Planning under Distribution Shifts with Causal POMDPs

Matteo Ceriscioli, Karthika Mohan
School of Electrical Engineering and Computer Science (EECS)
Oregon State University
Corvallis, OR 97331, USA
{ceriscim,karthika.mohan}@oregonstate.edu

Abstract

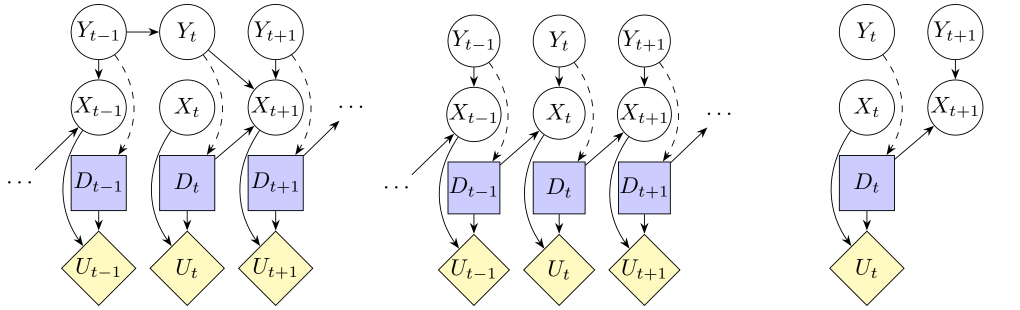
In the real world, planning is often challenged by distribution shifts. As such, a model of the environment obtained under one set of conditions may no longer remain valid as the distribution of states or the environment dynamics change, which in turn causes previously learned strategies to fail. In this work, we propose a theoretical framework for planning under partial observability using Partially Observable Markov Decision Processes (POMDPs) formulated using causal knowledge. By representing shifts in the environment as interventions on this causal POMDP, the framework enables evaluating plans under hypothesized changes and actively identifying which components of the environment have been altered. We show how to maintain and update a belief over both the latent state and the underlying domain, and we prove that the value function remains piecewise linear and convex (PWLC) in this augmented belief space. Preservation of PWLC under distribution shifts has the advantage of maintaining the tractability of planning via α-vector-based POMDP methods.

Introduction

Planning aims to compute an optimal way for an agent to act, assuming access to a complete and correct model of the environment. In stochastic settings, this often means identifying a policy that maps information states to actions and maximizes long-term reward Kaelbling et al. [1998]. Difficulties arise once the agent is deployed in a different environment and the underlying state distribution or transition dynamics change. Distribution shifts undermine the assumption that the transition and observation processes remain fixed and accurately describe how states evolve.

To illustrate this issue, consider a delivery rover whose policy and world model were obtained in an environment with specific conditions, such as urban areas with temperate weather. When the same rover is deployed in a different setting, relevant factors, such as friction, visibility, or mechanical stress, may change. A conventional POMDP planner cannot identify which mechanisms have shifted or why performance has degraded. A causally informed model can represent such changes as interventions on specific components, reason about their effects on the environment and the sensors, and support effective planning under the resulting distribution shift.

This paper examines whether planning remains tractable in the presence of distribution shifts. The tractability of finite-horizon POMDPs relies on the fact that their value functions are PWLC in the belief state Smallwood and Sondik [1973]. When the environment is affected by an unknown shift, the resulting ambiguity in the transition dynamics introduces an additional source of uncertainty into the model. Although certain structured forms of uncertainty are known to preserve the PWLC structure [Osogami, 2015], general forms of ambiguity can cause finite-horizon POMDPs to lose their PWLC properties [Saghafian, 2018]. We consider a causal formulation of POMDPs in which



(a) Non-time-homogeneous Causal POMDP

(b) Time-homogeneous Causal POMDP

(c) Compact representation

Figure 1: On the left is an example of a non–time-homogeneous causal POMDP. If the causal model is faithful, that is, every edge in the causal graph corresponds to an actual direct dependency between two variables, then any change in the graph structure across timesteps implies a change in the transition function. In the center is a CID that is compatible with a time-homogeneous causal POMDP. On the right is a compact representation of a time-homogeneous causal POMDP compatible with the CID shown in the center.

distribution shifts are represented as interventions. Under this formulation, we show that the value function retains its PWLC structure despite the presence of unknown shifts. Thus this ensures that standard α-vector–based POMDP planning algorithms remain applicable.

