Frameworks for Representing and Learning Abstract Causal Models

Riccardo **Massidda**
ELLIS Ph.D. Student @ UniPi+UvA

Università di Pisa

ellis

UvA

Introduction

Background

Causal Abstraction with Soft Interventions

Causal Abstraction on Linear Models

Conclusions

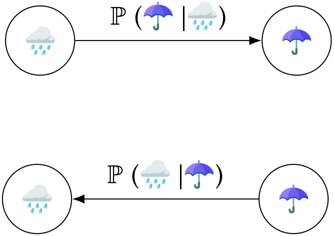
Probabilistic models, such as **Bayesian Networks**, enable the representation of joint probabilities

$\mathbb{P}($🌧️, ☂️$)$

Probabilistic models, such as **Bayesian Networks**, enable the representation of joint probabilities

$$\mathbb{P}(🌧️, ☂️)$$

Causal ordering is **not** necessary for probabilistic modelling.

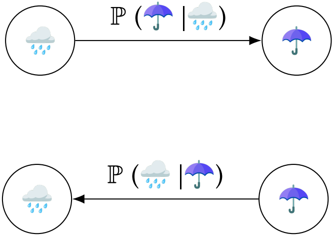


Probabilistic models, such as **Bayesian Networks**, enable the representation of joint probabilities

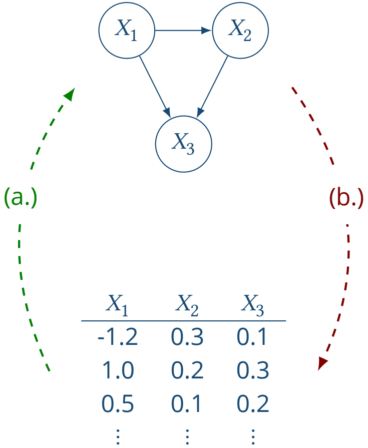
$\mathbb{P}($🌧️, ☂️$)$

Causal ordering is **not** necessary for probabilistic modelling.

...but it's needed to predict the effect of **interventions**!

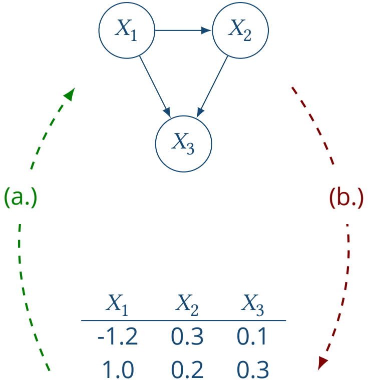


Learning causal models (a.) is challenging and generally requires non-observational data.



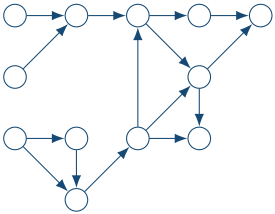
Learning causal models (a.) is challenging and generally requires non-observational data.

We can address it by restricting the **data generating process** (b.).



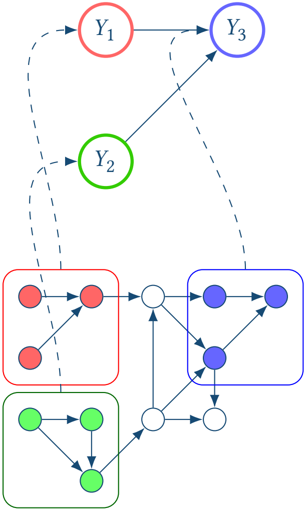
X_1	X_2	X_3
-1.2	0.3	0.1
1.0	0.2	0.3
0.5	0.1	0.2
⋮	⋮	⋮

Causal relations might not be defined on the same **level of detail** of the observed variables.



Causal relations might not be defined on the same **level of detail** of the observed variables.

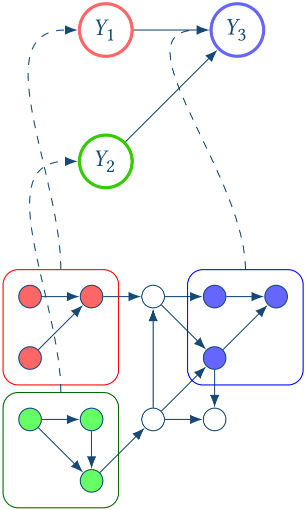
Causal Abstraction assumes the existence of higher-level aggregated *abstract* variables.



Causal relations might not be defined on the same **level of detail** of the observed variables.

Causal Abstraction assumes the existence of higher-level aggregated *abstract* variables.

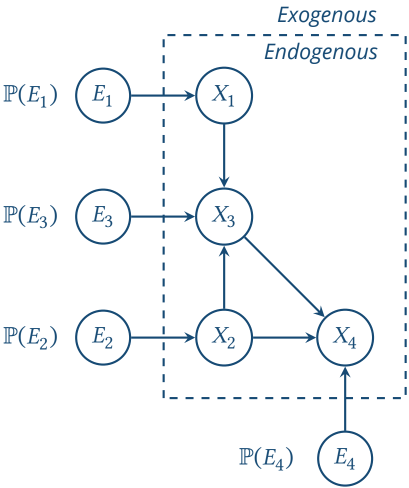
Can we use this to understand or interpret **large** models?



A Structural Causal Model

$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

specifies the deterministic mechanisms $\boldsymbol{f}$ between a set of endogenous variables $\boldsymbol{X}$ and a set of exogenous variables $\boldsymbol{E}$ with distribution $\mathbb{P}_{\boldsymbol{E}}$.



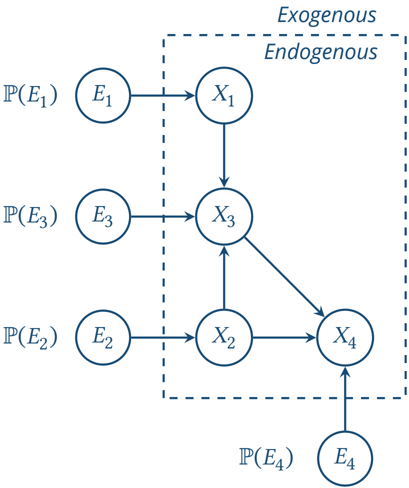
To each *endogenous* variable $X \in \boldsymbol{X}$, we assign an *exogenous* variable $E_X \in \boldsymbol{E}$.
The endogenous mechanism f_X of X is then defined as a function

$$f_X : \mathcal{D}(\mathrm{Pa}(X) \cup E_X) \to \mathcal{D}(X).$$

We define the model reduction

$$\mathcal{M} : \mathcal{D}(\boldsymbol{E}) \to \mathcal{D}(\boldsymbol{X}),$$

whenever the model is **acyclic**.



