Returning the Favour:
When Regression Benefits from Probabilistic Causal Knowledge

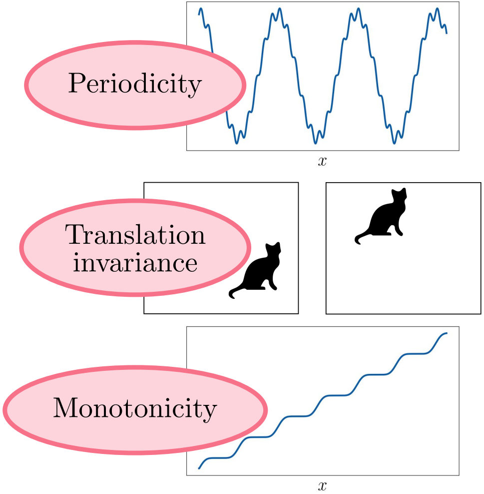
Shahine Bouabid[*1] Jake Fawkes[*1] Dino Sejdinovic[2]

* Equal contribution

[1] Department of Statistics, University of Oxford

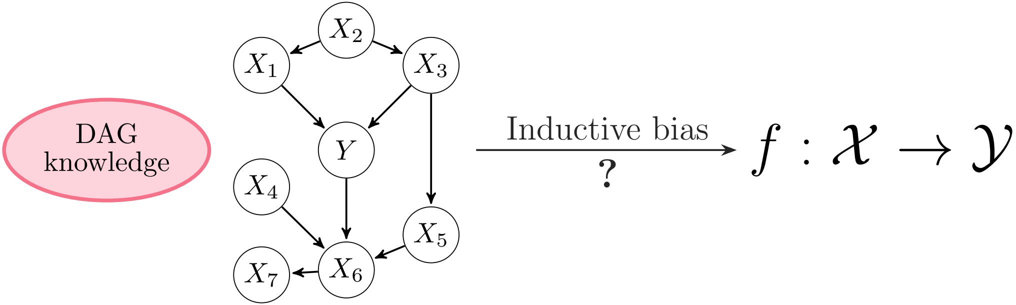
[2] School of CMS & AIML, University of Adelaide, Adelaide

Central Question

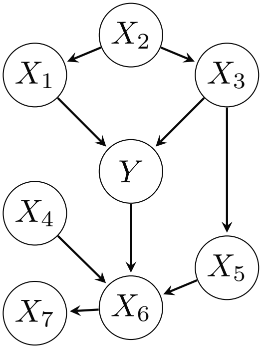


$$\xrightarrow{\text{Inductive bias}} f : \mathcal{X} \to \mathcal{Y}$$

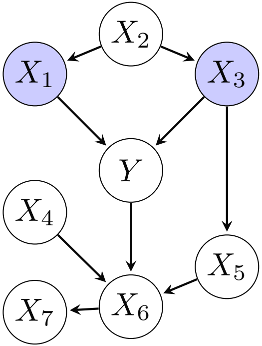
Central Question



What does a DAG tell us about $\mathbb{P}(Y|X)$?

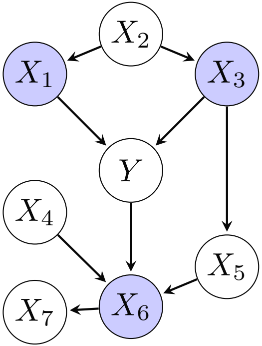


What does a DAG tell us about $\mathbb{P}(Y|X)$?



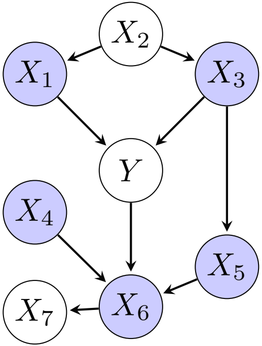
Parents of Y

What does a DAG tell us about $\mathbb{P}(Y|X)$?