Causal modeling is widely used to support robust decision making under perturbations of the data-generating process [Pearl and Bareinboim, 2011, Pearl, 2018, Schölkopf, 2022, Ceriscioli and Mohan, 2025], as it offers a principled framework for analyzing how distribution shifts affect outcomes. In particular, such shifts can be naturally represented as interventions in a causal model [Richens and Everitt, 2024, Ceriscioli and Mohan, 2025], which allows systematic reasoning about their consequences.

The primary contribution of this paper is a causal framework for planning under partial observability in the presence of distribution shifts, which preserves the PWLC structure of the value function. The framework enables the evaluation of policies under specified shifts and supports the detection and identification of changes in the environment.

All proofs are included in Appendix A.

1 Preliminaries

Let X be a node in a graph G. In this paper we denote the set of parent nodes of X as $Pa(X)$. Random variables are represented with upper-case letters, e.g. X, and the corresponding instantiations with lower-case letters, e.g. x. The set of observable values of a random variable X is called $dom(X)$. Given a set $\mathbf{X}$, $\Pi(\mathbf{X})$ is the set of all probability distributions over $\mathbf{X}$.

Factored POMDPs. Partially Observable Markov Decision Processes (POMDPs) Åström [1965], Kaelbling et al. [1998] model sequential decision-making in which the agent lacks full observability of the environment. When each state in a POMDP is represented as a vector $s = (v_1, \ldots, v_n)$ of instantiations of random variables $V_1, \ldots, V_n$, the model is called a factored POMDP Boutilier and Poole [1996]. This structure allows the transition function to be decomposed into sub-functions involving only subsets of the state variables.

Causal Influence Diagrams. In this paper, we study factored POMDPs whose intra- and inter-temporal causal structure is expressed using Causal Influence Diagrams (CIDs) Heckerman [1995], Everitt et al. [2021]. CIDs extend Causal Bayesian Networks (CBNs) Pearl [2009] by incorporating decision-theoretic elements, offering a structured way to represent causal relationships while specifying what the agent observes and what it influences through its actions.

Definition 1 (Causal Influence Diagram [Everitt et al., 2021])**.** A *Causal Influence Diagram* (CID) is a Causal Bayesian Network $M = (G = (\mathbf{V}, \mathbf{E}), P)$, where P is a joint probability distribution compatible with the conditional independences encoded in the DAG G. The nodes in $\mathbf{V}$ are partitioned into decision ($\mathbf{D}$), utility ($\mathbf{U}$), and chance ($\mathbf{C}$) nodes, $\mathbf{V} = (\mathbf{D}, \mathbf{U}, \mathbf{C})$. Each utility node U_i is assigned a utility function $f_i : dom(Pa(U_i)) \to \mathbb{R}$.

In a CID, the environment is described by the set of chance nodes $\mathbf{C}$. Each chance node $C \in \mathbf{C}$ corresponds to a random variable, The set of decision nodes $\mathbf{D}$ contains all the variables which value is set by an agent, for each $D \in \mathbf{D}$ it is possible to assign a policy $\pi : dom(Pa(D)) \to \mathcal{A}$.

2 Causal POMDPs

There exist various causal representations of POMDPs, including models that integrate causal structure into online planning or representation learning. For example, CAR-DESPOT incorporates causal information into the AR-DESPOT planner to reason about confounding in robotic environments Cannizzaro and Kunze [2023]. Earlier work has proposed a causal POMDP model for causal representation learning and has shown its usefulness for zero-shot learning in complex tasks [Sontakke et al., 2021]. Other approaches use causal modeling to recover hidden causal dynamics within partially observable environments [Liang and Boularias, 2021], or employ causal abstractions to improve long-horizon reasoning [Gao et al., 2025].

In this work, we adapt the formulation of Ceriscioli and Mohan [2025], which models a POMDP as an infinite CID unrolled over time containing a decision node and a utility node at each timestep t, corresponding respectively to the action and reward at time t.

Definition 2 (Causal POMDP). A causal POMDP is a tuple $(\mathbf{V}, \mathcal{A}, \mathbf{T}, R, \mathbf{V}_o, O, \gamma)$ where:

1. $\mathbf{V} = (V_1, \ldots, V_n)$ is the ordered set of state variables.
2. $\mathcal{A}$ is the finite set of actions.
3. $T : dom(\mathbf{V}) \times \mathcal{A} \to \Pi(dom(\mathbf{V}))$ is the state-transition function. Given $\mathbf{V}^{(t)}$ the set of state variables before the transition, and $\mathbf{V}^{(t+1)}$ the set of state variables after the transition, it is possible to decompose T as follows:

$$T(v^{(t)}, a^{(t)}, v^{(t+1)}) = \prod_{i=1}^{n} T_i(v_i^{(t+1)} \mid a^{(t)}, pa(v_i^{(t+1)})) \tag{1}$$

   with $Pa(V_i^{(t+1)}) \subseteq \mathbf{V}^{(t)} \cup (\bigcup_{k<i} V_k^{(t+1)})$. Each function T_i governs the transition of the state variable V_i.

