Causal Abstraction II

Jack Heaney

The University of Edinburgh
Post-Graduate Research in Mathematics

November 6, 2025

Thermostat

$$\begin{tikzcd}
\mathrm{dial} \arrow[dr] \arrow[ddr] & & \mathrm{temp}(t_0) \arrow[dl] \\
& \mathrm{heat}(t_0) \arrow[r] & \mathrm{temp}(t_1) \arrow[dl] \\
& \mathrm{heat}(t_1) & \vdots \\
& & \mathrm{temp}(t_n)
\end{tikzcd}$$

$$\mathrm{dial} \longrightarrow \mathrm{temp}$$

Summary

- Definitions
  1. Recall from Last Time: Causal models, Interventions etc.
  2. Formal notion of model alignment
- Example
  1. Section 2.6: Mech. Interp. Alignment

When is a Causal Abstract Model "Correct"?

Queries

$$A \longrightarrow B$$

- Probabilistic
- Counterfactual

Definitions

Variables/Values

- $\mathbb{V}$
- $X \in \mathbb{V},\ \mathrm{Val}_X \neq \emptyset$

$$\Sigma = (\mathbb{V}, \mathrm{Val})$$

Partial/Total Settings

$$\mathbb{x} \subseteq \bigcup_{X \in \mathbb{X}} \mathrm{Val}_X$$

$$\mathbb{x} \in \mathrm{Val}_{\mathbb{X}}$$ partial

$$\mathbb{v} \in \mathrm{Val}_{\mathbb{V}}$$ total

Projection

$$\mathbb{X} \subseteq \mathbb{Y} \subseteq \mathbb{V}$$

$$\mathbb{y} \longrightarrow \mathrm{Proj}_{\mathbb{X}}(\mathbb{y})$$

Causal Model

$$\mathcal{M} = (\Sigma, \{\mathcal{F}_X\}_{X \in \mathbb{V}})$$

$$\mathcal{F}_X : \mathrm{Val}_{\mathbb{V}} \longrightarrow \mathrm{Val}_X$$

$$\mathcal{F}_{\mathbb{X}} := \{\mathcal{F}_X\}_{X \in \mathbb{X}}$$

Hard Interventions

$$\mathbb{I} \subseteq \mathbb{V}, \quad \mathbb{i} \in \mathrm{Val}_{\mathbb{I}}$$

Replace $\mathcal{F}_X$ with constants functions

$$\mathbb{v} \longmapsto \mathrm{Proj}_X(\mathbb{i}), \forall X \in \mathbb{I}$$

$$\mathrm{Hard}_{\mathbb{I}} := \mathrm{Val}_{\mathbb{I}}$$

A Solution to a Model

Given $\mathcal{M} = (\mathbb{V}, \{\mathcal{F}_X\}_{X\in\mathbb{V}})$, the set of solutions, denoted $\mathrm{Solve}(\mathcal{M})$, is the set of all $\mathbb{v} \in \mathrm{Val}_{\mathbb{V}}$ such that all the equations

$$\mathrm{Proj}_X(\mathbb{v}) = \mathcal{F}_X(\mathbb{v}), \quad X \in \mathbb{V}.$$

Intervention Algebras

Let Λ be a set and $\oplus$ be a binary operation on Λ. We define $(\Lambda, \oplus)$ to be an *intervention algebra* if there exists a signature Σ such that $(\Phi, \circ) \simeq (\Lambda, \oplus)$.

Ordering on Intervention Algebras

Let $(\Lambda, \oplus)$ be an intervention algebra. Define an ordering $\leq$ on elements of Λ as follows:

$$\lambda \preceq \lambda' \quad \Longleftrightarrow \quad \lambda' \oplus \lambda = \lambda'$$

Equivalently,

$$\begin{aligned} \mathbb{x} \preceq \mathbb{y} \quad &\Longleftrightarrow \quad \mathbb{X} \subseteq \mathbb{Y} \text{ and } \mathbb{x} = \mathrm{Proj}_{\mathbb{X}}(\mathbb{y}) \\ &\Longleftrightarrow \quad \mathbb{x} \subseteq \mathbb{y} \end{aligned}$$

Exact transformations

Let $\mathcal{M}$, $\mathcal{M}^\star$ be causal models and let $(\Psi, \circ)$ and $(\Psi, \Box)$ be two intervention algebras where Ψ and $\Psi^\star$ are sets of interventionals on $\mathcal{M}$ and $\mathcal{M}^\star$, respectively.

Furthermore, let $\tau : \mathrm{Val}_{\mathbb{V}} \longrightarrow \mathrm{Val}_{\mathbb{V}^\star}$ and $\omega : \Psi \longrightarrow \Psi^\star$ be two partial surjective functions where ω is $\preceq$-preserving.