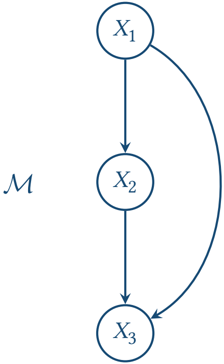
Given an SCM

$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

a subset of variables $\boldsymbol{V} \subset \boldsymbol{X}$ and a setting $\boldsymbol{v} \in \mathcal{D}(\boldsymbol{V})$, a *hard* intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ results in a SCM $\mathcal{M}^i = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}^i, \mathbb{P}_{\boldsymbol{E}})$, where

$$f^i_X = \begin{cases} v_X & X \in \boldsymbol{V} \\ f_X & X \notin \boldsymbol{V}, \end{cases}$$

for each endogenous variable $X \in \boldsymbol{X}$.



Given an SCM

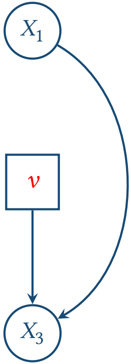
$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

a subset of variables $\boldsymbol{V} \subset \boldsymbol{X}$ and a setting $\boldsymbol{v} \in \mathcal{D}(\boldsymbol{V})$, a *hard* intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ results in a SCM $\mathcal{M}^i = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}^i, \mathbb{P}_{\boldsymbol{E}})$, where

$$f_X^i = \begin{cases} v_X & X \in \boldsymbol{V} \\ f_X & X \notin \boldsymbol{V}, \end{cases}$$

for each endogenous variable $X \in \boldsymbol{X}$.

$\mathcal{M}^i$



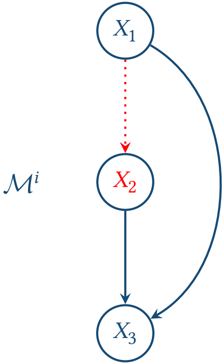
Given an SCM

$$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}}),$$

a subset of variables $\boldsymbol{V} \subset \boldsymbol{X}$ and a set of functions $\boldsymbol{h}$, a *soft* intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{h})$ results in a SCM $\mathcal{M}^i = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}^i, \mathbb{P}_{\boldsymbol{E}})$, where

$$f_X^i = \begin{cases} h_X & X \in \boldsymbol{V} \\ f_X & X \notin \boldsymbol{V}, \end{cases}$$

for each endogenous variable $X \in \boldsymbol{X}$.



Additive Noise Models

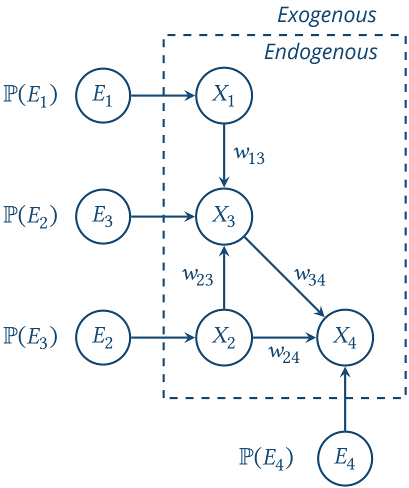
In a linear ANM, endogenous mechanisms have form

$$x_j = \sum_{X_i \in \mathrm{Pa}(X_j)} w_{ij} x_i + e_j,$$

for each $X_j \in \boldsymbol{X}$.

The model reduction is

$$\mathcal{M}(\boldsymbol{e}) = (\mathbf{I} - \mathbf{W})^{-1} \boldsymbol{e}$$
$$= \mathbf{F}^\top \boldsymbol{e}.$$

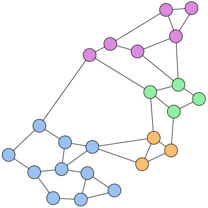


Low-Level SCM

Defined on variables $\boldsymbol{X}$ with exogenous noise $\mathbb{P}_{\boldsymbol{E}}$, structural functions $\boldsymbol{f}$, and interventions $\boldsymbol{I}$.

Low-Level SCM

Defined on variables $\boldsymbol{X}$ with exogenous noise $\mathbb{P}_{\boldsymbol{E}}$, structural functions $\boldsymbol{f}$, and interventions $\boldsymbol{I}$.

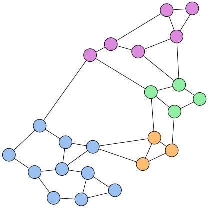
$\mathcal{L} =$ 

Sensor data, raw measurements, or high-dimensional data.

Low-Level SCM

Defined on variables $\boldsymbol{X}$ with exogenous noise $\mathbb{P}_{\boldsymbol{E}}$, structural functions $\boldsymbol{f}$, and interventions $\boldsymbol{I}$.

$\mathcal{L} =$



Sensor data, raw measurements, or high-dimensional data.

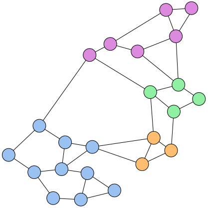
High-Level SCM

Defined on variables $\boldsymbol{Y}$ with exogenous noise $\mathbb{P}_{\boldsymbol{U}}$, structural functions $\boldsymbol{g}$, and interventions $\boldsymbol{J}$.

Low-Level SCM

Defined on variables $\boldsymbol{X}$ with exogenous noise $\mathbb{P}_{\boldsymbol{E}}$, structural functions $\boldsymbol{f}$, and interventions $\boldsymbol{I}$.

$\mathcal{L} =$

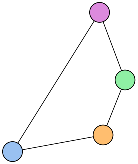


Sensor data, raw measurements, or high-dimensional data.

High-Level SCM

Defined on variables $\boldsymbol{Y}$ with exogenous noise $\mathbb{P}_{\boldsymbol{U}}$, structural functions $\boldsymbol{g}$, and interventions $\boldsymbol{J}$.

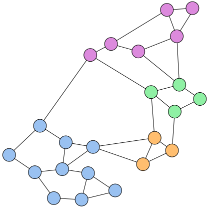
$\mathcal{H} =$



Summary statistics, overviews, or low-dimensional data.

Low-Level SCM

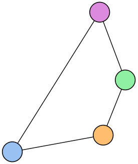
Defined on variables $\boldsymbol{X}$ with exogenous noise $\mathbb{P}_{\boldsymbol{E}}$, structural functions $\boldsymbol{f}$, and interventions $\boldsymbol{I}$.

$\mathcal{L} =$ 

Sensor data, raw measurements, or high-dimensional data.

High-Level SCM

Defined on variables $\boldsymbol{Y}$ with exogenous noise $\mathbb{P}_{\boldsymbol{U}}$, structural functions $\boldsymbol{g}$, and interventions $\boldsymbol{J}$.