4. $R : \mathbf{V}_R \times \mathcal{A} \to \mathbb{R}$ is a reward function, where $\mathbf{V}_R \subseteq \mathbf{V}$.
5. $\mathbf{V}_o = \{V_{o,1}, \ldots, V_{o,p}\}$ is a set of observable variables, forming the agent's observation.
6. $\mathbf{O} = \{O_1, \ldots, O_p\}$ $O : dom(\mathbf{V}) \times \mathcal{A} \times dom(\mathbf{V}_o) \to [0, 1]$ is the set of conditional observation probabilities.
7. $\gamma \in [0, 1)$ is the discount factor.

The set of state variables induces the set of states $\mathbf{S} \coloneqq dom(\mathbf{V}) = dom(V_1) \times \cdots \times dom(V_n)$. Similarly, the set of observations is $\mathbf{\Omega} \coloneqq dom(\mathbf{V}_o)$. The agent receives a reward $R(s, a)$ after taking an action $a \in \mathcal{A}$ in state $s \in \mathbf{S}$.

Observable variables can be included in the state ($\mathbf{V}_o \subseteq \mathbf{V}$) without loss of expressiveness: for any causal POMDP where $\mathbf{V}_o \cap \mathbf{V} = \emptyset$, there exists an equivalent causal POMDP with $\mathbf{V}_o' \subseteq \mathbf{V}$ where transitions and rewards do not depend on $\mathbf{V}_o$.

Assumption 1. The observable variables are also state variables, i.e. $\mathbf{V}_o \subseteq \mathbf{V}$.

When Assumption 1 holds, the state fully determines the observation. Therefore, the conditional observation probability function simplifies to $O(v', a, v_o') = 1$ if v_o' is compatible with v' and 0 otherwise, are the new state and observation after action a. Since v' determines v_o', we can write $O(v', v_o')$ instead of $O(v', a, v_o')$.

Under Assumption 1, a causal POMDP $(\mathbf{V}, \mathcal{A}, \mathbf{T}, R, \mathbf{V}_o, O, \gamma)$ induces a CID with an infinite DAG $G = (\mathbf{V}', E)$, where $\mathbf{V} = \bigcup_{t=0}^{\infty} \mathbf{V}^{(t)}$ and $\mathbf{V}^{(t)} := \mathbf{V}$ (see Figure 1b). Time-homogeneity of the transition function allows a compact CID representation including only two consecutive timesteps t and $t+1$, with decisions and utilities at t, and edges involving D and U, between timesteps, and within $t+1$ (see Figure 1c). If all state variables are observable, this reduces to a causal MDP.

Distribution shifts generally alter the state transition function, and if the distribution shift is unknown, then estimating it requires planning using a non-time-homogeneous POMDP, non-time-homogeneous causal POMDPs can also be represented with infinite CIDs, as illustrated in Figure 1a.

3 Planning under Distribution Shifts

Being able to discern the causal relationships governing the environment allows us to infer the state dynamics under changing conditions. In a causal model, *interventions* are deliberate alterations of components or mechanisms of the model. A *distribution shift* is any change in the probability distribution of a random variable. In this paper, we focus on shifts that affect the variables describing the environment in which we plan, specifically, the variables used to factorize the POMDP state in the causal POMDP model. A *domain* denotes the probabilistic configuration of the environment under a particular shift, including the original one.

In this section, we develop the use of causal POMDPs for planning under potential distribution shifts. We first formalize distribution shifts as interventions within the underlying causal model. A key consequence of modeling distribution shifts as interventions is that they can be embedded directly into the agent's belief and planning process. We proceed by describing how to evaluate a policy under a specified distribution shift. Finally, we address the problem of planning while concurrently identifying the distribution shift, showing how to update the agent's belief over states and domains and that the resulting belief value function remains piecewise linear and convex.

