Summary of *Robust Agents Learn Causal World Models*

J. Richens and T. Everitt, ICLR 2024

Basic setup

Let $\boldsymbol{V} = \{V_1, \ldots, V_n\}$ be a finite collection of random variables. A Bayesian network on $\boldsymbol{V}$ is a pair $M = (G, P)$, where G is a DAG on the V_i and the joint law factorizes according to G:

$$P(v_1, \ldots, v_n) = \prod_i P(v_i \mid pa_i),$$

where pa_i denotes the values of the parents $\mathrm{Pa}(V_i)$ of V_i.

In the paper the V_i come in three types: environmental (chance) variables $\boldsymbol{C} = \{C_1, \ldots, C_m\}$, a decision node D, and a utility node U. This models a decision problem: the agent observes the parents $\mathrm{Pa}(D)$ of the decision node and chooses the value of D according to some policy; each outcome carries a utility given by the value of U.

In simplified terms, the paper asks: **under what circumstances does the agent's family of policies contain enough information to reconstruct the underlying causal model** (G, P)**?**

Interventions

The paper's answer is essentially that the agent must be sufficiently *robust* to certain perturbations of its environment.

The basic perturbation is a *local intervention*. For instance, we may replace the local conditional law at node C_i with a point mass:

$$P_{do(C_i = c_i')}(c_1, \ldots, c_n) = \mathbf{1}_{\{c_i = c_i'\}} \prod_{j \neq i} P(c_j \mid pa_j).$$

Note that $do(\cdot)$ is *not* conditioning under P; it specifies a new interventional law. In general a local intervention has the form $\sigma = do(C_i = f(C_i))$, under which

$$P(c_i \mid pa_i; \sigma) = \sum_{c_i' :\, f(c_i') = c_i} P(c_i' \mid pa_i),$$

with a constant f giving a *hard* intervention. We also allow *finite* mixtures: finitely many local interventions $\sigma_1, \ldots, \sigma_m$ with weights $q_j \geq 0$, $\sum_j q_j = 1$, meaning σ_j is applied with probability q_j, so that $P(\cdot \mid \sigma) = \sum_j q_j P(\cdot \mid \sigma_j)$. The class Σ of domain shifts is the set of all such finite mixtures. The weights are real numbers that may be varied *continuously*, and the theorems assume the agent is near-optimal for *every* $\sigma \in \Sigma$—a continuum of shifts. Only two-component mixtures are needed below, but the continuum in the weight is essential.

A policy is a conditional law $\pi(d \mid pa_D)$ over decisions given the observed variables, and the utility is a real-valued function $U = U(pa_U)$. An optimal policy satisfies $\pi^* \in \arg\max_\pi \mathbb{E}^\pi[U]$.

The paper assumes an *unmediated decision task*, $\mathrm{Desc}(D) \cap \mathrm{Anc}(U) = \varnothing$: the decision influences utility only directly (typically $D \in \mathrm{Pa}(U)$), never through intermediate environmental variables.

Domain shifts and regret

The paper studies generalization beyond the i.i.d. setting. Rather than being tested under the training law P, the environment may shift to an intervention-induced law P_σ, while the utility function stays fixed. These are the *domain shifts*.

Under a shift σ, let π^*_σ be the optimal policy and π_σ the agent's policy. Its regret is

$$\mathrm{Regret}_\sigma(\pi_\sigma) = \mathbb{E}^{\pi^*_\sigma}_\sigma[U] - \mathbb{E}^{\pi_\sigma}_\sigma[U],$$

and the policy is δ-optimal if $\mathrm{Regret}_\sigma(\pi_\sigma) \le \delta$. Regret is an external measure: the agent need not know the optimum.

A second assumption is *domain dependence*: there must exist compatible environment distributions under which the optimal policy differs. Otherwise the policy is the same everywhere and cannot reveal anything about the environment. The central object is therefore the family $\{\pi_\sigma\}_{\sigma\in\Sigma}$, where Σ is the class of mixtures of local interventions.

Main theorems

Theorem 1 (exact reconstruction). For almost all causal influence diagrams satisfying the assumptions above, knowing the optimal policy $\pi^*_\sigma(d \mid pa_D)$ for every $\sigma \in \Sigma$ determines the causal graph and the local conditional laws on the ancestors of the utility:

$$\boxed{\{\pi^*_\sigma\}_{\sigma\in\Sigma} \Longrightarrow (G, P) \text{ on } \mathrm{Anc}(U).}$$

The exceptional cases are degenerate parameter choices for which some causal relation never affects expected utility, and is therefore invisible to optimal decisions.

Theorem 2 (approximate reconstruction). If $\mathrm{Regret}_\sigma(\pi_\sigma) \le \delta$ for all $\sigma \in \Sigma$, then one can reconstruct an approximate model $M' = (G', P')$ with

$$|P'(v_i \mid pa_i) - P(v_i \mid pa_i)| \le \gamma(\delta), \qquad \gamma(0) = 0, \quad \gamma(\delta) = O(\delta)$$

for small δ. Weak causal edges may be missed when their effect on expected utility falls below the regret scale.

