Jointly Learning Consistent Causal Abstractions Over Multiple Interventional Distributions

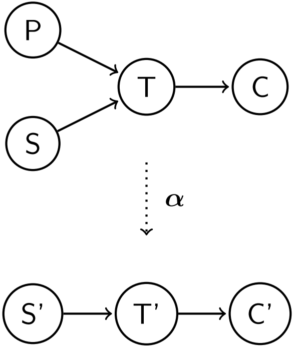
F.M. Zennaro, M. Drávucz, G. Apachitei,
W.D. Widanage, T. Damoulas

University of Warwick

2nd Conference on Causal Learning and Reasoning (CLeaR)

Given two SCMs [2, 3] $\mathcal{M}, \mathcal{M}'$ representing the same system at different levels of detail [5, 1, 4], we want to learn an **abstraction** $\boldsymbol{\alpha}$ between them.

- ✓ rely on multi-scale representations
- ✓ transfer data between different resolutions
- ✓ scale computational expense

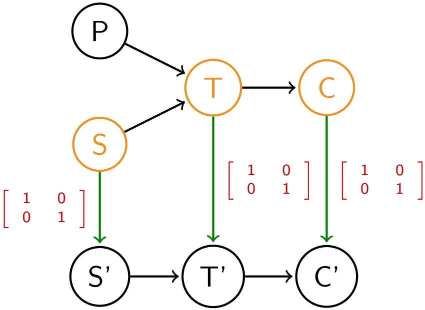


An **abstraction** $\boldsymbol{\alpha}$ is a tuple

$$\langle R, a, \alpha_i \rangle$$

where:

- $R \subseteq \mathcal{X}$ are relevant variables;
- $a : R \to \mathcal{X}'$ is a surjective function between *variables*;
- $\alpha_i : \mathcal{M}[a^{-1}(X_i')] \to \mathcal{M}'[X_i']$ is a collection of surjective functions between *outcomes*.



$$\boldsymbol{\alpha} : \begin{cases} R = \{S, C, T\} \\ a(S) \mapsto S', a(T) \mapsto T', a(C) \mapsto C' \\ \alpha_{S'}(s) \mapsto s, \alpha_{T'}(t) \mapsto t, \alpha_{C'}(c) \mapsto c \end{cases}$$

Abstraction Error [4]

Given two (disjoint set of) variables in $\mathcal{X}'$, we evaluate **abstraction error** in terms of *interventional consistency* $E_{\alpha}(X', Y')$ as the maximum *distance between interventional distributions*.

$$
\begin{array}{ccc}
\mathcal{M}[S] & \xrightarrow[P_{\mathcal{M}_{\iota}}(T|do(S))]{\mu} & \mathcal{M}[T] \\
\Big\downarrow \alpha_{S'} & & \Big\downarrow \alpha_{C'} \\
\mathcal{M}'[S'] & \xrightarrow[P_{\mathcal{M}'_{\iota'}}(T'|do(S'))]{\nu} & \mathcal{M}'[T']
\end{array}
$$

$$E_{\alpha}(S', T') = \max_{\iota} D_{JSD}(\alpha_{T'} \cdot \mu, \nu \cdot \alpha_{S'})$$

Abstraction Errors [4]

An abstraction implies multiple *abstraction errors*.

$$
\begin{array}{ccc}
\mathcal{M}[S] & \xrightarrow{\mu} & \mathcal{M}[T] \\
{\scriptstyle \alpha_{S'}}\downarrow & & \downarrow{\scriptstyle \alpha_{T'}} \\
\mathcal{M}'[S'] & \xrightarrow{\nu} & \mathcal{M}'[T']
\end{array}
\qquad
\begin{array}{ccc}
\mathcal{M}[T] & \xrightarrow{\mu'} & \mathcal{M}[C] \\
{\scriptstyle \alpha_{T'}}\downarrow & & \downarrow{\scriptstyle \alpha_{C'}} \\
\mathcal{M}'[T'] & \xrightarrow{\nu'} & \mathcal{M}'[C']
\end{array}
$$

$$
\begin{array}{ccc}
\mathcal{M}[S] & \xrightarrow{\mu' \circ \mu} & \mathcal{M}[C] \\
{\scriptstyle \alpha_{S'}}\downarrow & & \downarrow{\scriptstyle \alpha_{C'}} \\
\mathcal{M}'[S'] & \xrightarrow{\nu' \circ \nu} & \mathcal{M}'[C']
\end{array}
$$