3.1 Modeling Distribution Shifts as Interventions

As shown in previous work Richens and Everitt [2024], Ceriscioli and Mohan [2025], interventions are an effective way to represent distribution shifts in a causal model. Applying an intervention σ to a set of variables may change the joint distribution $P(\mathbf{V})$. When the model and intervention are known, it is possible to compute the updated distribution $P(\mathbf{V}; \sigma)$ which is called the *interventional distribution*, opposed to the *observational distribution* $P(\mathbf{V})$, which is the one we observe when no intervention is applied.

We introduce stochastic shifts, a family of soft interventions.

Definition 3 (Stochastic Shift Intervention)**.** Let $M = (G, \Theta)$ be a CBN, and X be a discrete random variable with $dom(X) = \{x_1, \dots, x_m\}$. A *Stochastic Shift Intervention* σ is an intervention associated with a matrix:

$$A_\sigma = \begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mm} \end{pmatrix} \quad \text{with} \sum_j p_{ji} = 1 \text{ for every } i. \tag{2}$$

such that when applied to a random variable X, its conditional distribution is updated as follows:

$$P(X = x_i \mid pa(X); \sigma) = \sum_{j=1}^{m} p_{ji} P(X = x_j \mid pa(X)) \tag{3}$$

The condition $\sum_j p_{ji} = 1$ is equivalent to requiring that each row in A_σ must sum up to one. Observe that if the probability distribution of X is represented with a probability vector, e.g. $P(X|pa(X)) = (P(x_1 \mid pa(X)), \dots, P(x_m \mid pa(X)))^T$, then we can apply the stochastic shift intervention σ by matrix multiplication, i.e.

$$P(X \mid pa(X); \sigma) = A_\sigma^T P(X|pa(X)) \tag{4}$$

A_σ is the identity matrix iff σ is the identity intervention σ_{id}, which leaves the domain and distribution unchanged.

Example. Let $X \sim Unif(\{1,2,3\})$ and σ be a stochastic shift s.t. $p_{11} = p_{22} = 1$, and $p_{31} = p_{32} = \frac{1}{2}$. Each time X takes the value 3, σ remaps it to 1 or 2 with equal probability. So $P(X = x; \sigma) = \frac{1}{2}$ if $x \in \{1,2\}$ and 0 if $x = 3$.

A stochastic shift intervention σ maps probability distributions over a finite set to other distributions. The following result shows that, for any starting distribution, there exists a σ that maps it to any target distribution.

Proposition 1. *Given a conditional distribution $P(X \mid pa(X))$ and an arbitrary target conditional distribution $P'(X \mid pa(X))$, it is possible to define a stochastic shift intervention σ s.t. $P'(X \mid pa(X)) = P(X \mid pa(X); \sigma)$.*

Note that even if the original distribution $P(X \mid pa(X))$ and the shifted distribution $P(X \mid pa(X); \sigma)$ are both observed, in the general case it is not always possible to uniquely determine the distribution shift σ, as it is possible that two different shifts applied to the same distribution generate the same distribution. Suppose X is the outcome of a fair coin flip, i.e., $X \sim Bern(\frac{1}{2})$, also suppose that after altering the coin we observe $Bern(\frac{3}{4})$ as the new distribution, then both

$$A_\sigma = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix} \qquad A_{\sigma'} = \begin{pmatrix} 0.5 & 0.5 \\ 1 & 0 \end{pmatrix} \tag{5}$$

correspond to shifts that map $Bern(\frac{1}{2})$ to $Bern(\frac{3}{4})$.

3.2 Evaluating Policies under Distribution Shifts

Let $M = (\mathbf{V}, \mathcal{A}, \mathbf{T}, R, \mathbf{V}_o, O, \gamma)$ be a causal POMDP, $\pi : dom(\mathbf{V}_o) \to \mathcal{A}$ a policy, and σ a stochastic shift intervention representing a distribution shift. When planning under partial observability, the state is generally not completely available to the agent and therefore it maintains a belief b_S about the state that updates whenever it receives an observation Kaelbling et al. [1998]. Evaluating a policy consists in computing its expected return, represented by the belief value function $V^\pi(b)$ using the Bellman equation for causal POMDPs:

$$V^\pi(b_S) = \sum_{s \in S} R(s, \pi(b_S)) b_S(s) + \gamma \sum_{s', o'} O(s', o') \sum_{s} \prod_{V_i \in \mathbf{V}^{(t+1)}} T_i(v_i \mid \pi(b_S), pa(v_i)) V^\pi(b_S^{a,o}) \tag{6}$$

Known shift σ. We write $V^\pi(b_S; \sigma)$ to denote the value function when the environment is affected by the shift σ. Its expression is identical to that in Equation 6, except that each transition term $T_i(v_i \mid \pi(b_S), pa(v_i))$ is replaced by $T_i(v_i \mid \pi(b_S), pa(v_i); \sigma)$. Note that if $\sigma = \sigma_{id}$ then $V^\pi(b_S; \sigma) = V^\pi(b_S)$. Once the shift σ is fixed and the transition functions are updated via Equation 3, the problem of computing the value function of a causal POMDP reduces to that of a standard POMDP.