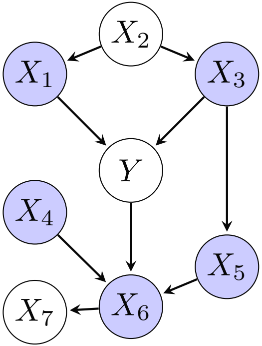
Parents of Y $\cup$ Children of Y

What does a DAG tell us about $\mathbb{P}(Y|X)$?



Parents of Y $\cup$ Children of Y $\cup$ Spouses of Y

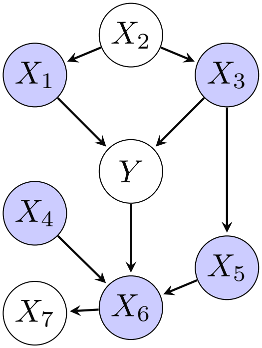
What does a DAG tell us about $\mathbb{P}(Y|X)$?



Parents of Y $\cup$ Children of Y $\cup$ Spouses of Y

= Markov Boundary of Y

What does a DAG tell us about $\mathbb{P}(Y|X)$?

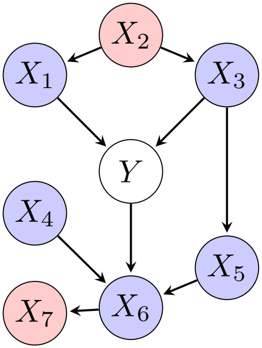


Parents of Y $\cup$ Children of Y $\cup$ Spouses of Y

= Markov Boundary of Y

- The Markov Boundary contains all relevant information for $\mathbb{P}(Y|X)$

What does a DAG tell us about $\mathbb{P}(Y|X)$?

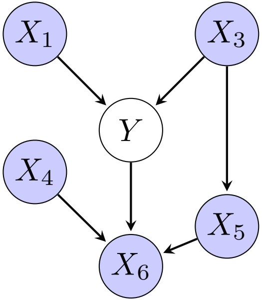


Parents of Y $\cup$ Children of Y $\cup$ Spouses of Y

$=$ Markov Boundary of Y

- The Markov Boundary contains all relevant information for $\mathbb{P}(Y|X)$
- X_2 and X_7 do not inform $\mathbb{P}(Y|X)$

What does a DAG tell us about $\mathbb{P}(Y|X)$?

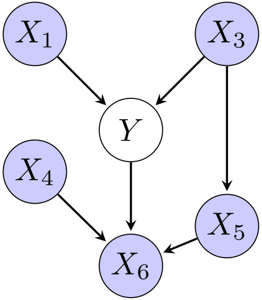


Parents of Y $\cup$ Children of Y $\cup$ Spouses of Y

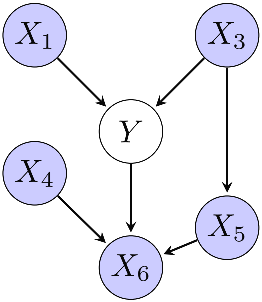
$=$ Markov Boundary of Y

- The Markov Boundary contains all relevant information for $\mathbb{P}(Y|X)$
- X_2 and X_7 do not inform $\mathbb{P}(Y|X)$

Further information in the Markov Boundary?

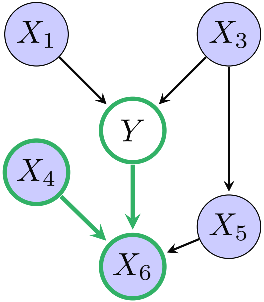


Further information in the Markov Boundary?



- Presence of **collider** X_6

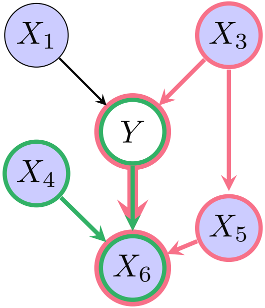
Further information in the Markov Boundary?



- Presence of **collider** X_6

$$Y \perp\!\!\!\perp X_4$$

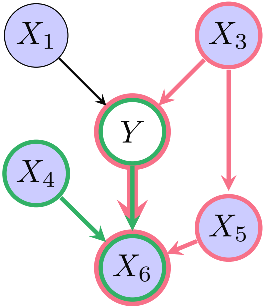
Further information in the Markov Boundary?



