The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Provably Efficient Causal Model-Based Reinforcement Learning for Systematic Generalization

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

[1]Politecnico di Milano
[2]Università di Bologna
[3]ETH Zurich
[4]Imperial College London
[5]Twitter
[6]University of Oxford
mirco.mutti@polimi.it, rdesanti@ethz.ch

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 that is shared. Consider a set of environments that share the laws of motion as an 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 a tractable formulation of systematic generalization by employing a *causal* viewpoint. Then, under specific structural assumptions, we provide a simple learning algorithm that guarantees 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 (Jaksch, Ortner, and Auer 2010; Lange, Gabel, and Riedmiller 2012). 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.

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 (Jin et al. 2020a), 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

[*]These authors contributed equally. The appendix of this paper can be found at https://arxiv.org/abs/2202.06545.

	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	$Pr(\lVert \widehat{P}_\mathcal{G} - P_\mathcal{G} \rVert_1 \geq \epsilon) \leq \delta$	$\widetilde{O}\big(d_S^3 n^{3Z+1}/\epsilon^2\big)$	$\widetilde{O}\big(S^2 2^{2Z}/\epsilon^2\big)$
Systematic Generalization	$Pr(V_1^* - V_1^\pi \geq \epsilon_\lambda + \epsilon) \leq \delta$	$\widetilde{O}(MH^6 d_S^3 Z^2 n^{5Z+3}/\epsilon^2)$	$\widetilde{O}(MH^6 S^4 A^2 Z^2/\epsilon^2)$

Table 1: Sample complexity overview.

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 $\mathcal{G}$. 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 $\mathcal{G}$ from a mixture of the environments. Finally, the causal transition model is employed by a planning oracle to provide an approximately optimal policy for an unknown 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, 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:

(c1) The first tractable formulation of the systematic generalization problem in RL, thanks to structural assumptions motivated by causal considerations (§ 3);
(c2) A provably efficient algorithm to learn systematic generalization over an infinite set of environments (§ 4.1);
(c3) The sample complexity of estimating the causal structure underlying a discrete MDP (§ 4.2);
(c4) The sample complexity of estimating the Bayesian network underlying a discrete MDP (§ 4.3);
(c5) A brief numerical validation of the main results (§ 5).

On a technical level, (c3, c4) require the adaptation of known results in causal discovery (Wadhwa and Dong 2021) and Bayesian network estimation (Dasgupta 1997) to the specific MDP setting, which are then employed as building blocks to obtain the rate for systematic generalization (c2). See Table 1 for a summary of the main sample complexity results.

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

2 Preliminaries

We start with some notions about graphs, causality, and Markov decision processes for later use. We 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 $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_a$ 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 $\bigotimes$ is 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 $\lVert P - Q \rVert_1 = \sum_{A \in \mathcal{A}} |P(A) - Q(A)|$, and their Kullback-Leibler (KL) divergence is $d_{KL}(P||Q) = \sum_{A \in \mathcal{A}} P(A) \log(P(A)/Q(A))$.

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 (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 A *tabular* episodic Markov Decision Process (MDP, Puterman 2014) 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

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

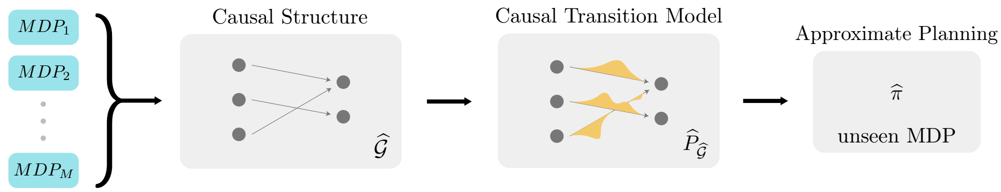


Figure 1: High-level illustration of the causal model-based approach to systematic generalization.

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.,

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

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 $d_A = A$.

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 that 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 $\mathbb{U}$. 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 unknown 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 $V_{\mathcal{M},r}^* - V_{\mathcal{M},r}^\pi \leq \epsilon$ 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 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 $\mathcal{G}$, without loosing 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.[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

[2]Note that one can always take $P_{\mathcal{G}}(Y|Z) = 1, \forall (X, Y)$.