$\mathcal{H} =$ 

Summary statistics, overviews, or low-dimensional data.

$$|\boldsymbol{X}| \gg |\boldsymbol{Y}|$$

Given two SCMs

- $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}})$ with admissible interventions $\boldsymbol{I}$,
- $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{g}, \mathbb{P}_{\boldsymbol{U}})$ with admissible interventions $\boldsymbol{J}$,

Causal Abstraction consists of two *surjective* functions

Given two SCMs

- $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}})$ with admissible interventions $\boldsymbol{I}$,
- $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{g}, \mathbb{P}_{\boldsymbol{U}})$ with admissible interventions $\boldsymbol{J}$,

Causal Abstraction consists of two *surjective* functions

- $\tau : \mathcal{D}(\boldsymbol{X}) \to \mathcal{D}(\boldsymbol{Y})$ (Endogenous Map)

Given two SCMs

- $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}})$ with admissible interventions $\boldsymbol{I}$,
- $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{g}, \mathbb{P}_{\boldsymbol{U}})$ with admissible interventions $\boldsymbol{J}$,

Causal Abstraction consists of two *surjective* functions

- $\tau : \mathcal{D}(\boldsymbol{X}) \to \mathcal{D}(\boldsymbol{Y})$ (Endogenous Map)
- $\gamma : \mathcal{D}(\boldsymbol{E}) \to \mathcal{D}(\boldsymbol{U})$ (Exogenous Map)

Given two SCMs

- $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{E}})$ with admissible interventions $\boldsymbol{I}$,
- $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{g}, \mathbb{P}_{\boldsymbol{U}})$ with admissible interventions $\boldsymbol{J}$,

Causal Abstraction consists of two *surjective* functions

- $\tau : \mathcal{D}(\boldsymbol{X}) \to \mathcal{D}(\boldsymbol{Y})$ (Endogenous Map)
- $\gamma : \mathcal{D}(\boldsymbol{E}) \to \mathcal{D}(\boldsymbol{U})$ (Exogenous Map)

that induce a unique intervention map $\omega : \boldsymbol{I} \to \boldsymbol{J}$ such that

$$\omega(i) = j \iff \mathrm{Rst}(j) = \{\tau(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{Rst}(i)\}$$

$$\mathrm{Rst}(\boldsymbol{V} \leftarrow \boldsymbol{v}) = \{\boldsymbol{x} \mid \boldsymbol{x}_{\boldsymbol{V}} = \boldsymbol{v}\}$$

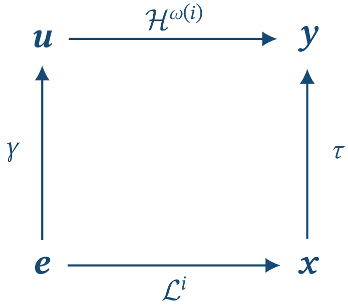
$\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$.

$$\iff$$

$$\tau \circ \mathcal{L}^i = \mathcal{H}^{\omega(i)} \circ \gamma$$

$\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$.

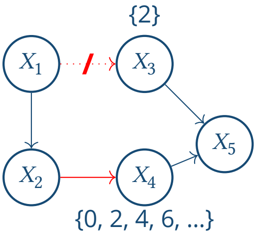
$$\iff$$



Intervention Map

🚨 The intervention map is defined for **hard** interventions only.

🚨 The intervention map does not have an **explicit form**.

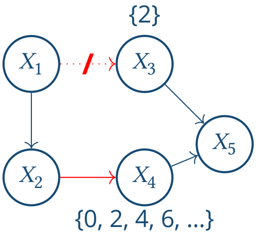


$i = (X_3 \leftarrow 2, X_4 \leftarrow 2X_2)$

The **Soft Restriction** of an intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{h})$ contains all the values that an intervened model can assume.

$$\begin{aligned}\mathrm{SoftRst}(\mathcal{M}^i) &= \{\boldsymbol{x} \in \mathbb{R}^5 \mid x_3 = 2, x_4 \in \mathrm{Image}(\lambda x.2x)\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^5 \mid x_3 = 2, \mathrm{Even}(x_4)\}.\end{aligned}$$

$$\begin{bmatrix} 0.5 \\ -0.2 \\ 2 \\ 16 \\ 9.4 \end{bmatrix} \in \mathrm{SoftRst}(\mathcal{M}^i), \begin{bmatrix} 0.5 \\ -0.2 \\ 2 \\ 7 \\ 9.4 \end{bmatrix} \notin \mathrm{SoftRst}(\mathcal{M}^i)$$



$i = (X_3 \leftarrow 2, X_4 \leftarrow 2X_2)$

The **Soft Restriction** of an intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{h})$ contains all the values that an intervened model can assume.

$$\begin{aligned}\mathrm{SoftRst}(\mathcal{M}^i) &= \{\boldsymbol{x} \in \mathbb{R}^5 \mid x_3 = 2, x_4 \in \mathrm{Image}(\lambda x.2x)\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^5 \mid x_3 = 2, \mathrm{Even}(x_4)\}.\end{aligned}$$

$$\begin{bmatrix} 0.5 \\ -0.2 \\ 2 \\ 16 \\ 9.4 \end{bmatrix} \in \mathrm{SoftRst}(\mathcal{M}^i), \begin{bmatrix} 0.5 \\ -0.2 \\ 2 \\ 7 \\ 9.4 \end{bmatrix} \notin \mathrm{SoftRst}(\mathcal{M}^i)$$

Then, we define ω as

$$\omega(i) = j \iff \mathrm{SoftRst}(\mathcal{H}^j) = \left\{\tau(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}^i)\right\}.$$

Abstraction Ambiguity

Massidda, Geiger, et al. (2023)