3.3 Planning under Unknown Distribution Shifts

Now we consider the task of planning in an environment affected by an unknown distribution shift.

3.3.1 State and Domain Estimation.

As part of planning with causal POMDP under an unknown distribution shift, it is no longer sufficient for the agent to keep track of its belief about the state, as it also needs to keep a belief about the unknown domain. It is possible to define a prior on the state and the domain separately as $b_S(s)$ and $b_\Sigma(\sigma)$, however, since in the general case at each timestep the observation depends both on the previous state and the domain, the state belief and the domain belief become coupled and therefore it is appropriate to keep track of them using a joint belief $b(s, \sigma)$. If we have no prior knowledge we can express that as uniform priors on both the states and the domain $b(s, \sigma) = b(\sigma|s)b(s) = b_\Sigma(\sigma)b_S(s)$ where $b_S(s) \sim$ Unif(S), and $b_\Sigma(\sigma) \sim$ Unif(Σ).

The following proposition shows that the joint belief over states and domains admits a Bayesian update analogous to the standard POMDP filter, but extended to account for domain-dependent transition dynamics.

Proposition 2. *[State-Domain Joint Belief Update] Let b be the current joint belief over states and domains, a be the action taken at time t, and o' the observation received after performing a. Then the updated state state-domain joint belief b' is:*

$$b'_{o',a}(s',\sigma) = \frac{O(s',o')}{P(o' \mid a,b)} \sum_{s} b(s,\sigma) \prod_{V_i \in \mathbf{V^{(t+1)}}} T(v_i \,|\, pa(V_i); \sigma) \tag{7}$$

such that $s' = (v_1, \ldots, v_n)$ for $\{V_1, \ldots, V_n\} = \mathbf{V}^{(t+1)}$.

3.4 Preservation of PWLC under Distribution Shift

We establish that, despite the presence of an unknown shift, the structural properties that support tractable POMDP planning remain intact.

Value Function for Casual POMDPs under an Unknown Shift. To demonstrate that the value function of a causal POMDP subject to a distribution shift and planning horizon n is piecewise linear and convex, and that it admits a finite vector representation analogous to that of standard POMDPs, we propose an approach similar to Porta et al. [2005] and present a constructive proof based on the state–action value function.

The following lemma shows that the value function under an unknown distribution shift can still be expressed as a maximum over linear functionals of the joint belief.

Lemma 1. *Let Σ be a set of stochastic shift interventions, the value function of a causal POMDP with set of states $\mathbf{S}$ subject to an unknown distribution shift in the set Σ with planning horizon n can be expressed as:*

$$V_n(b) = \max_{\{\alpha_n^i\}_i} \sum_{s \in \mathbf{S}} \int_{\Sigma} \alpha_n^i(s,\sigma) b(s,\sigma)\, d\sigma \tag{8}$$

for some functions $\alpha_n^i : S \times \Sigma \to \mathbb{R}$.

Using the characterization of $V_n(b)$ provided by Lemma 1, the following theorem completes our analysis by proving that the value function under distribution shifts is still piecewise linear and convex, thereby preserving the structural form that underlies α-vector–based planning.

Theorem 1. *Let Σ be a set of stochastic shift interventions, the value function of a causal POMDP subject to an unknown distribution shift in the set Σ with planning horizon n is piecewise linear and convex in the joint state-domain belief b.*

This confirms that distribution shifts do not compromise the representational tractability of the planning problem.

Conclusions

Recognizing distribution shifts as a critical source of uncertainty in planning, this work introduces a framework for decision-making under such shifts using causal POMDPs. By representing the state in a factorized form and distribution shifts as interventions on the underlying causal model, causal POMDPs enable both the evaluation of policies under specified shifts and planning in an unknown domain by maintaining a joint belief over states and domains. We show that, even under the uncertainty in the transition dynamics that arises from an unknown distribution shift, the finite-horizon value function remains piecewise linear and convex with respect to the joint state–domain belief, thereby preserving the structural properties that support α-vector–based planning in standard POMDPs.