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

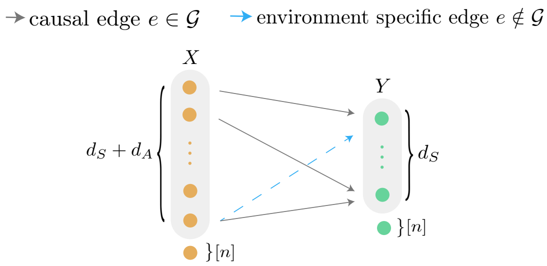


Figure 2: Causal structure $\mathcal{G}$ of $\mathbb{U}$.

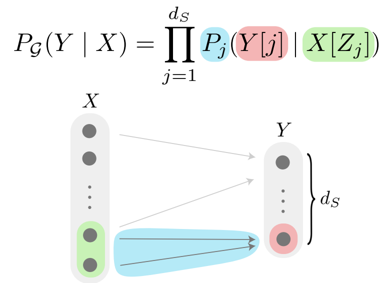


Figure 3: Causal transition model $P_{\mathcal{G}}$ of $\mathbb{U}$.

structure $\mathcal{G}$, $P_{\mathcal{G}}$ 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 the causal parents of $Y[j]$, i.e., $(X[z], Y[j]) \in \mathcal{G}, \forall z \in Z_j$. 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., $P_{\mathcal{G}}^{(h)} = P_{\mathcal{G}}^{(1)}$ for any step $h \in [H]$. In this work, we assume that the causal transition model is non-vacuous and that it explains a significant part of the transition dynamics of $\mathcal{M}_i \in \mathbb{U}$.

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

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). 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 structure 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 $\mathbb{M} \subset \mathbb{U}$ a class of size M. We say that $\mathbb{M}$ is* causally even *if*[5]

$$\mathbb{E}_{i \sim \mathcal{U}_{[M]}} \big[F_i(Y[j]|X)\big] = 1, \ \ \forall j \in [d_S].$$

In this paper we assume that $\mathbb{M}$ is *diverse* and *even* by design, while we leave as future work the problem of selecting such a class from active interactions with $\mathbb{U}$, which would add to our formulation flavors of active learning and experimental design (Hauser and Bühlmann 2014; Kocaoglu, Shanmugam, and Bareinboim 2017; Ghassami et al. 2018).

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 (Hauser and Bühlmann 2014).

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

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] 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 $\mathcal{M}_i \in \mathbb{U} \setminus \mathbb{M}$.

[5] We denote by $\mathcal{U}_{[M]}$ the uniform distribution over $[M]$.

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 an (unknown) 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 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 $\mathcal{M}_i$.[6]

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 2) to obtain $\widehat{\mathcal{G}}$, and an estimation procedure of the Bayesian network over $\widehat{\mathcal{G}}$ (Algorithm 3) 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 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*

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

An analogous result can be derived for 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*

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

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.

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 causal structure $\mathcal{G}_i$, which we assume to be $2Z$-sparse (Assumption 1).

[6]The planning oracle can be substituted with a principled approximate planning solver (see Jin et al. 2020a, Section 3.3).

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 $i = 1, \dots, M$ **do**
  let $P_i(Y|X)$ the transition model of $\mathcal{M}_i \in \mathbb{M}$
  $\widehat{\mathcal{G}}_i \leftarrow$ *Causal Structure Estimation* $(P_i, P(X), K')$
end for
let $\widehat{\mathcal{G}} = \cap_{i=1}^M \widehat{\mathcal{G}}_i$
let $K'' = C''\big(d_S^3 n^{3Z+1} \log(4 d_S n^Z/\delta)/\epsilon^2\big)$
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* $(P_{\mathbb{M}}, \widehat{\mathcal{G}}, K'')$
Output: causal transition model $\widehat{P}_{\widehat{\mathcal{G}}}$