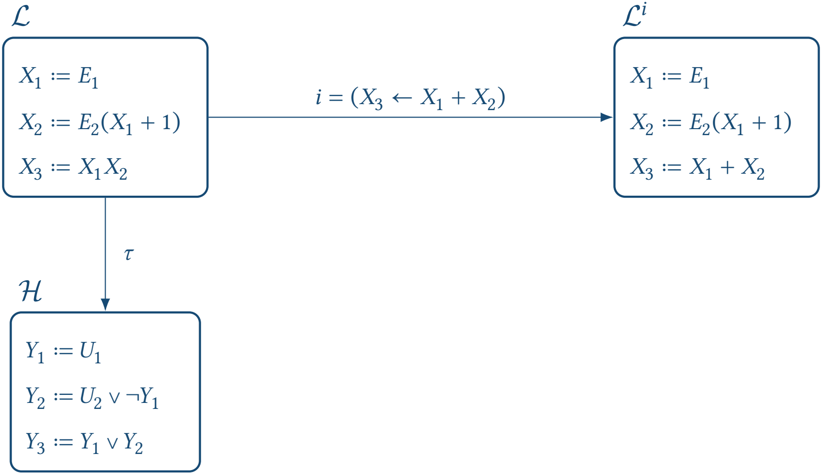
$\mathcal{L}$

$$X_1 := E_1$$
$$X_2 := E_2(X_1 + 1)$$
$$X_3 := X_1 X_2$$

$\downarrow \tau$

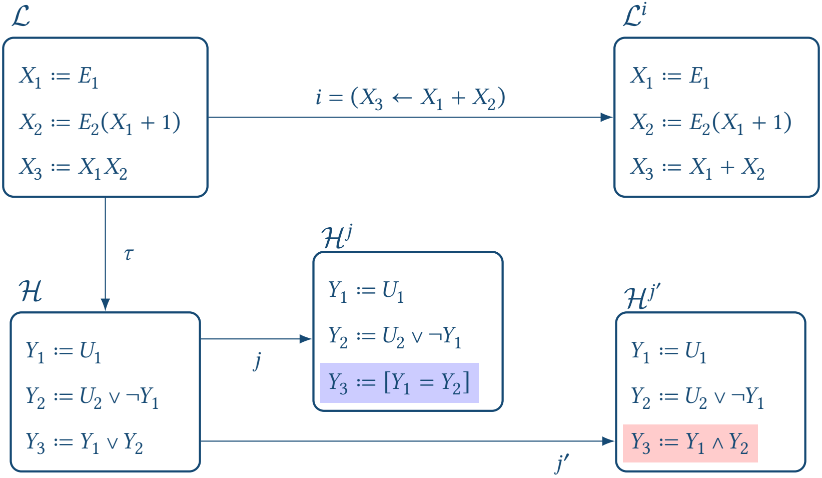
$\mathcal{H}$

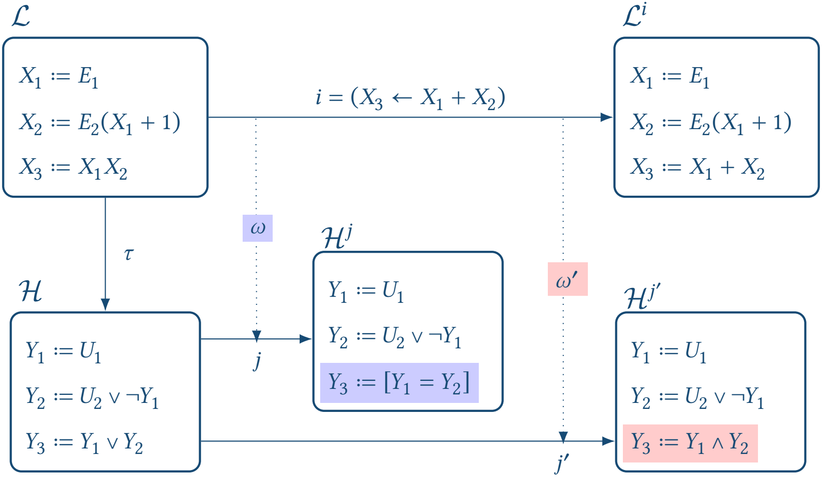
$$Y_1 := U_1$$
$$Y_2 := U_2 \vee \neg Y_1$$
$$Y_3 := Y_1 \vee Y_2$$


$\mathcal{L}$
$X_1 := E_1$
$X_2 := E_2(X_1 + 1)$
$X_3 := X_1 X_2$
$i = (X_3 \leftarrow X_1 + X_2)$
$\mathcal{L}^i$
$X_1 := E_1$
$X_2 := E_2(X_1 + 1)$
$X_3 := X_1 + X_2$
τ
$\mathcal{H}$
$Y_1 := U_1$
$Y_2 := U_2 \vee \neg Y_1$
$Y_3 := Y_1 \vee Y_2$

Abstraction Ambiguity

Massidda, Geiger, et al. (2023)




$\mathcal{L}$
$X_1 := E_1$
$X_2 := E_2(X_1 + 1)$
$X_3 := X_1 X_2$
$i = (X_3 \leftarrow X_1 + X_2)$
$\mathcal{L}^i$
$X_1 := E_1$
$X_2 := E_2(X_1 + 1)$
$X_3 := X_1 + X_2$
τ
ω
ω'
$\mathcal{H}$
$Y_1 := U_1$
$Y_2 := U_2 \vee \neg Y_1$
$Y_3 := Y_1 \vee Y_2$
j
$\mathcal{H}^j$
$Y_1 := U_1$
$Y_2 := U_2 \vee \neg Y_1$
$Y_3 := [Y_1 = Y_2]$
j'
$\mathcal{H}^{j'}$
$Y_1 := U_1$
$Y_2 := U_2 \vee \neg Y_1$
$Y_3 := Y_1 \wedge Y_2$

Y_1	Y_2	$Y_1 \wedge Y_2$	$[Y_1 = Y_2]$
F	F	F	T
F	T	F	F
T	F	F	F
T	T	T	T

The two high-level interventions differ only for values that are **never** reached by the model for any exogenous configuration.

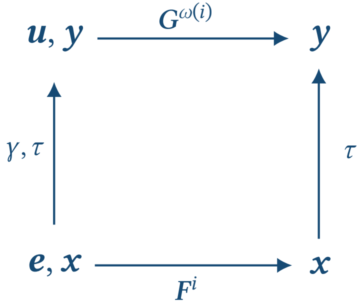
$\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$ on **soft** interventions

$$\iff$$

$$\tau \circ \boldsymbol{F}^i = \boldsymbol{G}^{\omega(i)} \circ [\gamma, \tau]$$

$\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$ on **soft** interventions

$$\iff$$



By generalizing the restriction set and testing consistency for each abstract variable, we can uniquely define ω for *soft interventions* such that

$$\omega(i) = (Y \leftarrow \tau_Y \circ \boldsymbol{F}^i \circ \tau^{-1}_{\mathrm{Pa}(Y)})$$

Questions in Causal Abstraction

🚨 Which causal **graphs** are consistent with abstraction?

🚨 Which causal **mechanisms** are consistent with abstraction?

$\mathcal{H}$ is a $\mathbf{T}$-abstraction of $\mathcal{L}$

$$\iff$$

$\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$ and $\tau(\boldsymbol{x}) = \mathbf{T}^\top \boldsymbol{x}$, where $\mathbf{T} \in \mathbb{R}^{|\boldsymbol{X}| \times |\boldsymbol{Y}|}$.

Relevant Variables

$$\mathbf{T} = \begin{vmatrix} \mathbf{0.43} & 0 & 0 \\ 0 & 0 & 0 \\ \mathbf{0.71} & 0 & 0 \\ 0 & \mathbf{0.52} & 0 \\ 0 & \mathbf{-0.12} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{0.98} \end{vmatrix}$$

The set of relevant variables of an abstract variable $Y \in \boldsymbol{Y}$ is the subset of concrete variables $\Pi_R(Y)$ on which it depends through $\mathbf{T}$.

Relevant Variables

$$\mathbf{T} = \begin{vmatrix} \mathbf{0.43} & 0 & 0 \\ 0 & 0 & 0 \\ \mathbf{0.71} & 0 & 0 \\ 0 & \mathbf{0.52} & 0 \\ 0 & \mathbf{-0.12} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{0.98} \end{vmatrix}$$

The set of relevant variables of an abstract variable $Y \in \boldsymbol{Y}$ is the subset of concrete variables $\Pi_R(Y)$ on which it depends through $\mathbf{T}$.

Lemma 6.3.1, p. 4:
Relevant variables **must** be disjoint.