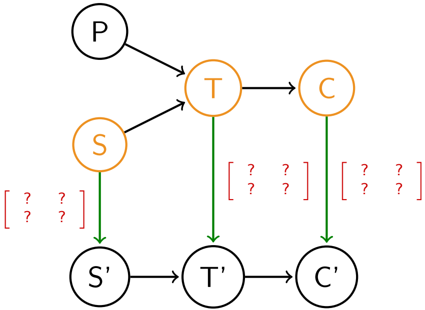
(Global) abstraction error $e(\boldsymbol{\alpha})$ is the maximum abstraction error over all disjoint sets of variable.

$$e(\boldsymbol{\alpha}) = \sup_{\mathbf{X}', \mathbf{Y}' \subseteq \mathcal{X}'} E_{\boldsymbol{\alpha}}(\mathbf{X}', \mathbf{Y}')$$

Problem statement

Given a partially define *abstraction* $\boldsymbol{\alpha}$ in terms of $\langle R, a \rangle$ can I learn α_i as:

$$\min_{\boldsymbol{\alpha}} \; e(\boldsymbol{\alpha})$$



Challenges

$$\alpha_{S'} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}, \alpha_{T'} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}, \alpha_{C'} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}$$

(i) *Multiple related problems*

(ii) *Combinatorial optimization*

(iii) *Surjectivity constraints*

$$\begin{array}{ccc} \mathcal{M}[S] & \xrightarrow{\mu} & \mathcal{M}[T] \\ \downarrow \alpha_{S'} & & \downarrow \alpha_{T'} \\ \mathcal{M}'[S'] & \xrightarrow{\nu} & \mathcal{M}'[T'] \end{array} \qquad \begin{array}{ccc} \mathcal{M}[T] & \xrightarrow{\mu'} & \mathcal{M}[C] \\ \downarrow \alpha_{T'} & & \downarrow \alpha_{C'} \\ \mathcal{M}'[T'] & \xrightarrow{\nu'} & \mathcal{M}'[C'] \end{array}$$

$$\begin{array}{ccc} \mathcal{M}[S] & \xrightarrow{\mu' \circ \mu} & \mathcal{M}[C] \\ \downarrow \alpha_{S'} & & \downarrow \alpha_{C'} \\ \mathcal{M}'[S'] & \xrightarrow{\nu' \circ \nu} & \mathcal{M}'[C'] \end{array}$$

Baselines: parallel or sequential approaches.

Relaxation and parametrization

We address *(ii) combinatorial optimization* by *relaxing* and *parametrizing* all α_i.

$$\min_{\boldsymbol{\alpha}(\mathbf{W})} \; e(\boldsymbol{\alpha}(\mathbf{W}))$$

$$\alpha_{S'}, \alpha_{T'}, \alpha_{C'} \in \mathbb{R}^{2\times 2}$$

$$\begin{bmatrix} 0.7 & 1.2 \\ -0.2 & 3.3 \end{bmatrix}$$

We add *tempering* $t(W) = \frac{e^{\frac{W_ij}{T}}}{\sum_i e^{\frac{W_ij}{T}}}$ along the matrix columns to binarize them.

$$\mathcal{L}_1 : \min_{\boldsymbol{\alpha}(\mathbf{W})} \; e(\boldsymbol{\alpha}(t(\mathbf{W})))$$

$$\alpha_{S'}, \alpha_{T'}, \alpha_{C'} \in [0,1]^{2\times 2}$$

$$t\left(\begin{bmatrix} 0.7 & 1.2 \\ -0.2 & 3.3 \end{bmatrix}\right) = \begin{bmatrix} 0.99 & 0.02 \\ 0.01 & 0.98 \end{bmatrix}$$

