Quantifying Consistency and Information Loss for Causal Abstraction Learning

F.M. Zennaro, P. Turrini, T. Damoulas

University of Warwick

32nd International Joint Conference on Artificial Intelligence

Structural Causal Models

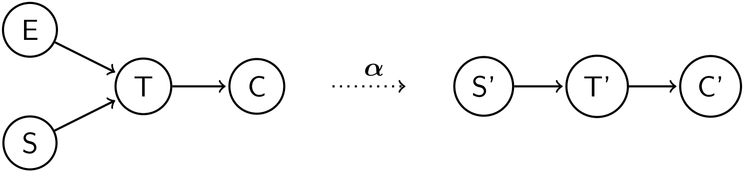
A *structural causal model* (SCM) $\mathcal{M} = \langle \mathcal{X}, \mathcal{U}, \mathcal{F}, \mathcal{P} \rangle$ is a mathematical object representing a causal system [2, 3].

A SCM is associated with a *directed acyclic graph* (DAG)

$$\begin{tikzcd} E \arrow[dr] & & \\ & T \arrow[r] & C \\ S \arrow[ur] & & \end{tikzcd}$$

Abstractions

The same causal system may be represented at different *levels of abstraction* [1].



Given two SCMs we want a formal **abstraction map** α between them.

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

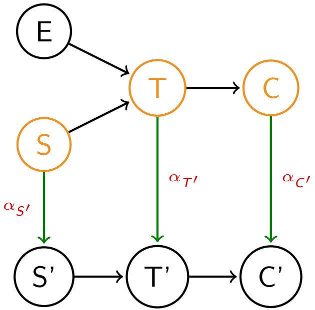
Abstraction Theory [4]

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

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

where:

- R is a set of relevant nodes/variables;
- a is a surjective function between *variables*;
- α_i is a collection of surjective functions between *outcomes*.



Abstraction Error [4]

We evaluate the *quality* of an abstraction in terms of *interventional consistency*.

The **abstraction error** wrt $P(\mathbf{Y}'|do(\mathbf{X}'))$ is the maximum *distance between interventional distributions* in the base and abstracted model.

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu", "P_{\mathcal{M}_\iota}(T|do(S))"'] \arrow[d, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, "\alpha_{T'}"] \\
\mathcal{M}'[S'] \arrow[r, "\nu", "P_{\mathcal{M}'_{\iota'}}(T'|do(S'))"'] & \mathcal{M}'[T']
\end{tikzcd}
$$

$$E(\boldsymbol{\alpha}, S', T') = \max_{s \in \mathcal{M}[S]} D_{JSD}(\alpha_{T'} \cdot \mu, \nu \cdot \alpha_{S'})$$

An abstraction implies multiple *abstraction errors*.

(Global) abstraction error $e(\alpha)$ is the maximum abstraction error over all disjoint sets of variables.

$$\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, "\alpha_{T'}"] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] & \mathcal{M}'[T']
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathcal{M}[T] \arrow[r, "\mu'"] \arrow[d, "\alpha_{T'}"'] & \mathcal{M}[C] \arrow[d, "\alpha_{C'}"] \\
\mathcal{M}'[T'] \arrow[r, "\nu'"] & \mathcal{M}'[C']
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu' \circ \mu"] \arrow[d, "\alpha_{S'}"'] & \mathcal{M}[C] \arrow[d, "\alpha_{C'}"] \\
\mathcal{M}'[S'] \arrow[r, "\nu' \circ \nu"] & \mathcal{M}'[C']
\end{tikzcd}$$

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

Generalizing Abstraction Error

The abstraction error can be expressed more generally as:

$$E_{\boldsymbol{\alpha}}(\mathbf{X}', \mathbf{Y}') = \operatorname*{agg}_{x' \in \mathbf{X}'} D(p, q)$$

$$e(\boldsymbol{\alpha}) = \operatorname*{agg}_{(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}} E_{\boldsymbol{\alpha}}(\mathbf{X}', \mathbf{Y}')$$

parametrized by aggregation functions, distances, paths, intervention sets, and pseudo-inverse.

$$\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}$$

A new family of errors

Interventional consistency (IC)

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}
$$

Consistency projected on the abstracted model.

Interventional information loss (IIL)

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^{+}_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^{+}_{T'}"']
\end{tikzcd}
$$

Loss in abstracting and reconstructing.

A new family of errors

Interventional superresolution information loss (ISIL)

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^+_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^+_{T'}"']
\end{tikzcd}
$$

Loss in reconstructing and abstracting.

Interventional superresolution consistency (ISC)

$$
\begin{tikzcd}
\mathcal{M}[S] \arrow[r, "\mu"] \arrow[d, bend right, "\alpha_{S'}"'] & \mathcal{M}[T] \arrow[d, bend right, "\alpha_{T'}"'] \\
\mathcal{M}'[S'] \arrow[r, "\nu"] \arrow[u, bend right, "\alpha^+_{S'}"'] & \mathcal{M}'[T'] \arrow[u, bend right, "\alpha^+_{T'}"']
\end{tikzcd}
$$

Consistency projected on the base model.

In the paper...

https://arxiv.org/abs/2305.04357

- Properties of the errors (IC, IIL, ISIL, ISC)
- Discussion of other error measure parameters
- Algorithms for evaluating and learning abstractions
- Empirical evaluation

https://github.com/FMZennaro/CausalAbstraction/tree/main/papers/2023-quantifying-consistency-and-infoloss

Thanks!

Thank you for your attention!

References I

[1] Luciano Floridi. The method of levels of abstraction. *Minds and machines*, 18(3):303–329, 2008.

[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.