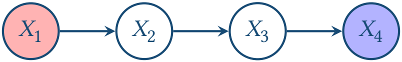
Intervention Map

As a consequence of disjointness and linearity,
the intervention map ω is uniquely defined.

$$\omega(\boldsymbol{V} \leftarrow \boldsymbol{v}) = (Y \leftarrow y) \iff \boldsymbol{V} = \Pi_R(Y) \text{ and } y = \boldsymbol{t}_Y^\top \boldsymbol{v}.$$

$\mathbf{T}$-direct Path

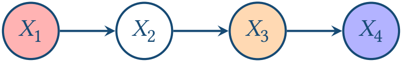
A directed path between two variables is $\mathbf{T}$-direct
if and only if any other variable on the path is not relevant.



$$X_1 \xrightarrow{\mathbf{T}} X_4$$

$\mathbf{T}$-direct Path

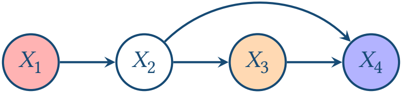
A directed path between two variables is $\mathbf{T}$-direct
if and only if any other variable on the path is not relevant.



$$X_1 \overset{\mathbf{T}}{\not\to} X_4$$

$\mathbf{T}$-direct Path

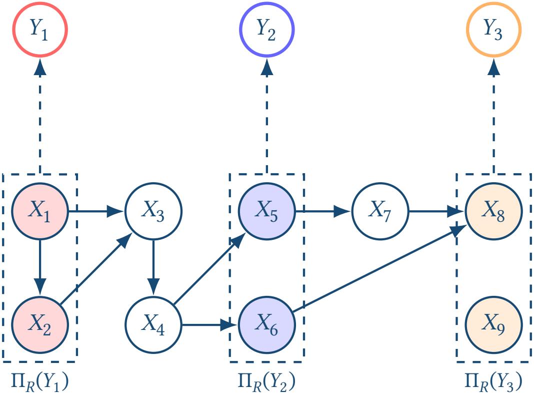
A directed path between two variables is $\mathbf{T}$-direct
if and only if any other variable on the path is not relevant.



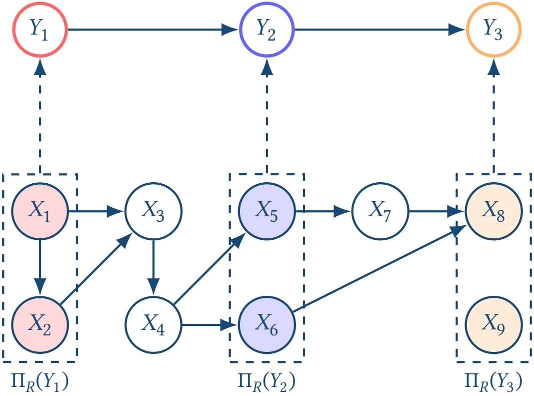
$$X_1 \xrightarrow{\mathbf{T}} X_4$$

Lemma 6.3.3, p. 77:

Let $X_1 \in \Pi_R(Y_1)$ and $X_2 \in \Pi_R(Y_2)$. If $X_1 \xrightarrow{\mathbb{T}} X_2$ in $\mathcal{L}$, then

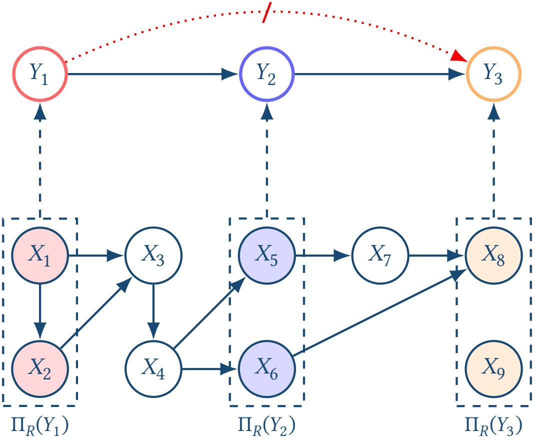


Lemma 6.3.3, p. 77:
Let $X_1 \in \Pi_R(Y_1)$ and $X_2 \in \Pi_R(Y_2)$. If $X_1 \xrightarrow{\mathrm{T}} X_2$ in $\mathcal{L}$, then $Y_1 \to Y_2$ in $\mathcal{H}$.

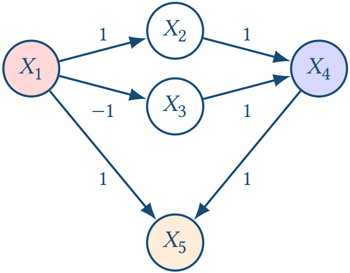


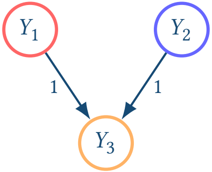
Lemma 6.3.3, p. 77:

Let $X_1 \in \Pi_R(Y_1)$ and $X_2 \in \Pi_R(Y_2)$. If $X_1 \xrightarrow{\mathrm{T}} X_2$ in $\mathcal{L}$, then $Y_1 \to Y_2$ in $\mathcal{H}$.



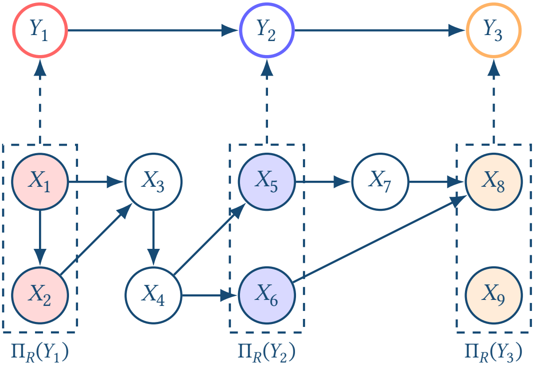
With cancelling paths, things get nasty!





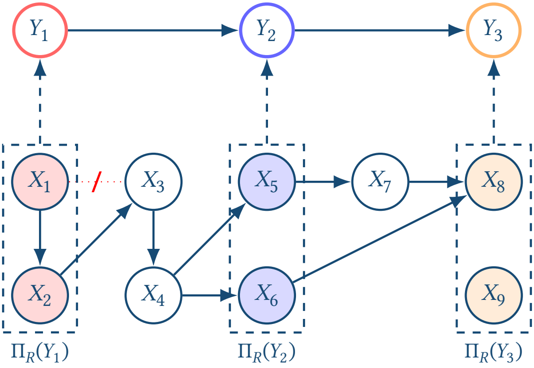
Theorem 6.3.5, p. 78:

Let $Y_1 \to Y_2$ in $\mathcal{H}$. Then, $\forall X_1 \in \Pi_R(Y_1)$, $\exists X_2 \in \Pi_R(Y_2)$ s.t. $X_1 \xrightarrow{\mathbb{T}} X_2$ in $\mathcal{L}$.