Theorem 4.3. *Let $\delta \in (0,1)$ and $\epsilon > 0$. For an unknown 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 $Pr\big(V^*_{\mathcal{M}_i,r} - V^{\widehat{\pi}}_{\mathcal{M}_i,r} \geq \epsilon_\lambda + \epsilon\big) \leq \delta$, 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 $\epsilon' = \epsilon/2H^3 n^{Z+1}$.*

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 Theorem 4.3 holds with $\epsilon_\lambda = 2\lambda SAH^3$, $\epsilon' = \epsilon/2SAH^3$.*

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 $\widetilde{O}(MH^6 d_S^3 Z^2 n^{5Z+3})$ samples 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., Wadhwa and Dong 2021), 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

Algorithm 2 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 $\widehat{\mathcal{G}}$

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

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

Lemma 4.5 (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* 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 $O(n \log(1/\delta)/\epsilon^2)$.*

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 $\mathcal{M}$ a discrete MDP with 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 $K = O\big(n \log(d_S^2 d_A/\delta)/\epsilon^2\big)$.*

Corollary 4.7. *Let $\mathcal{M}$ a tabular MDP. The result of Theorem 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., Dasgupta 1997) 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 (Dasgupta 1997). 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 $\mathcal{M}$ a discrete MDP with causal structure $\mathcal{G}$, 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 $K = O\big(d_S^3 n^{3Z+1} \log(d_S n^Z/\delta)/\epsilon^2\big)$.*

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

Algorithm 3 MDP Bayesian Network Estimation

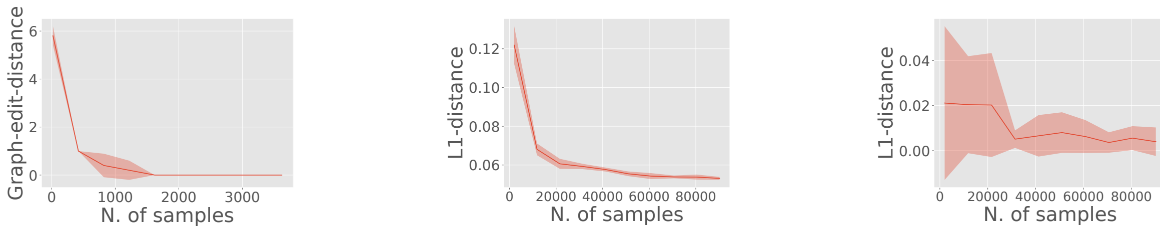
Input: sampling model $P(Y|X)$, dependency graph $\mathcal{G}$, batch parameter K
let $K' = \lceil K/d_S n^Z \rceil$
for $j = 1, \ldots, d_S$ **do**
  let Z_j the scopes $(X[Z_j], Y[j]) \subseteq \mathcal{G}$
  initialize the counts $N(X[Z_j], Y[j]) = 0$
  for each value $x \in [n]^{|Z_j|}$ **do**
    for $k = 1, \ldots, K'$ **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 $\widehat{P}_\mathcal{G}$

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 values 0 (below average), 1 (average) or 2 (above average). See Appendix B for 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 $\text{GED}(\mathcal{G}, \widehat{\mathcal{G}})$ 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, the distance converges to zero as we increase the number of samples, and we can recover the exact causal graph (Figure 4a).

Causal Transition Model Estimation Figure 4b shows the L_1-distance between the estimated and ground-truth



(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.

causal transition model, as a function of the number of samples ($K' + K''$ in Algorithm 1). As the samples grow, the L_1-distance shrinks towards 0.05, which is due to the environments not fully respecting the evenness assumption.

Value Function Estimation Finally, we investigate whether we can approximate the optimal value function for an 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

Finally, we revise the relevant literature and discuss how it relates with our problem formulation and results.

Causal Discovery and Bayesian Networks On a technical level, our work is related to previous efforts on the sample complexity of causal discovery (Wadhwa and Dong 2021) and Bayesian network estimation (Friedman and Yakhini 1996; Dasgupta 1997; Bhattacharyya, Canonne, and Yang 2022). None of these works consider the MDP setting. Instead, we account for the peculiar MDP structure to get sharper rates w.r.t. a blind application of previous results.