Enforcing surjectivity

We address *(iii) surjective constraints* through a *penalty function*:

$$\mathcal{L}_2 : \min_{\mathbf{W}} \sum_{\mathbf{W}} \sum_i \left(1 - \max_j t(W)_{ij}\right)$$

$$\alpha_{S'}, \alpha_{T'}, \alpha_{C'} \in [0,1]^{2\times 2}$$

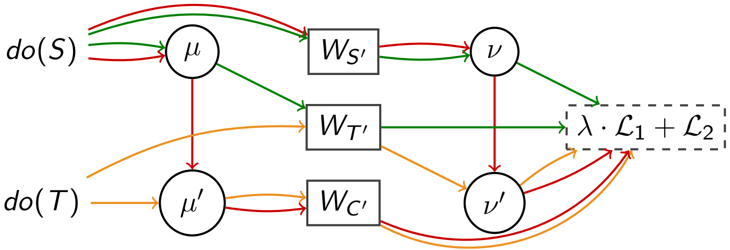
$$\begin{bmatrix} 0.99 & 0.02 \\ 0.01 & 0.98 \end{bmatrix} \overset{\mathcal{L}_2}{\rightsquigarrow}$$

$$(1-0.99)+(1-0.98)$$

Solution by gradient descent

We address *(i) multiple related problems* by *jointly* solving all the problems via *gradient descent*:

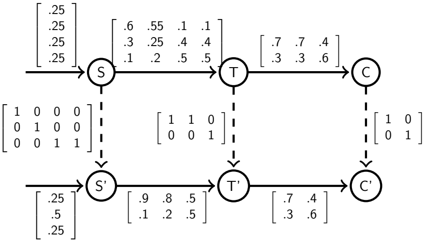
$$\begin{array}{ccc} \mathcal{M}[S] & \xrightarrow{\mu} & \mathcal{M}[T] \\ \downarrow{\scriptstyle \alpha_{S'}} & & \downarrow{\scriptstyle \alpha_{T'}} \\ \mathcal{M}'[S'] & \xrightarrow{\nu} & \mathcal{M}'[T'] \end{array} \qquad \begin{array}{ccc} \mathcal{M}[T] & \xrightarrow{\mu'} & \mathcal{M}[C] \\ \downarrow{\scriptstyle \alpha_{T'}} & & \downarrow{\scriptstyle \alpha_{C'}} \\ \mathcal{M}'[T'] & \xrightarrow{\nu'} & \mathcal{M}'[C'] \end{array} \qquad \begin{array}{ccc} \mathcal{M}[S] & \xrightarrow{\mu' \circ \mu} & \mathcal{M}[C] \\ \downarrow{\scriptstyle \alpha_{S'}} & & \downarrow{\scriptstyle \alpha_{C'}} \\ \mathcal{M}'[S'] & \xrightarrow{\nu' \circ \nu} & \mathcal{M}'[C'] \end{array}$$

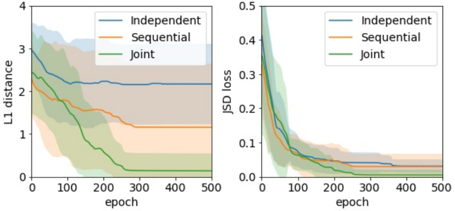


Synthetic Experiments

We evaluated our learning method:

- On multiple synthetic models;
- Against *independent* and *sequential* approach;
- Monitoring *loss functions*, *L1-dist from ground truth*, *wall-clock time*.





Real-World Experiments

We evaluated our learning methods:

- On battery manufacturing data collected by two research groups (LRCS, WMG) using different model;
- Performing abstraction of data from base to abstracted (WMG $\to$ LRCS);
- Evaluating change in performance using aggregated data when predicting out-of-sample (k).

	Training set	**Test Set**	**MSE**
(a)	LRCS[$CG \neq k$]	LRCS[$CG = k$]	1.86 ± 1.75
(b)	LRCS[$CG \neq k$] + WMG	LRCS[$CG = k$]	0.22 ± 0.26
(c)	LRCS[$CG \neq k$] + WMG[$CG \neq k$]	LRCS[$CG = k$] + WMG[$CG = k$]	1.22 ± 0.95

Thanks!

Thank you for listening!

[1] Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

[2] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[3] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

[4] Eigil Fjeldgren Rischel. The category theory of causal models. 2020.

[5] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.