References

Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. URL `https://api.semanticscholar.org/CorpusID:121222106`.

Craig Boutilier and David Poole. Computing optimal policies for partially observable decision processes using compact representations. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1168–1175. AAAI Press, 1996. ISBN 026251091X.

Ricardo Cannizzaro and Lars Kunze. Car-despot: Causally-informed online pomdp planning for robots in confounded environments. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2018–2025, 2023. doi: 10.1109/IROS55552.2023.10342223.

Matteo Ceriscioli and Karthika Mohan. Agents robust to distribution shifts learn causal world models even under mediation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=JvHif4fyeP`.

Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. Agent incentives: A causal perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13): 11487–11495, May 2021. doi: 10.1609/aaai.v35i13.17368. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17368`.

Haichuan Gao, Tianrun Xu, Tianren Zhang, Yuqing Guo, Chujie Zhao, Jinsheng Ren, Yizhou Jiang, Shangqi Guo, and Feng Chen. Causal dreamer for partially observable model-based reinforcement learning. *Neurocomputing*, 652:131012, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.131012. URL `https://www.sciencedirect.com/science/article/pii/S0925231225016844`.

David Heckerman. A bayesian approach to learning causal networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 285–295, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL `https://www.sciencedirect.com/science/article/pii/S000437029800023X`.

Junchi Liang and Abdeslam Boularias. Inferring time-delayed causal relations in pomdps from the principle of independence of cause and mechanism. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1944–1950. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/268. URL `https://doi.org/10.24963/ijcai.2021/268`. Main Track.

Takayuki Osogami. Robust partially observable markov decision process. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 106–115, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/osogami15.html`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 3, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3176182. URL `https://doi-org.oregonstate.idm.oclc.org/10.1145/3159652.3176182`.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):247–254, Aug. 2011. doi: 10.1609/aaai.v25i1.7861. URL `https://ojs.aaai.org/index.php/AAAI/article/view/7861`.

Josep M. Porta, Matthijs T. J. Spaan, and Nikos Vlassis. Robot planning in partially observable continuous domains. In *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2005. doi: 10.15607/RSS.2005.I.029.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=pOoKI3ouv1`.

Soroush Saghafian. Ambiguous partially observable markov decision processes: Structural results and applications. *Journal of Economic Theory*, 178:1–35, 2018. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2018.08.006. URL `https://www.sciencedirect.com/science/article/pii/S0022053118304770`.

Bernhard Schölkopf. *Causality for Machine Learning*, page 765–804. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, October 1973. doi: 10.1287/opre.21.5.1071. URL `https://ideas.repec.org/a/inm/oropre/v21y1973i5p1071-1088.html`.

Sumedh A Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal curiosity: Rl agents discovering self-supervised experiments for causal representation learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9848–9858. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/sontakke21a.html`.

A Proof

Proposition 1. *Given a conditional distribution $P(X \mid pa(X))$ and an arbitrary target conditional distribution $P'(X \mid pa(X))$, it is possible to define a stochastic shift intervention σ s.t. $P'(X \mid pa(X)) = P(X \mid pa(X); \sigma)$.*

Proof. Let $dom(X) = \{x_1, \ldots, x_m\}$. We define a stochastic shift interventions σ with parameters $p_{ji} = P'(X = x_i \mid pa(X))$. Observe that for all $i \in \{1 \ldots, m\}$:

$$\begin{aligned} P(X = x_i \mid pa(X); \sigma) &= \sum_j p_{ji} P(X = x_j \mid pa(X)) \\ &= P'(X = x_i \mid pa(X)) \sum_j P(X = x_j \mid pa(X)) \\ &= P'(X = x_i \mid pa(X)) \end{aligned} \tag{9}$$

Hence, the intervention σ transforms the original conditional distribution $P(X \mid pa(X))$ exactly into the target distribution $P'(X \mid pa(X))$, which concludes the proof. □

Proposition 2 (State-Domain Joint Belief Update)**.** *Let b be the current joint belief over states and domains, a be the action taken at time t, and o' the observation received after performing a. Then the updated state state-domain joint belief b' is:*

$$b'_{o',a}(s', \sigma) = \frac{O(s', o')}{P(o' \mid a, b)} \sum_s b(s, \sigma) \prod_{V_i \in \mathbf{V^{(t+1)}}} T(v_i | pa(V_i); \sigma) \tag{10}$$

such that $s' = (v_1, \ldots, v_n)$ for $\{V_1, \ldots, V_n\} = \mathbf{V}^{(t+1)}$.