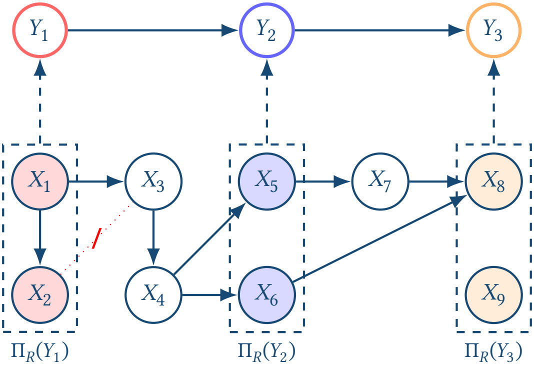
$\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$

Theorem 6.3.5, p. 78:
Let $Y_1 \to Y_2$ in $\mathcal{H}$. Then, $\forall X_1 \in \Pi_R(Y_1)$, $\exists X_2 \in \Pi_R(Y_2)$ s.t. $X_1 \xrightarrow{\mathbb{T}} X_2$ in $\mathcal{L}$.



$\mathcal{H}$ is **not** a τ-abstraction of $\mathcal{L}$

Theorem 6.3.5, p. 78:
Let $Y_1 \to Y_2$ in $\mathcal{H}$. Then, $\forall X_1 \in \Pi_R(Y_1)$, $\exists X_2 \in \Pi_R(Y_2)$ s.t. $X_1 \xrightarrow{\mathbb{T}} X_2$ in $\mathcal{L}$.



$\mathcal{H}$ is **not** a τ-abstraction of $\mathcal{L}$

The exogenous abstraction function is a linear transformation

$$\gamma(\boldsymbol{e}) = \mathbf{S}^{\top}\boldsymbol{e},$$

where $\mathbf{S} = \mathbf{F}\mathbf{T}\mathbf{G}^{-1}$.

$$\mathbf{S} = \begin{vmatrix} \mathbf{0.43} & 0 & 0 \\ \mathbf{0.22} & 0 & 0 \\ \mathbf{0.71} & 0 & 0 \\ 0 & \mathbf{0.52} & 0 \\ 0 & \mathbf{-0.12} & 0 \\ 0 & 0 & \mathbf{-1.02} \\ 0 & 0 & \mathbf{0.98} \end{vmatrix}$$

The exogenous abstraction function is a linear transformation

$$\gamma(\boldsymbol{e}) = \mathbf{S}^\top \boldsymbol{e},$$

where $\mathbf{S} = \mathbf{F}\mathbf{T}\mathbf{G}^{-1}$.

$$\mathbf{S} = \begin{vmatrix} \mathbf{0.43} & 0 & 0 \\ \mathbf{0.22} & 0 & 0 \\ \mathbf{0.71} & 0 & 0 \\ 0 & \mathbf{0.52} & 0 \\ 0 & \mathbf{-0.12} & 0 \\ 0 & 0 & \mathbf{-1.02} \\ 0 & 0 & \mathbf{0.98} \end{vmatrix}$$

The set of **block** variables of an abstract variable $Y \in \boldsymbol{Y}$ is the subset of concrete variables $\Pi(Y)$ on which it depends through $\mathbf{S}$.

Properties of $\mathbb{T}$-abstraction

The exogenous abstraction function is a linear transformation

$$\gamma(\boldsymbol{e}) = \mathbf{S}^\top \boldsymbol{e},$$

where $\mathbf{S} = \mathbf{F}\mathbf{T}\mathbf{G}^{-1}$.

$$\mathbf{S} = \begin{bmatrix} \mathbf{0.43} & 0 & 0 \\ \mathbf{0.22} & 0 & 0 \\ \mathbf{0.71} & 0 & 0 \\ 0 & \mathbf{0.52} & 0 \\ 0 & \mathbf{-0.12} & 0 \\ 0 & 0 & \mathbf{-1.02} \\ 0 & 0 & \mathbf{0.98} \end{bmatrix}$$

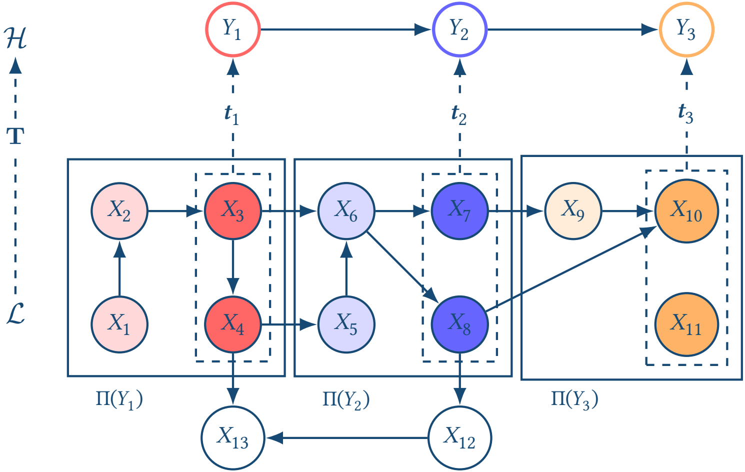
The set of **block** variables of an abstract variable $Y \in \boldsymbol{Y}$ is the subset of concrete variables $\Pi(Y)$ on which it depends through $\mathbf{S}$.

Lemma 6.3.8, p. 81:
Relevant variables are a subset of block variables.

Lemma 6.3.9, p. 82:
Block variables must be disjoint.

Theorem 6.3.10, p. 82:
Block variables follow abstract ordering.

Properties of $\mathbb{T}$-abstraction



$\mathcal{H}$ is a $\mathbf{T}$-abstraction of $\mathcal{L}$

$$\Longleftrightarrow$$

$$\tau \circ \mathcal{L}^i = \mathcal{H}^{\omega(i)} \circ \gamma$$

$\mathcal{H}$ is a $\mathbf{T}$-abstraction of $\mathcal{L}$

$$\iff$$

$$Y_i \prec_{\mathcal{H}} Y_j \iff \Pi(Y_i) \prec_{\mathcal{L}} \Pi(Y_j)$$

$$\mathbf{W}_{ij}\boldsymbol{s}_j = m_{ij}\boldsymbol{t}_i$$

$\mathcal{H}$ is a $\mathbf{T}$-abstraction of $\mathcal{L}$

$$\Longleftrightarrow$$

$$Y_i \prec_{\mathcal{H}} Y_j \iff \Pi(Y_i) \prec_{\mathcal{L}} \Pi(Y_j)$$

$$\mathbf{W}_{ij}(\mathbf{I} - \mathbf{W}_{jj})^{-1}\boldsymbol{t}_j = m_{ij}\boldsymbol{t}_i$$

Parametrical Properties of $\mathbf{T}$-abstraction

$\mathcal{H}$ is a $\mathbf{T}$-abstraction of $\mathcal{L}$

$$\iff$$

$$Y_i \prec_{\mathcal{H}} Y_j \iff \Pi(Y_i) \prec_{\mathcal{L}} \Pi(Y_j)$$

$$\mathbf{W}_{ij}(\mathbf{I} - \mathbf{W}_{jj})^{-1}\boldsymbol{t}_j = m_{ij}\boldsymbol{t}_i$$

This characterization enables testing for $\mathbf{T}$-abstraction in closed form.

