

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) (Sutton and Barto, 2018) as a powerful tool to address a wide range of sequential decision making problems, the curse of generalization (Kirk et al., 2021) 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. 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, and they are unlikely to be of any use for addressing different tasks. Instead, some universal causal relations generalize over the environments, and once learned 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 with a few, or zero, interactions.

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 document we provide the first tractable formulation of systematic generalization (Section 3), thanks to a set of structural assumptions that are motivated by a causal viewpoint. Especially, we consider a large, potentially infinite, set of reward-free environments, or a *universe*, the agent can freely interact with. 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 document, we provide a partially positive answer to

[*]Equal contribution [1]Politecnico di Milano [2]Università di Bologna [3]ETH Zurich [4]Twitter [5]Imperial College London [6]University of Oxford. Correspondence to: Mirco Mutti <mirco.mutti@polimi.it>, Riccardo De Santi <rdesanti@ethz.ch>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

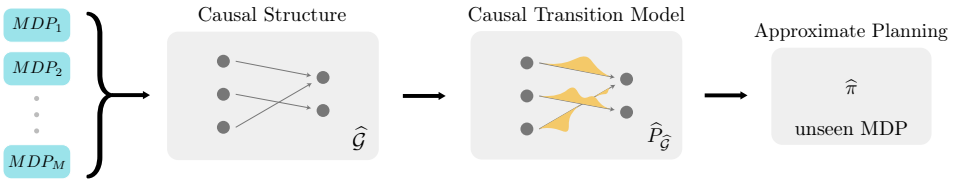

Figure 1: High-level illustration of causal model-based approach to systematic generalization.

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 $\mathcal{G}$ (Wadhwa and Dong, 2021). 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 $\mathcal{G}$ (Dasgupta, 1997) from a mixture of the environments. Then, the learned 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 allows achieving any desired planning error up to an unavoidable error term, which is inherent to the setting. Especially, we provide an analysis of the sample complexity (Section 4) of the proposed approach, which is polynomial in all the relevant quantities of the problem.

Finally, with this work we aim to connect several active research areas on reward-free RL (Jin et al., 2020), multi-task RL (Brunskill and Li, 2013), model-based RL (Sutton and Barto, 2018), factored MDPs (Rosenberg and Mansour, 2021), causal RL (Zhang et al., 2020), experimental design (Ghassami et al., 2018), independence testing (Canonne et al., 2018), into a general framework where individual progresses can be enhanced beyond the sum of their parts.

2. Notation

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 any $A \in \mathcal{A}$, we denote with $A[Z]$ the vector $(A_i)_{i \in Z}$. Given two probability measures P and Q over a discrete space $\mathcal{A}$, their L_1-distance is $\|P - Q\|_1 = \sum_{A \in \mathcal{A}} |P(A) - Q(A)|$. We will denote by $\mathcal{U}_{\mathcal{A}}$ the uniform distribution over $\mathcal{A}$.

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. 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.

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 (Dean and Kanazawa, 1989) 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 We define a *discrete* episodic Markov Decision Process (MDP) (Puterman, 2014) as $\mathcal{M} := ((\mathcal{S}, d_{\mathcal{S}}, n), (\mathcal{A}, d_{\mathcal{A}}, n), P, H, r)$, where $\mathcal{S}$ is a set of $|\mathcal{S}| = S$ states and $\mathcal{A}$ is a set of $|\mathcal{A}| = A$ actions, such that every $s \in \mathcal{S}$ can be represented through a d_S-dimensional vector of discrete features taking value in $[n]$, and $a \in \mathcal{A}$ through a d_A-dimensional vector of discrete features taking value in $[n]$.[2] 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 horizon of an episode, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic reward function. A stochastic *policy* $\pi_h(a|s)$ 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 $V_h^{\pi}(s) := \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right]$. We will write $V_{\mathcal{M},r}^{\pi}$ to denote V_1^{π} in the MDP $\mathcal{M}$ with reward function r.

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, which we call a *universe*.

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

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

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

[2] Note that any *tabular* MDP can be formulated under this alternative formalism by taking $n = 2$, $d_S = S$, and $d_A = A$.