Then $\mathcal{M}^\star$ is an *exact transformation* of $\mathcal{M}$ if, for each interventional $\mathfrak{I} \in \Psi$

$$\tau\left(\mathrm{Solve}(\mathcal{M}_{\mathfrak{I}})\right) = \mathrm{Solve}\left(\mathcal{M}^\star_{\omega(\mathfrak{I})}\right)$$

Alignment

An *alignment* between signatures $\Sigma_{\mathcal{L}}$ and $\Sigma_{\mathcal{H}}$ is given by a pair $\langle \Pi, \pi \rangle$ of a partition of $\mathbb{V}_{\mathcal{L}}$

$$\Pi = \{\Pi_{X_{\mathcal{H}}}\}_{X_{\mathcal{H}} \in \mathbb{V}_{\mathcal{H}} \cup \{\perp\}}$$

and a family

$$\pi = \{\pi_{X_{\mathcal{H}}}\}_{X_{\mathcal{H}} \in \mathbb{V}_{\mathcal{H}}}$$

of maps, such that:

1. The partition Π of $\mathbb{V}_{\mathcal{L}}$ consists of non-overlapping, non-empty cells $\Pi_{X_{\mathcal{H}}} \subseteq \mathbb{V}_{\mathcal{L}}$ for each $X_{\mathcal{H}} \in \mathbb{V}_{\mathcal{H}}$, in addition to a (possibly empty) cell $\Pi_{\perp}$;
2. There is a partial surjective map $\pi_{X_{\mathcal{H}}} : \mathrm{Val}_{\Pi_{X_{\mathcal{H}}}} \longrightarrow \mathrm{Val}_{X_{\mathcal{H}}}$ for each $X_{\mathcal{H}} \in \mathbb{V}_{\mathcal{H}}$.

Canonical map for an Alignment

An alignment $\langle \Pi, \pi \rangle$ induces a unique partial function ω^{π} that maps from low-level hard interventions to high-level hard interventions. For $\mathbb{x}_{\mathcal{L}} \in \mathrm{Val}_{\Pi_{X_{\mathcal{H}}}}$ where $\mathbb{X}_{\mathcal{H}} \subseteq \mathbb{V}_{\mathcal{H}}$ and $\Pi_{\mathbb{X}_{\mathcal{H}}} = \cup_{X \in \mathbb{X}_{\mathcal{H}}} \Pi_X$, we define

$$\omega^{\pi}(\mathbb{x}_{\mathcal{L}}) := \bigcup_{X \in \mathbb{V}_{\mathcal{H}}} \pi_{X_{\mathcal{H}}} \left(\mathrm{Proj}_{\Pi_{X_{\mathcal{H}}}}(\mathbb{x}_{\mathcal{L}}) \right)$$

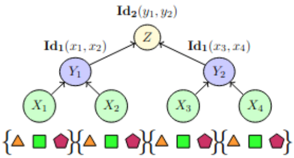
We also have a unique partial function τ^{π} which is ω^{π} restricted to only (low-level) total settings.

Constructive Abstraction

We say that $\mathcal{H}$ is a constructive abstraction of $\mathcal{L}$ under an alignment $\langle \Pi, \pi \rangle$ if and only if $\mathcal{H}$ is an exact transformation of $\mathcal{L}$ under $(\tau^{\pi}, \omega^{\pi})$.

Example

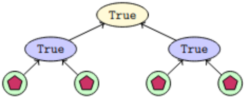
High-Level Model



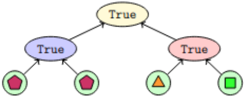
$\mathbf{Id_1}$	△	□	⬠
△	T	F	F
□	F	T	F
⬠	F	F	T

$\mathbf{Id_2}$	T	F
T	T	F
F	F	T

(a) The algorithm.



(b) The total setting of $\mathcal{L}$ determined by the empty intervention.



(c) The total setting of $\mathcal{L}$ determined by the intervention fixing X_3, X_4, and Y_2 to be $\triangle$, $\square$, and `True`.

$$\mathbb{V}_{\mathcal{H}} = \{X_1, X_2, X_3, X_4, Y_1, Y_2, Z\}$$

$$\mathrm{Val}_{X_i} = \{\triangle, \square, \pentagon\} \quad \mathrm{Val}_{Y_j} = \{T, F\} \quad \mathrm{Val}_Z = \{T, F\}$$

$$\mathcal{F}_{X_i} = \pentagon \quad \mathcal{F}_{Y_j}(x_{2j-1}, x_{2j}) = \mathbb{1}[x_{2j-1} = x_{2j}]$$

$$\mathcal{F}_Z(y_1, y_2) = \mathbb{1}[y_1 = y_2]$$

Low-level model

$$\mathbb{V}_{\mathcal{L}} = \{N_1, ..., N_8\}$$
$$\cup \{H_{(i,\ j)} \mid 1 \leq i \leq 2,\ 1 \leq j \leq 8\} \cup \{O_T, O_F\}$$