Causal Abstraction Learning with Non-Gaussian Noise

Assuming **non-Gaussian noise**, linear ANMs are identifiable from observational data (Shimizu et al. 2006).

$$\boldsymbol{e}^{(i)} \sim \text{Exponential for } i = 1, \ldots, |\mathcal{D}_{\mathcal{L}}|,$$

$$\boldsymbol{x}^{(i)} = \mathcal{L}(\boldsymbol{e}^{(i)}) \qquad \text{for } i = 1, \ldots, |\mathcal{D}_{\mathcal{L}}|,$$

Causal Abstraction Learning with Non-Gaussian Noise

Assuming **non-Gaussian noise**, linear ANMs are identifiable from observational data (Shimizu et al. 2006).

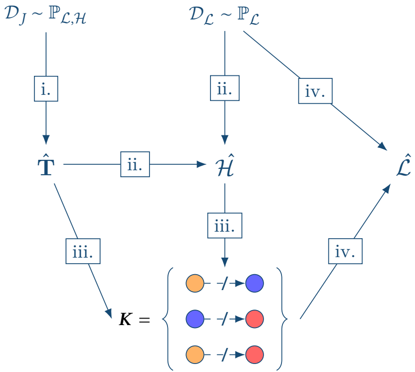
What about abstractions?

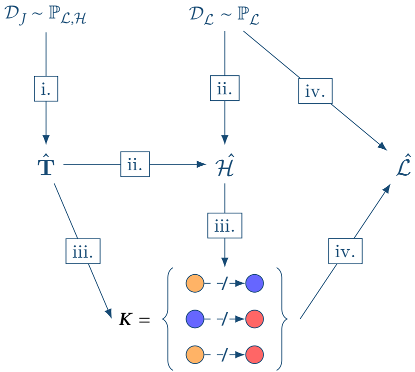
$$\boldsymbol{e}^{(i)} \sim \text{Exponential for } i = 1, \dots, |\mathcal{D}_{\mathcal{L}}|,$$

$$\boldsymbol{x}^{(i)} = \mathcal{L}(\boldsymbol{e}^{(i)}) \quad \text{for } i = 1, \dots, |\mathcal{D}_{\mathcal{L}}|,$$

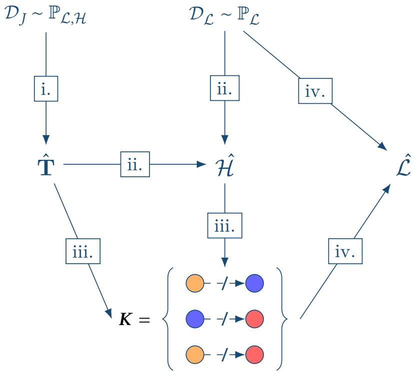
$$\boldsymbol{y}^{(i)} = \mathcal{H}(\gamma(\boldsymbol{e}^{(i)})) \quad \text{for } i = 1, \dots, |\mathcal{D}_{J}|,$$

such that $|\mathcal{D}_J| \ll |\mathcal{D}_{\mathcal{L}}|$.

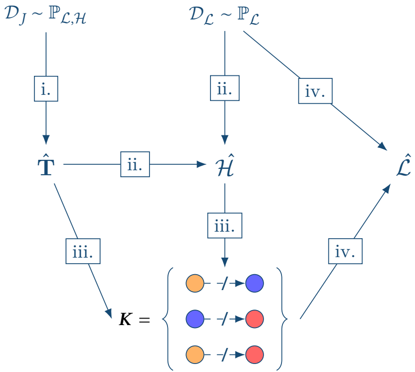

$\mathcal{D}_J \sim \mathbb{P}_{\mathcal{L},\mathcal{H}}$
$\mathcal{D}_\mathcal{L} \sim \mathbb{P}_\mathcal{L}$
i.
ii.
iv.
$\hat{\mathbf{T}}$
ii.
$\hat{\mathcal{H}}$
$\hat{\mathcal{L}}$
iii.
iii.
iv.
$K =$



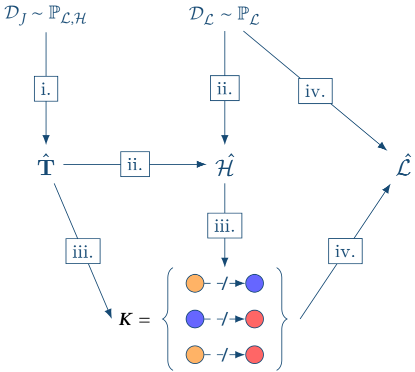
i. Recover $\hat{\mathbf{T}}$ from $\mathcal{D}_J$



i. Recover $\hat{\mathbf{T}}$ from $\mathcal{D}_J$
ii. Abstract $\mathcal{D}_\mathcal{L}$ and recover $\hat{\mathcal{H}}$



i. Recover $\hat{\mathbf{T}}$ from $\mathcal{D}_J$
ii. Abstract $\mathcal{D}_{\mathcal{L}}$ and recover $\hat{\mathcal{H}}$
iii. Define constraints K from $\hat{\mathcal{H}}, \hat{\mathbf{T}}$



i. Recover $\hat{\mathbf{T}}$ from $\mathcal{D}_J$
ii. Abstract $\mathcal{D}_\mathcal{L}$ and recover $\hat{\mathcal{H}}$
iii. Define constraints $\boldsymbol{K}$ from $\hat{\mathcal{H}}, \hat{\mathbf{T}}$
iv. Recover $\hat{\mathcal{L}}$ from $\mathcal{D}_\mathcal{L}$ and $\boldsymbol{K}$

Algorithm 1: Abs-LiNGAM

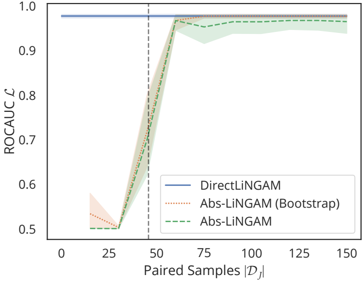
Input: Concrete Observational Dataset $\mathcal{D}_{\mathcal{L}}$,
Joint Observational Dataset $\mathcal{D}_J$.

Result: Abstraction function $\hat{\mathbf{T}} \in \mathbb{R}^{d \times b}$,
Abstract adjacency matrix $\hat{\mathbf{M}} \in \mathbb{R}^{b \times b}$,
Concrete adjacency matrix $\hat{\mathbf{W}} \in \mathbb{R}^{d \times d}$.