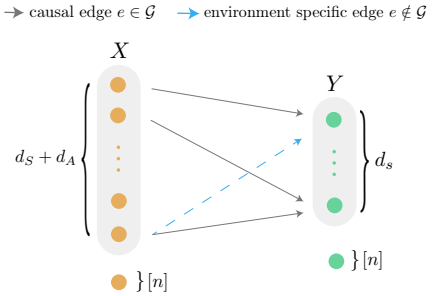

Figure 2: Illustration of the causal structure $\mathcal{G}$ of $\mathbb{U}$.

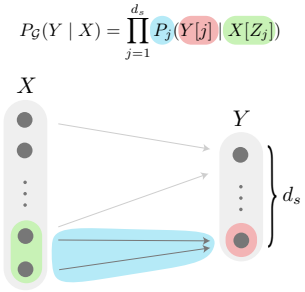

Figure 3: Illustration of the causal transition model $P_{\mathcal{G}}$.

$\mathcal{M} \in \mathbb{U}$ 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*.

Definition 1 (Systematic Generalization)**.** *For any latent MDP $\mathcal{M} \in \mathbb{U}$ and any given reward $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, 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 $\epsilon > 0$.*

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

To ease the notation, we denote the current state-action features with a random vector $X = (X_i)_{i \in [d_S + d_A]}$, and the next state features with a random vector $Y = (Y_i)_{i \in [d_S]}$. For each environment $\mathcal{M}_i \in \mathbb{U}$, the conditional dependencies between Y and X are represented through a bipartite dependency graph $\mathcal{G}_i$. 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. We assume the causal structure $\mathcal{G}$ is time-consistent, i.e., $\mathcal{G}^{(h)} = \mathcal{G}^{(1)}$ for any step $h \in [H]$, and 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)**.** *Let $Z \in \mathbb{N}$. 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,[3] we can express each transition model P_i as

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

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,[4] the Z_j are the set of indices z such that $(X[z], Y[j]) \in \mathcal{G}$. Since it represents the conditional probabilities due to universal causal relations in $\mathbb{U}$, we call $P_{\mathcal{G}}$ the *causal transition model* of $\mathbb{U}$, for which we show an illustration in Figure 3. We assume the causal transition model $P_{\mathcal{G}}$ is also time-consistent, i.e., $P^{(h)}_{\mathcal{G}} = P^{(1)}_{\mathcal{G}}, \forall h \in [H]$, and that it explains a significant part of the transition dynamics of $\mathcal{M}_i \in \mathbb{U}$.

Assumption 2 (λ-sufficiency)**.** *Let $\lambda \in [0, 1]$ be a constant. The causal transition model $P_{\mathcal{G}}$ is* causally λ-sufficient *if*

$$\sup_X \|P_{\mathcal{G}}(\cdot|X) - P_i(\cdot|X)\|_1 \leq \lambda, \qquad \forall P_i \in \mathcal{M}_i \in \mathbb{U}.$$

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 Section 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

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

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

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.

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

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

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 (Hauser and Bühlmann, 2014; Kocaoglu et al., 2017; Ghassami et al., 2018).

4. 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 can draw interactions with a generative model $P(X)$. We would like 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, for which we provide 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 $\mathcal{M}_i$. We provide an upper bound to the sample complexity of the Algorithm 1.

Algorithm 1 Causal Transition Model Estimation

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

Lemma 4.1. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^{M}$ be a class of M discrete MDPs, let $\delta \in (0,1)$, and let $\epsilon > 0$. The Algorithm 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*

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

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]

Theorem 4.2. *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 $\epsilon' = \epsilon/2H^3 n^{Z+1}$.

Theorem 4.2 establish the sample complexity of systematic generalization through Lemma 4.1. For the discrete MDP setting, we have that $\widetilde{O}(MH^6 d_S^3 Z^2 n^{5Z+3})$, 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

[5] 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).

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 $\mathcal{M}$.

References

Brunskill, E. and Li, L. (2013). Sample complexity of multi-task reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–57. IEEE.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*.

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

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150.

Dembo, A. and Zeitouni, O. (2009). Ldp for finite dimensional spaces. In *Large deviations techniques and applications*, pages 11–70. Springer.

Diakonikolas, I., Gouleakis, T., Kane, D. M., Peebles, J., and Price, E. (2021). Optimal testing of discrete distributions with high probability. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing*.

Ghassami, A., Salehkaleybar, S., Kiyavash, N., and Bareinboim, E. (2018). Budgeted experiment design for causal structure learning. In *Proceedings of the International Conference on Machine Learning*.