Proof. By definition:

$$b'_{o',a}(s', \sigma) = P(s', \sigma \mid o', a, b) \tag{11}$$

where b is the current belief, a is the action taken at the current timestep, and o' is the observation received after executing a. By Bayes rule:

$$= \frac{P(o' \mid s', a, \sigma, b)}{P(o' \mid a, b)} P(s', \sigma \mid a, b) \tag{12}$$

Note that $o' \perp\!\!\!\perp \{a, \sigma, b\} \mid s'$, so $P(o' \mid s', a, \sigma, b) = P(o' \mid s') = O(s', o')$.

$$= \frac{O(s', o')}{P(o' \mid a, b)} \sum_{s \in \mathbf{S}} P(s' \mid s, a, \sigma, b) P(s, \sigma \mid a, b) \tag{13}$$

As $s' \perp\!\!\!\perp b \mid \{s, a, \sigma\}$ and $\{s, \sigma\} \perp\!\!\!\perp a \mid b$

$$= \frac{O(s', o')}{P(o' \mid a, b)} \sum_{s \in \mathbf{S}} P(s' \mid s, a, \sigma) P(s, \sigma \mid b) \tag{14}$$

$$= \frac{O(s', o')}{P(o' \mid a, b)} \sum_{s \in \mathbf{S}} b(s, \sigma) \prod_{V_i \in \mathbf{V^{(t+1)}}} T(v_i \mid pa(V_i); \sigma) \tag{15}$$

which corresponds to the desired expression. □

Lemma 1. *Let Σ be a set of stochastic shift interventions, the value function of a causal POMDP with set of states $\mathbf{S}$ subject to an unknown distribution shift in the set Σ with planning horizon n can be expressed as:*

$$V_n(b) = \max_{\{\alpha_n^i\}_i} \sum_{s \in \mathbf{S}} \int_\Sigma \alpha_n^i(s, \sigma) b(s, \sigma) \, d\sigma \tag{16}$$

for some functions $\alpha_n^i : S \times \Sigma \to \mathbb{R}$.

Proof. We prove the statement by induction on the planning horizon n.
Base case. Let $n = 0$, since the plan ends after the execution of a single action, then the value function corresponds to:

$$V_0(b) = \max_{a \in \mathcal{A}} Q_0(b, a) = \max_{a \in \mathcal{A}} \sum_{s \in \mathbf{S}} \int_{\Sigma} R(s, a) b(s, \sigma)\, d\sigma \tag{17}$$

We define $\{\alpha_0^i(s, \sigma)\} = \{R(s, a)\}_{a \in \mathcal{A}}$ and hence the value function is in the desired form.
Induction step. Assume the statement holds for planning horizon $n - 1$. Let $b'_{o',a}$ be the belief obtained by updating the belief b after observing o' and executing action a according Equation 7. Observe that the value function for a planning horizon n can be defined recursively as:

$$V_n(b) = \max_{a \in \mathcal{A}} \{\sum_{s \in \mathbf{S}} \int_{\Sigma} b(s, \sigma) R(s, a)\, d\sigma + \gamma \sum_{o' \in \boldsymbol{\Omega}} P(o' \mid a, b) V_{n-1}(b'_{o',a})\} \tag{18}$$

by inductive hypothesis:

$$V_{n-1}(b'_{o',a}) = \max_{\{\alpha_{n-1}^j\}} \sum_{s' \in \mathbf{S}} \int_{\Sigma} \alpha_{n-1}^j(s', \sigma) b'_{o',a}(s', \sigma)\, d\sigma \tag{19}$$

by substituting the expression from Equation 19 into Equation 18, we obtain:

$$\begin{aligned} V_n(b) = \max_{a} \{ &\sum_{s \in \mathbf{S}} \int_{\Sigma} R(s, a) b(s, \sigma)\, d\sigma + \\ &+ \gamma \sum_{o' \in \boldsymbol{\Omega}} P(o' \mid a, b) \max_{\{\alpha_{n-1}^j\}} \sum_{s' \in \mathbf{S}} \int_{\Sigma} \alpha_{n-1}^j(s', \sigma) b'_{o',a}(s', \sigma)\, d\sigma\} \end{aligned} \tag{20}$$

by updating the belief according to Proposition 2 we get:

$$\begin{aligned} = \max_{a \in \mathcal{A}} \{ &\sum_{s \in \mathbf{S}} \int_{\Sigma} R(s, a) b(s, \sigma)\, d\sigma + \\ &+ \gamma \sum_{o' \in \boldsymbol{\Omega}} \max_{\{\alpha_{n-1}^j\}} \sum_{s \in \mathbf{S}} \int_{\Sigma} \left[\sum_{s' \in \mathbf{S}} O(s', o') P(s' \mid s, a, \sigma) \alpha_{n-1}^j(s', \sigma) \right] b(s, \sigma)\, d\sigma\} \end{aligned} \tag{21}$$