Reward-Free RL Reward-free RL (Jin et al. 2020a) is akin to a special case of our systematic generalization framework in which the set of MDPs is a singleton (Wang et al. 2020; Zanette et al. 2020; Kaufmann et al. 2021; Ménard et al. 2021; Zhang, Du, and Ji 2021; Qiu et al. 2021). It is worth comparing our sample complexity result to independent reward-free exploration for each MDP. Let $|\mathbb{U}| = U$, the latter 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 $\mathbb{U}$ (Jin et al. 2020a). This compares favorably with our rate $\widetilde{O}(MH^6S^4A^2/\epsilon^2)$ whenever U is small, but leveraging the inner structure of $\mathbb{U}$ becomes crucial as U grows to infinity, while M remains constant. Our approach pays this further generality with the additional error term ϵ_λ, which is unavoidable. It is an interesting direction to see whether additional factors in S, A, H are also unavoidable.

Hidden Structures in RL Previous works have considered learning an hidden structure of the MDP for sample efficient RL (Du et al. 2019; Misra et al. 2020a,b; Agarwal et al. 2020). 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., Jin et al. 2020b; Cai et al. 2020; Yin et al. 2022) 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 Model-based RL (Sutton and Barto 2018) prescribes learning an approximate model of the transition dynamics to extract an optimal policy. Theoretical works (e.g., Jaksch, Ortner, and Auer 2010; Ayoub et al. 2020) generally focus on the estimation of the approximate value functions obtained through the learned model, rather than the estimation of the model itself. A notable exception is (Tarbouriech et al. 2020), 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 MDP, instead of the shared transition dynamics of an infinite set of MDPs that we target in this paper.

Factored MDPs The factored MDP formalism (Kearns and Koller 1999) allows encoding transition dynamics that are the product of multiple independent factors. This is closely related to how we define the causal transition model 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 (Delgado, Sanner, and De Barros 2011; Xu and Tewari 2020; Talebi, Jonsson, and Maillard 2021; Tian, Qian, and Sra 2020), or by estimating its structure from data (Strehl, Diuk, and Littman 2007; Vigorito and Barto 2009; Chakraborty and Stone 2011; Osband and Van Roy 2014; Rosenberg and Mansour 2021). 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 (Zhang et al. 2020; Tomar et al. 2021; Gasse et al. 2021; Feng et al. 2022) address model-based RL from a causal perspective. The motivations behind (Zhang et al. 2020) are especially similar to ours, but 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. Similarly to our paper, Feng et al. (2022) employ causal structure learning to build a factored representation of the MDP, but they tackle non-stationary changes in the environment instead of systematic generalization. Finally, Lu, Meisami, and Tewari (2021) show how to exploit a known causal representation for sample efficient RL, which can complement our work on how to learn such representation.

References

Agarwal, A.; Kakade, S.; Krishnamurthy, A.; and Sun, W. 2020. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*.

Ayoub, A.; Jia, Z.; Szepesvari, C.; Wang, M.; and Yang, L. 2020. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*.

Bhattacharyya, A.; Canonne, C. L.; and Yang, J. Q. 2022. Independence testing for bounded degree Bayesian network. In *Advances in Neural Information Processing Systems*.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2020. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*.

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

Chakraborty, D.; and Stone, P. 2011. Structure learning in ergodic factored MDPs without knowledge of the transition function's in-degree. In *International Conference on Machine Learning*.

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.

Delgado, K. V.; Sanner, S.; and De Barros, L. N. 2011. Efficient solutions to factored MDPs with imprecise transition probabilities. *Artificial Intelligence*, 175(9-10): 1498–1527.

Dembo, A.; and Zeitouni, O. 2009. LDP for finite dimensional spaces. In *Large Deviations Techniques and Applications*, 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 *Annual ACM SIGACT Symposium on Theory of Computing*.

Du, S.; Krishnamurthy, A.; Jiang, N.; Agarwal, A.; Dudik, M.; and Langford, J. 2019. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*.

Feng, F.; Huang, B.; Zhang, K.; and Magliacane, S. 2022. Factored adaptation for non-stationary reinforcement learning. In *Advances in Neural Information Processing Systems*.