Hauser, A. and Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020). Reward-free exploration for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

Kirk, R., Zhang, A., Grefenstette, E., and Rocktäschel, T. (2021). A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*.

Kocaoglu, M., Shanmugam, K., and Bareinboim, E. (2017). Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. (2020). 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.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rosenberg, A. and Mansour, Y. (2021). Oracle-efficient regret minimization in factored mdps with unknown structure. In *Advances in Neural Information Processing Systems*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Wadhwa, S. and Dong, R. (2021). On the sample complexity of causal discovery and the value of domain expertise. *arXiv preprint arXiv:2102.03274*.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

Zhang, A., Lyle, C., Sodhani, S., Filos, A., Kwiatkowska, M., Pineau, J., Gal, Y., and Precup, D. (2020). Invariant causal prediction for block mdps. In *Proceedings of the International Conference on Machine Learning*.

A. Proofs

Proofs of Section 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 (Hauser and Bühlmann, 2014). 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: Causal Transition Model Estimation

Before reporting the proof of the main result in Theorem 4.2, it is worth considering a set of lemmas that will be instrumental to the main proof.

First, we state the existence of a principled independence testing procedure.

Lemma A.1 (Diakonikolas et al. (2021))**.** *There exists an (ϵ, δ)-independence tester $\mathbb{I}(A, B)$ for distributions $P_{A,B}$ on $[n] \times [n]$, which returns with probability at least $1 - \delta$*

- *yes, if A, B are independent,*
- *no, if $\inf_{Q \in \{\Delta_A \times \Delta_B\}} \|P_{A,B} - Q\|_1 \geq \epsilon$,*

with a sample complexity $O(n \log(1/\delta)/\epsilon^2)$.

Second, 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.2. *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

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

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

$$\|P_{\mathcal{G}} - P_{\mathcal{G}_\epsilon}\|_1 \leq \sum_{j=1}^{d_S} \left\| P_j(Y[j] \mid X[Z_{j,\epsilon}], \cup_{i=1}^{I} X[i]) - P_j(Y[j] \mid X[Z_{j,\epsilon}]) \right\|_1 \tag{2}$$

$$\leq \sum_{j=1}^{d_S} \sum_{i'=1}^{I} \left\| P_j(Y[j] \mid X[Z_{j,\epsilon}], \cup_{i=i'}^{I} X[i]) - P_j(Y[j] \mid X[Z_{j,\epsilon}], \cup_{i=i'+1}^{I} X[i]) \right\|_1 \tag{3}$$

$$\leq \sum_{j=1}^{d_S} \sum_{i'=1}^{I} \epsilon \leq d_S Z \epsilon, \tag{4}$$

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 A.10 and the Bayesian network estimation result of Theorem A.13.

Lemma A.3. *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

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

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.2 to have $d_s Z \epsilon' \leq \frac{\epsilon}{3}$, which gives $\epsilon' \leq \frac{\epsilon}{3 d_S Z}$. Then, we consider the middle term $(\bullet)$, for which we can write

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

We can now upper bound $(\bullet) \leq \delta/2$ through (6) by invoking Theorem A.10 with threshold $\epsilon' = \frac{\epsilon}{3 d_S Z}$ and confidence $\delta' = \frac{\delta}{2}$, which gives

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

Next, we can upper bound $(\star) \leq \delta/2$ by invoking Theorem A.13 with threshold $\epsilon' = \frac{\epsilon}{3}$ and confidence $\delta' = \frac{\delta}{2}$, which gives

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

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.3 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 $\delta \in (0,1)$, and let $\epsilon > 0$. The Algorithm 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*

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

Lemma A.4. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^M$ be a class of M discrete MDPs, 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*

$$K = O\bigg(\frac{M\ d_S^3\ Z^2\ (nm)^{3Z+1}\ \log\big(\frac{4M d_S^2 d_A (nm)^Z}{\delta}\big)}{\epsilon^2} \bigg).$$

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.3. 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 A.10 with threshold $\epsilon' = \frac{\epsilon}{3d_S Z}$ and confidence $\delta' = \frac{\delta}{2M}$, which gives