Let $\alpha_{a,o'}^j(s, \sigma) \coloneqq \sum_{s' \in \mathbf{S}} O(s', o') P(s' \mid s, a, \sigma) \alpha_{n-1}^j(s', \sigma)$, then Equation 21 can be rewritten as:

$$= \max_{a} \{\sum_{s \in \mathbf{S}} \int_{\Sigma} R(s, a) b(s, \sigma)\, d\sigma + \gamma \sum_{o' \in \boldsymbol{\Omega}} \max_{\{\alpha_{n-1}^j\}} \sum_{s \in \mathbf{S}} \int_{\Sigma} \alpha_{a,o'}^j(s, \sigma) b(s, \sigma)\, d\sigma\} \tag{22}$$

Let $\alpha_{a,o',b} \coloneqq \arg\max_{\{\alpha_{a,o'}^j\}} \sum_{s \in \mathbf{S}} \int_{\Sigma} \alpha_{a,o'}^j(s, \sigma) b(s, \sigma)\, d\sigma$, then:

$$V_n(b) = \max_{a \in \mathcal{A}} \left\{ \sum_{s \in \mathbf{S}} \int_{\Sigma} \left[R(s, a) + \gamma \sum_{o' \in \boldsymbol{\Omega}} \alpha_{a,o',b} \right] b(s, \sigma)\, d\sigma \right\} \tag{23}$$

Finally, we define:

$$\{\alpha_n^i\}_i \coloneqq \{R(s, a) + \gamma \sum_{o' \in \boldsymbol{\Omega}} \alpha_{a,o',b}\}_{a \in \mathcal{A}} \tag{24}$$

Then:

$$V_n(b) = \max_{\{\alpha_n^i\}_i} \sum_{s \in \mathbf{S}} \int_{\Sigma} \alpha_n^i(s, \sigma) b(s, \sigma)\, d\sigma \tag{25}$$

proving the induction step. □

Theorem 1. *Let Σ be a set of stochastic shift interventions, the value function of a causal POMDP subject to an unknown distribution shift in the set Σ with planning horizon n is piecewise linear and convex in the joint state-domain belief b:*

Proof. By Lemma 1 for every planning horizon n we know that:

$$V_n(b) = \max_{\{\alpha_n^i\}_i} \sum_{s \in \mathbf{S}} \int_{\Sigma} \alpha_i^n(s, \sigma) b(s, \sigma)\, d\sigma \tag{26}$$

Let $V_n^i(b) \coloneqq \sum_s \int_\Sigma \alpha_i^n(s, \sigma) b(s, \sigma)\, d\sigma$ and b_1, b_2 be two joint state-domain beliefs. Let $\lambda_1, \lambda_2 \in \mathbb{R}$, then:

$$\begin{aligned} V_n^i(\lambda_1 b_1 + \lambda_2 b_2) &= \sum_{s \in \mathbf{S}} \int_\Sigma \alpha_i^n(s, \sigma)(\lambda_1 b_1(s, \sigma) + \lambda_2 b_2(s, \sigma))\, d\sigma \\ &= \lambda_1 \sum_{s \in \mathbf{S}} \int_\Sigma \alpha_i^n(s, \sigma) b_1(s, \sigma)\, d\sigma + \lambda_2 \sum_{s \in \mathbf{S}} \int_\Sigma \alpha_i^n(s, \sigma) b_2(s, \sigma)\, d\sigma \\ &= \lambda_1 V_n^i(b_1) + \lambda_2 V_n^i(b_2) \end{aligned} \tag{27}$$

Therefore $V_n^i(b)$ is linear in b. Since, for any belief b, the value function $V_n(b)$ is given by the $V_n^i(b)$ with the largest value, and because there is a finite number of linear functions $V_n^i(b)$, it follows that $V_n(b)$ is piecewise linear. Since linear functions are convex and $V_n(b)$ is obtained as the pointwise maximum of linear functions, $V_n(b)$ is convex as well. □