Friedman, N.; and Yakhini, Z. 1996. On the sample complexity of learning Bayesian networks. In *International Conference on Uncertainty in Artificial Intelligence*.

Gasse, M.; Grasset, D.; Gaudron, G.; and Oudeyer, P.-Y. 2021. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421*.

Ghassami, A.; Salehkaleybar, S.; Kiyavash, N.; and Bareinboim, E. 2018. Budgeted experiment design for causal structure learning. In *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.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4).

Jin, C.; Krishnamurthy, A.; Simchowitz, M.; and Yu, T. 2020a. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020b. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

Kaufmann, E.; Ménard, P.; Domingues, O. D.; Jonsson, A.; Leurent, E.; and Valko, M. 2021. Adaptive reward-free exploration. In *Algorithmic Learning Theory*.

Kearns, M.; and Koller, D. 1999. Efficient reinforcement learning in factored MDPs. In *International Joint Conference on Artificial Intelligence*.

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*.

Lange, S.; Gabel, T.; and Riedmiller, M. 2012. Batch reinforcement learning. In *Reinforcement Learning*, 45–73.

Lu, Y.; Meisami, A.; and Tewari, A. 2021. Efficient reinforcement learning with prior causal knowledge. In *Conference on Causal Learning and Reasoning*.

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.

Ménard, P.; Domingues, O. D.; Jonsson, A.; Kaufmann, E.; Leurent, E.; and Valko, M. 2021. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*.

Misra, D.; Henaff, M.; Krishnamurthy, A.; and Langford, J. 2020a. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International Conference on Machine Learning*.

Misra, D.; Liu, Q.; Jin, C.; and Langford, J. 2020b. Provable rich observation reinforcement learning with combinatorial latent states. In *International Conference on Learning Representations*.

Osband, I.; and Van Roy, B. 2014. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*.

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

Qiu, S.; Ye, J.; Wang, Z.; and Yang, Z. 2021. On reward-free RL with kernel and neural function approximations: Single-agent MDP and Markov game. In *International Conference on Machine Learning*.

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

Strehl, A. L.; Diuk, C.; and Littman, M. L. 2007. Efficient structure learning in factored-state MDPs. In *AAAI Conference on Artificial Intelligence*.

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

Talebi, M. S.; Jonsson, A.; and Maillard, O. 2021. Improved exploration in factored average-reward mdps. In *International Conference on Artificial Intelligence and Statistics*.

Tarbouriech, J.; Shekhar, S.; Pirotta, M.; Ghavamzadeh, M.; and Lazaric, A. 2020. Active model estimation in markov decision processes. In *International Conference on Uncertainty in Artificial Intelligence*.

Tian, Y.; Qian, J.; and Sra, S. 2020. Towards minimax optimal reinforcement learning in factored markov decision processes. In *Advances in Neural Information Processing Systems*.

Tomar, M.; Zhang, A.; Calandra, R.; Taylor, M. E.; and Pineau, J. 2021. Model-invariant state abstractions for model-based reinforcement learning. *arXiv preprint arXiv:2102.09850*.

Vigorito, C. M.; and Barto, A. G. 2009. Incremental structure learning in factored MDPs with continuous states and actions. *University of Massachusetts Amherst-Department of Computer Science, Technical Report*.

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

Wang, R.; Du, S. S.; Yang, L.; and Salakhutdinov, R. R. 2020. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems*.

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, Technical Report*.

Xu, Z.; and Tewari, A. 2020. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. In *Advances in Neural Information Processing Systems*.

Yin, D.; Hao, B.; Abbasi-Yadkori, Y.; Lazić, N.; and Szepesvári, C. 2022. Efficient local planning with linear function approximation. In *Algorithmic Learning Theory*.

Zanette, A.; Lazaric, A.; Kochenderfer, M. J.; and Brunskill, E. 2020. Provably efficient reward-agnostic navigation with linear value iteration. In *Advances in Neural Information Processing Systems*.

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 *International Conference on Machine Learning*.

Zhang, Z.; Du, S.; and Ji, X. 2021. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*.