Theorem 3 (converse). If (G, P) is known, one can compute an optimal policy under every allowed intervention; and if an approximate model has local probability error at most ε, the induced policy has regret $\delta = O(\varepsilon)$. Schematically,

$$\boxed{\text{accurate causal model} \Longleftrightarrow \text{low-regret adaptation across domain shifts.}}$$

Reconstruction from policy behaviour

The key idea is to turn the points at which the agent's decision *switches* into numerical causal information. Fix two local interventions σ_1, σ_2 and two candidate decisions d_1, d_2, and consider the one-parameter family of mixtures

$$\sigma(q) = q\,\sigma_1 + (1-q)\,\sigma_2, \qquad q \in [0,1],$$

so $q = 0$ is pure σ_2 and $q = 1$ is pure σ_1. Define the *utility gap* of d_1 over d_2 under this mixture,

$$\Delta(q) \;:=\; \mathbb{E}[U \mid d_1, \sigma(q)] - \mathbb{E}[U \mid d_2, \sigma(q)].$$

An optimal agent chooses d_1 exactly when $\Delta(q) > 0$ and d_2 when $\Delta(q) < 0$. Because a mixture law is a convex combination, $P(\cdot \mid \sigma(q)) = qP(\cdot \mid \sigma_1) + (1-q)P(\cdot \mid \sigma_2)$, expected utility is affine in q, hence so is the gap:

$$\Delta(q) = q\,\Delta_1 + (1-q)\,\Delta_2, \qquad \Delta_i := \Delta \text{ under the pure intervention } \sigma_i.$$

Choose σ_1, σ_2 so that they favour different decisions, i.e. Δ_1 and Δ_2 have opposite signs. Then Δ is a straight line crossing zero at exactly one point $q_* \in (0,1)$:

$$\Delta(q_*) = 0 \quad \Longleftrightarrow \quad q_*\Delta_1 + (1-q_*)\Delta_2 = 0 \quad \Longleftrightarrow \quad \Delta_1 = -\frac{1-q_*}{q_*}\,\Delta_2.$$

Observationally, q_* is the mixing weight at which the optimal policy flips from d_2 to d_1. An external observer who can query the (near-)optimal policy at every q locates this flip, and thereby learns the *ratio* Δ_1/Δ_2. If σ_2 is a *hard* intervention that fixes the utility-relevant variables to known values, then Δ_2 is computable directly from the known utility function, and the observed switching point turns the ratio into a number: Δ_1 is revealed. This is where the continuum of mixtures matters: locating q_* exactly requires the policy at all $q \in [0,1]$. With only finitely many weights one can only bracket q_*, which bounds Δ_1 to an interval—the approximate regime of Theorem 2.

From utility gaps to probabilities. Δ_1 is a linear function of the unknown interventional probabilities $P(\cdot \mid \sigma_1)$ with coefficients given by the utility function. By taking σ_1 to be a hard intervention on *every* chance variable except one, C_i, all other probabilities collapse to point masses and the single unknown is the local law $P(C_i = c_i \mid do(\boldsymbol{C} \setminus C_i))$, which is thereby recovered. Repeating over i, and over the values at which the other variables are fixed, recovers every utility-relevant local conditional law.

Reading off the graph. These laws are *interventional*, and an intervention is felt, generically, exactly along directed paths. For almost all parameterisations,

$$\begin{aligned} &P(Y \mid do(X = x)) \neq P(Y \mid do(X = x')) \text{ for some } x \neq x' \\ \Longleftrightarrow \quad &\text{there is a directed path from } X \text{ to } Y; \end{aligned}$$

in the binary case, $P(Y \mid do(X=0)) \neq P(Y \mid do(X=1))$ iff X is an ancestor of Y. This detects ancestors. To isolate the *direct parents*, hold every other variable fixed by hard intervention: writing $\boldsymbol{C}_{-ij}$ for all chance variables other than C_i, C_j,

$$P\big(C_i \mid do(C_j = c,\ \boldsymbol{C}_{-ij} = \boldsymbol{c})\big) \text{ depends on } c \quad \Longleftrightarrow \quad C_j \in \mathrm{Pa}(C_i),$$

because fixing the remaining variables blocks every path from C_j to C_i except a direct edge. So the parent set of each C_i, and hence the DAG on $\mathrm{Anc}(U)$, is read off from which fixed values change which recovered laws. The word "generically" is exactly Theorem 1's "almost all": the excluded parameter choices are those for which an edge is present in G yet has no effect on the corresponding conditional (a cancellation), and such an edge is invisible to any decision. Thus

$$\boxed{\text{policy switch} \to \text{utility gap } \Delta_1 \to \text{interventional probability} \to \text{parents} \to (G, P).}$$

Interpretation

The theorem does not say that the agent literally stores a symbolic DAG. Rather, if it stays near-optimal across a sufficiently rich family of intervention-induced domains, then its policy behaviour contains enough information for an *external observer* to reconstruct the utility-relevant causal model. In this information-theoretic sense,

$$\boxed{\text{robust OOD behaviour} \Longrightarrow \text{implicit causal knowledge.}}$$

Two limits are worth keeping in view for discussion. First, this is an identifiability result, not a statement about learning: it says the causal model is *recoverable from* a robust policy, not that any particular training procedure produces one. Second, it is scoped to unmediated tasks and to shifts generated by (mixtures of) local interventions; how far the conclusion survives mediated decisions or richer shift classes is left open.