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

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.,

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

which concludes the proof. □

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

Lemma A.5. *Let $\mathbb{M} = \{\mathcal{M}_i\}_{i=1}^M$ be a class of M tabular MDPs. The sample complexity of Lemma 4.1 reduces to*

$$K = O\bigg(\frac{M\ S^2\ Z^2\ 4^{2Z}\ \log\big(\frac{4MS^2 A 4^Z}{\delta}\big)}{\epsilon^2} \bigg).$$

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 A.11 (instead of Theorem A.10) with threshold $\epsilon' = \frac{\epsilon}{3SZ}$ and confidence $\delta' = \frac{\delta}{2M}$, which gives

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

Similarly, we upper bound $(\star) \leq \delta/2$ by invoking Corollary A.15 (instead of Theorem A.1) with threshold $\epsilon' = \frac{\epsilon}{3}$ and confidence $\delta' = \frac{\delta}{2}$, which gives

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

Finally, we combine 11 with 12 to obtain

$$K = MK' + K'' \leq \frac{M\ S^2\ Z^2\ 2^{2Z}\ \log\left(\frac{4MS^2A2^Z}{\delta}\right)}{\epsilon^2}$$

□

Proofs of Section 4: Planning

For the sake of notational clarity within the following proofs we express

Theorem 4.2. *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 $\epsilon' = \epsilon/2H^3n^{Z+1}$.

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 (Jin et al., 2020, Theorem 3.5), we can write the following decomposition, where $V^* := V^{\pi^*}$.

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

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.6. 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:

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

□

Lemma A.6. *Under the preconditions of Theorem 4.2, with probability $1-\delta$, for any reward function r and policy π, we can bound the value function estimation error as follows.*

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

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 (Jin et al., 2020). We first recall (Dann et al., 2017, Lemma E.15), which we restate in Lemma A.9. In this prof 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

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

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

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

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

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

We can now bound the first term of (16):

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

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:

$$P^{\pi^\dagger}_{\mathcal{G}}(X[Z_j])\leq n^Z\cdot P_{\mathcal{G}}(X[Z_j])$$

Replacing this in (17) and marginalizing over $Y\setminus Y[j]$ we obtain:

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

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

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

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:

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

Therefore the initial expression in (14) becomes:

$$\begin{aligned}
\left|\mathbb{E}_{s\sim P}\left[\widehat{V}^\pi_1(s)-V^\pi_1(s)\right]\right| &\leq \sum_{h=1}^{H}\mathbb{E}_X\Big|(\widehat{P}_{\widehat{\mathcal{G}}}-P)\widehat{V}^\pi_{h+1}(X)\Big| \qquad (18)\\
&\leq \sum_{h=1}^{H}[n^{Z+1}H^2\epsilon'+n^{2Z+1}d_S H^2\lambda] \qquad (19)\\
&\leq \underbrace{n^{Z+1}H^3\epsilon'}_{\epsilon}+\underbrace{n^{2Z+1}d_S H^3\lambda}_{\epsilon_\lambda} \qquad (20)
\end{aligned}$$

□

Corollary A.7. *For a tabular MDP $\mathcal{M} \in \mathbb{M}$, the result of Theorem 4.2 holds with $\epsilon_\lambda = 2\lambda SAH^3$, $\epsilon' = \epsilon/2SAH^3$.*

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 (Jin et al., 2020, Theorem 3.5), we can write the following decomposition, where $V^* := V^{\pi^*}$.

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

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.8. In order to get $2SAH^3\epsilon' = \epsilon$ we have to set $\epsilon' = \frac{\epsilon}{2SAH^3}$. Considering the sample complexity result in Lemma A.5 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.8. *Under the preconditions of Corollary A.7, with probability $1-\delta$, for any reward function r and policy π, we can bound the value function estimation error as follows.*

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

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 (Jin et al., 2020). We first recall (Dann et al., 2017, Lemma E.15), which we restate in Lemma A.9. In this prof 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

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

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

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

where ϵ' is the approximation error between $\widehat{P}_{\widehat{\mathcal{G}}}$ and P_G studied in Lemma 4.1, and in the penultimate 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| \\
&= \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] - \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] \\
&+ \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] \\
&- \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'}\Big[\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\Big] + \dots \\
&\leq \epsilon' + 2\epsilon' + \dots
\end{aligned}
\tag{24}
$$

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:

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

□

In the following we restate (Dann et al., 2017, Lemma E.15) for the case of stationary transition model.