- Presence of **collider** X_6

$$Y \perp\!\!\!\perp X_4$$

$$Y \perp\!\!\!\perp X_5 \mid X_3$$



- Presence of **collider** X_6

$$Y \perp\!\!\!\perp X_4$$
$$Y \perp\!\!\!\perp X_5 \mid X_3$$

- Colliders provide additional information on $\mathbb{P}(Y|X)$ which is unused

Inductive biases from simple collider

Simple collider

$$Y \rightarrow X_1 \leftarrow X_2$$

Inductive biases from simple collider

Simple collider

$$Y \rightarrow X_1 \leftarrow X_2$$

$$\boxed{Y \perp\!\!\!\perp X_2, \qquad Y \not\!\perp\!\!\!\perp X_2 \mid X_1}$$

Inductive biases from simple collider

Simple collider

$$Y \longrightarrow X_1 \longleftarrow X_2$$

$$\boxed{Y \perp\!\!\!\perp X_2, \qquad Y \not\!\perp\!\!\!\perp X_2 \mid X_1}$$

Squared-loss regression problem

$$\boxed{\operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_{1i}, x_{2i})\right)^2 + \lambda \Omega(f)}$$

Simple collider

$Y \rightarrow X_1 \leftarrow X_2$

Simple collider

$Y \rightarrow X_1 \leftarrow X_2$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y | X_1 = x_1, X_2 = x_2]$

Simple collider

$Y \rightarrow X_1 \leftarrow X_2$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]$

$$\mathbb{E}[f^*(X_1, X_2)|X_2]$$

Simple collider

$Y \rightarrow X_1 \leftarrow X_2$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]$

$$\mathbb{E}[f^*(X_1, X_2)|X_2] = \mathbb{E}\left[\mathbb{E}[Y|X_1, X_2] \mid X_2\right]$$

Simple collider

$$Y \rightarrow X_1 \leftarrow X_2$$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y | X_1 = x_1, X_2 = x_2]$

$$\begin{aligned}
\mathbb{E}[f^*(X_1, X_2) | X_2] &= \mathbb{E}\left[\mathbb{E}[Y | X_1, X_2] \mid X_2\right] \\
&= \mathbb{E}[Y \mid X_2] && \text{(Tower property)}
\end{aligned}$$

Simple collider

$Y \rightarrow X_1 \leftarrow X_2$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y | X_1 = x_1, X_2 = x_2]$

$$
\begin{aligned}
\mathbb{E}[f^*(X_1, X_2) | X_2] &= \mathbb{E}\left[\mathbb{E}[Y | X_1, X_2] \mid X_2\right] \\
&= \mathbb{E}[Y \mid X_2] && \text{(Tower property)} \\
&= \mathbb{E}[Y] && (Y \perp\!\!\!\perp X_2)
\end{aligned}
$$

Simple collider

$$Y \rightarrow X_1 \leftarrow X_2$$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y|X_1 = x_1, X_2 = x_2]$

$$
\begin{aligned}
\mathbb{E}[f^*(X_1, X_2)|X_2] &= \mathbb{E}\left[\mathbb{E}[Y|X_1, X_2] \mid X_2\right] \\
&= \mathbb{E}[Y \mid X_2] && \text{(Tower property)} \\
&= \mathbb{E}[Y] && (Y \perp\!\!\!\perp X_2) \\
&= 0 && \text{(wlog)}.
\end{aligned}
$$