All sets of a values are the real numbers

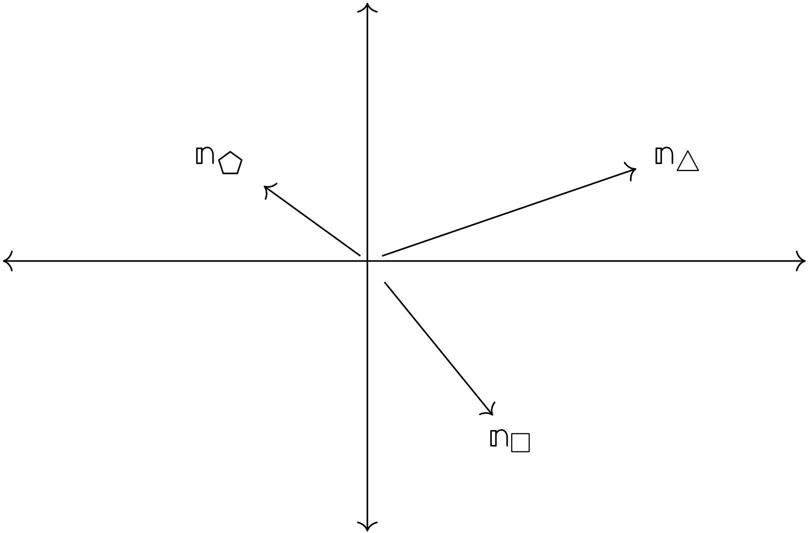
$$\mathcal{F}_{N_k} = 0, \quad 1 \leq k \leq 8$$

$$W_1, W_2 \in \mathbb{R}^{8\times 8}, W_3 \in \mathbb{R}^{8\times 2}$$

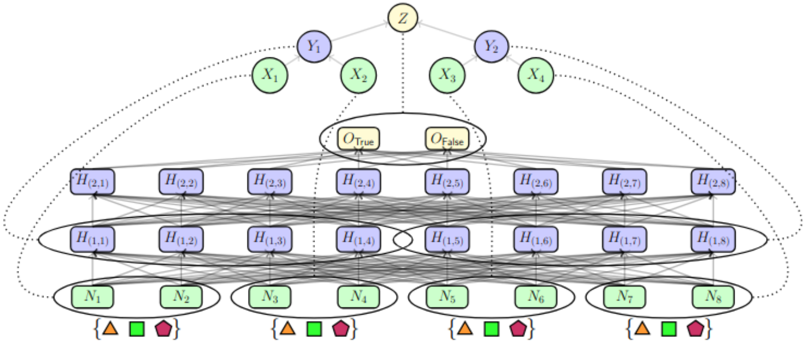
$$\mathcal{F}_{H_{(1,j)}}(\mathbb{n}) = \mathrm{ReLU}((\mathbb{n} W_1)_j) \quad \mathcal{F}_{H_{(1,j)}}(\mathbb{h}_1) = \mathrm{ReLU}((\mathbb{h}_1 W_2)_j)$$

$$\mathcal{F}_{\mathcal{O}_{\mathcal{T}}}(\mathbb{h}_2) = \mathrm{ReLU}((\mathbb{h}_2 W_3)_1) \quad \mathcal{F}_{\mathcal{O}_{\mathcal{F}}}(\mathbb{h}_2) = \mathrm{ReLU}((\mathbb{h}_2 W_3)_2)$$

Representation of Inputs



Alignment Example Visually



Alignment Example Formally

$$\Pi_Z = \{O_T, O_F\} \quad \Pi_{X_k} = \{N_{2k-1}, N_{2k}\} \quad \Pi_{Y_1} = \{H_{(1,j)} : 1 \leq j \leq 4\}$$

$$\Pi_{Y_2} = \{H_{(1,j)} : 5 \leq j \leq 8\} \quad \Pi_\perp = \mathbb{V} \setminus (\Pi_Z \cup \Pi_{Y_1} \cup \Pi_{Y_2} \cup \Pi_{X_1} \cup \Pi_{X_2} \cup \Pi_{X_3} \cup \Pi_{X_4} \cup \Pi_Z)$$

$$\pi_Z(o_T, o_F) = \begin{cases} T & o_T > o_F \\ F & \text{otherwise} \end{cases} \qquad \pi_{X_k}(n_{2k-1}, n_{2k}) = \begin{cases} \triangle & (n_{2k-1}, n_{2k}) = \mathfrak{n}_\triangle \\ \square & (n_{2k-1}, n_{2k}) = \mathfrak{n}_\square \\ \pentagon & (n_{2k-1}, n_{2k}) = \mathfrak{n}_\pentagon \\ \textbf{Undefined} & \text{otherwise} \end{cases}$$

Conclusion

Constructive abstraction will only hold if these stipulated alignments to intermediate variables do not violate the causal laws of $\mathcal{L}$.

"If I had more time,
I would have written a shorter letter."[1]

[1] Unknown (Various)

Thank You!

s2902284@ed.ac.uk