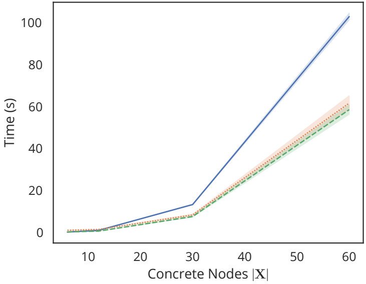
$\hat{\mathbf{T}} \leftarrow \arg\min_{\mathbf{T} \in \mathbb{R}^{b \times d}} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_J} \| \boldsymbol{x}^\top \mathbf{T} - \boldsymbol{y}^\top \|_2^2$;

for $Y_i \in \boldsymbol{Y}$ **do** ▷ Select Relevant Variables
  $\hat{\Pi}_R(Y_i) \leftarrow \{X_k \in \boldsymbol{X} \mid [\hat{\boldsymbol{t}}_i]_k \neq 0\}$
end

$\mathcal{D}_{\hat{\mathcal{H}}} \leftarrow \{\hat{\mathbf{T}}^\top \boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}_{\mathcal{L}}\}$ ▷ Create Abstract Dataset
$\hat{\mathbf{M}} \leftarrow$ DirectLiNGAM$(\mathcal{D}_{\hat{\mathcal{H}}}, \varnothing)$ ▷ Abstract Discovery
$\boldsymbol{K} \leftarrow \varnothing$
for $Y_i, Y_j \in \boldsymbol{Y}$ **do** ▷ Collect Prior Knowledge
  if $Y_i \not\dashrightarrow Y_j$ **then** ▷ Check Ancestorship in $\hat{\mathbf{M}}$
    for $X_k \in \hat{\Pi}_R(Y_i)$, $X_h \in \hat{\Pi}_R(Y_j)$ **do**
      $\boldsymbol{K} \leftarrow \boldsymbol{K} \cup \{X_k \not\dashrightarrow X_h\}$
    end
  end
end
$\hat{\mathbf{W}} \leftarrow$ DirectLiNGAM$(\mathcal{D}_{\mathcal{L}}, \boldsymbol{K})$ ▷ Concrete Discovery



(a) Performance over Paired Samples $|\mathcal{D}_J|$



(b) Execution Time (s) over Graph Size $|X|$

Introducing abstract information in the LiNGAM pipeline, we gain significant speedup (2x) in execution time (b, *right*) without performance loss (a, *left*) on the retrieval of the concrete model ($|X| \in [25, 50], |Y| = 5$).

Introduction

Background

Causal Abstraction with Soft Interventions

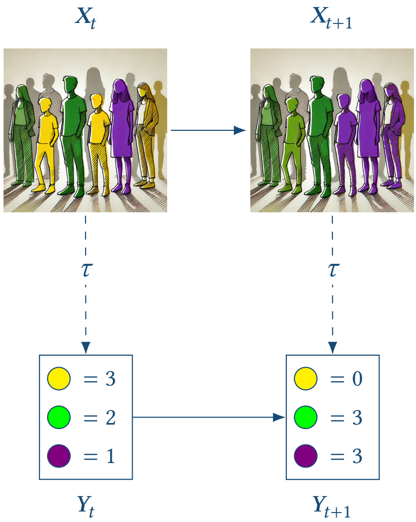
Causal Abstraction on Linear Models

Conclusions

Surrogate models are usually trained for **predictive** tasks.

Causal Abstraction enables the training of interventionally consistent surrogate models.

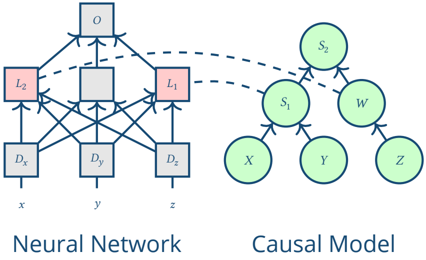
Dyer et al. (2024) shows how to fasten policy evaluation by abstracting SIRS epidemiological models.



The **interpretation** of neural networks is strongly related to causal queries: why? what if?

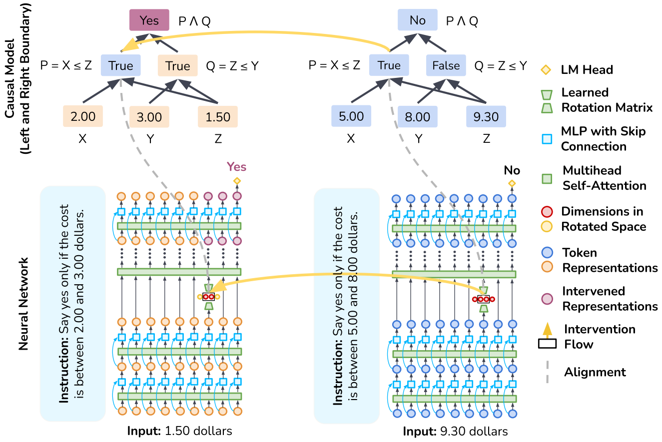
Causal Abstraction provides a framework to determine whether a neural network implements a causal model.

Geiger, Wu, et al. (2021) also shows how to enforce causal constraints when training neural networks.



Applications: Causal Mechanisms in LLMs

Wu et al. (2024)



It works for LLMs too!

Conclusion

📜 **Recap:**

Conclusion

📜 **Recap:**

- **Causal Abstraction** enables concise representation of complex causal relations.

Conclusion

📜 **Recap:**

- **Causal Abstraction** enables concise representation of complex causal relations.
- τ-abstraction provides a and explicit **intervention map** for generic causal models.

Conclusion

📜 **Recap:**

- **Causal Abstraction** enables concise representation of complex causal relations.
- τ-abstraction provides a and explicit **intervention map** for generic causal models.
- For **linear** models, we have sound guarantees on both graphical and functional properties.

Conclusion

📜 **Recap:**

- **Causal Abstraction** enables concise representation of complex causal relations.
- τ-abstraction provides a and explicit **intervention map** for generic causal models.
- For **linear** models, we have sound guarantees on both graphical and functional properties.
- Applications exploit **abstract** causal properties to understand and interpret complex models.

Beckers, Sander and Joseph Y. Halpern (2019). **"Abstracting Causal Models".** In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. Vol. 33. AAAI'19/IAAI'19/EAAI'19. Honolulu, Hawaii, USA: AAAI Press, pp. 2678–2685. isbn: 978-1-57735-809-1. doi: `10.1609/aaai.v33i01.33012678`. url: `https://doi.org/10.1609/aaai.v33i01.33012678`.

Dyer, Joel et al. (2024). **"Interventionally Consistent Surrogates for Complex Simulation Models".** In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Geiger, Atticus, Hanson Lu, et al. (2021). **"Causal abstractions of neural networks".** In: *Advances in Neural Information Processing Systems* 34, pp. 9574–9586.

Geiger, Atticus, Zhengxuan Wu, et al. (2021). **“Inducing causal structure for interpretable neural networks”.** In: *arXiv preprint arXiv:2112.00826*.

Massidda, Riccardo, Atticus Geiger, et al. (2023). **“Causal abstraction with soft interventions”.** In: *Conference on Causal Learning and Reasoning*. PMLR, pp. 68–87.

Massidda, Riccardo, Sara Magliacane, and Davide Bacciu (2024). **“Learning Causal Abstractions of Linear Structural Causal Models”.** In: *The 40th Conference on Uncertainty in Artificial Intelligence*. url: `https://openreview.net/forum?id=XlFqI9TMhf`.

Shimizu, Shohei et al. (2006). **“A linear non-Gaussian acyclic model for causal discovery.”.** In: *Journal of Machine Learning Research* 7.10.

Wu, Zhengxuan et al. (2024). **“Interpretability at scale: Identifying causal mechanisms in alpaca”.** In: *Advances in Neural Information Processing Systems* 36.