Lemma A.9. *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:*

$$
V'_h(s) - V''_h(s) = \mathbb{E}_{\mathcal{M}'',\pi}\Big[\sum_{i=h}^{H}[r'(s_i,a_i) - r''(s_i,a_i) + (P' - P'')V'_{i+1}(s_i,a_i)] \mid s_h = s\Big] \tag{25}
$$

Algorithm 2 Causal Structure Estimation for an MDP

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 a dependency is found add (X_z, Y_j) to $\widehat{\mathcal{G}}$
end for
Output: causal dependency graph $\widehat{\mathcal{G}}$

Proofs of Section 4: Causal Discovery

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

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 $P_{A,B}$*

$$(A, B) \in \mathcal{G}_\epsilon \quad \textit{iff} \quad \inf_{Q \in \{\Delta_A \times \Delta_B\}} \|P_{A,B} - Q\|_1 \geq \epsilon.$$

Theorem A.10. *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 2 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 2 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.,

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

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 A.1, whereas the Algorithm 2 in Diakonikolas et al. (2021) 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

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

where C is a sufficiently large universal constant (Diakonikolas et al., 2021, 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

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

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 is a direct consequence of Theorem A.10 by letting $n = 2, d_S = S, d_A = A$.

Corollary A.11. *Let $\mathcal{M}$ be a tabular MDP. The result of Theorem A.10 reduces to $K = O\big(\log(S^2 A/\delta)/\epsilon^2\big)$.*

Algorithm 3 Bayesian Network Estimation for an MDP

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 $\widehat{P}_j(Y[j]|X[Z_j]) = N(X[Z_j], Y[j])/K'$
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 $\widehat{P}_{\mathcal{G}}$

Proofs of Section 4: Bayesian Network Estimation

We first report a useful concentration inequality for the L1-norm between the empirical distribution computed over K samples and the true distribution (Weissman et al., 2003, Theorem 2.1).

Lemma A.12 (Weissman et al. (2003))**.** *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 A.13. *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 3 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 3 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

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

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

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

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

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.12 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 (Dembo and Zeitouni, 2009, Theorem 2.2.3)[6], which we report for convenience in the following lemma.

Lemma A.14 (Dembo and Zeitouni (2009))**.** *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 A.15.

Corollary A.15. *Let $\mathcal{M}$ be a tabular MDP. The result of Theorem A.13 reduces to $K = O\big(S^2 2^{2Z}\log(S2^Z/\delta)/\epsilon^2\big)$.*

Proof. We aim to obtain the number of samples K for which Algorithm 3 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

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

[6] Also reported in (Mardia et al., 2020, Example 1).

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

$$Pr\Big(d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}}) \geq \epsilon'\Big) \leq Pr\Big(\bigcup_{j=1}^{S} d_{KL}(\widehat{P}_j||P_j) \geq \frac{\epsilon'}{S}\Big) \tag{35}$$

$$\leq \sum_{j=1}^{S} Pr\Big(d_{KL}(\widehat{P}_j||P_j) \geq \frac{\epsilon'}{S}\Big) \tag{36}$$

$$\leq \sum_{j=1}^{S} Pr\Big(\bigcup_{x\in[2]^{|Z_j|}} d_{KL}\big(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)\big) \geq \frac{\epsilon'}{S2^{|Z_j|}}\Big) \tag{37}$$

$$\leq \sum_{j=1}^{S} \sum_{x\in[2]^{|Z_j|}} Pr\Big(d_{KL}\big(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)\big) \geq \frac{\epsilon'}{S2^{|Z_j|}}\Big) \tag{38}$$

$$\leq \sum_{j=1}^{S} \sum_{x\in[2]^{|Z_j|}} Pr\Big(d_{KL}\big(\widehat{P}_j(\cdot|x)||P_j(\cdot|x)\big) \geq \frac{\epsilon'}{S2^{Z}}\Big), \tag{39}$$

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.14 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 $Pr(d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}}) \geq \epsilon') \leq \delta'$, 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 $\|\widehat{P}_{\mathcal{G}} - P_{\mathcal{G}}\|_1 \leq \sqrt{2d_{KL}(\widehat{P}_{\mathcal{G}}||P_{\mathcal{G}})}$ (Csiszár, 1967) 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.,

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

□