Simple collider

$Y \rightarrow X_1 \leftarrow X_2$

Optimal regressor: $f^*(x_1, x_2) = \mathbb{E}[Y | X_1 = x_1, X_2 = x_2]$

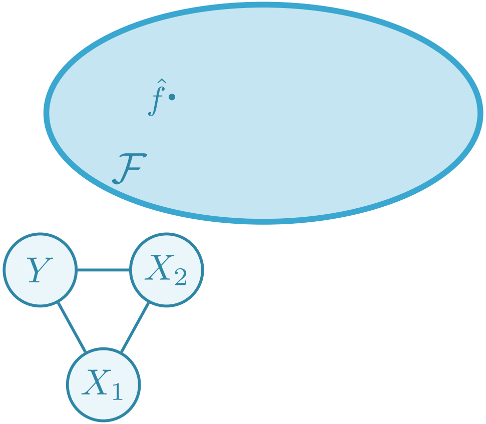
$$
\begin{aligned}
\mathbb{E}[f^*(X_1, X_2) | X_2] &= \mathbb{E}\left[\mathbb{E}[Y | X_1, X_2] \mid X_2\right] \\
&= \mathbb{E}[Y \mid X_2] && \text{(Tower property)} \\
&= \mathbb{E}[Y] && (Y \perp\!\!\!\perp X_2) \\
&= 0 && \text{(wlog).}
\end{aligned}
$$

Objective

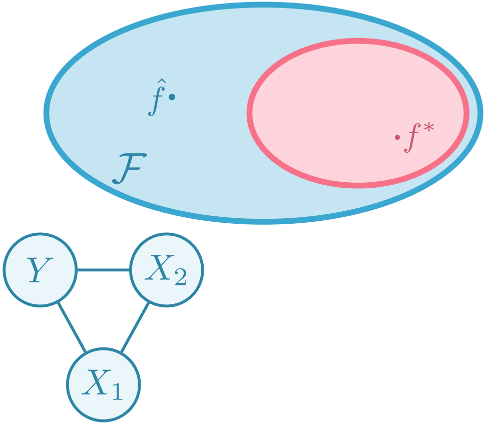
Find a regressor $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ in hypothesis space $\mathcal{F}$ that satisfies

$$
\boxed{\hat{f} \in \{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2) \mid X_2] = 0\}\,.}
$$

Constraining the hypothesis space $\mathcal{F}$

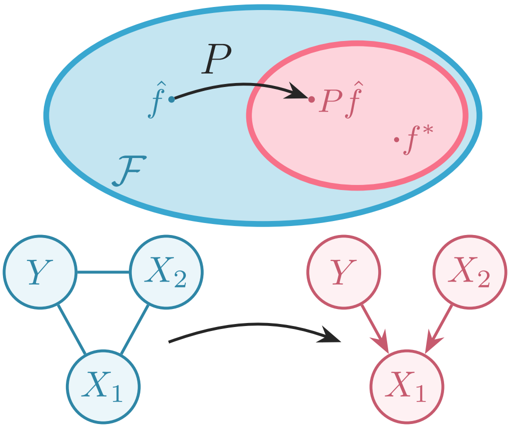


Constraining the hypothesis space $\mathcal{F}$



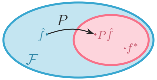
$$\{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2) \mid X_2] = 0\}$$

Constraining the hypothesis space $\mathcal{F}$

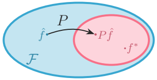


$$\left\{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2) \mid X_2] = 0\right\} = \mathrm{Range}(P)$$

Collider regression in $\mathcal{F} = L^2(X)$



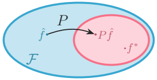
Collider regression in $\mathcal{F} = L^2(X)$



Take $P : L^2(X) \to L^2(X)$,

$$Pf(x_1, x_2) = f(x_1, x_2) - \mathbb{E}[f(X_1, X_2) | X_2 = x_2].$$

Collider regression in $\mathcal{F} = L^2(X)$

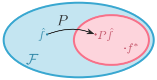


Take $P : L^2(X) \to L^2(X)$,

$$Pf(x_1, x_2) = f(x_1, x_2) - \mathbb{E}[f(X_1, X_2)|X_2 = x_2].$$

- P is an orthogonal projection

Collider regression in $\mathcal{F} = L^2(X)$

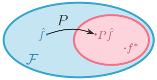


Take $P : L^2(X) \to L^2(X)$,

$$Pf(x_1, x_2) = f(x_1, x_2) - \mathbb{E}[f(X_1, X_2)|X_2 = x_2].$$

- P is an orthogonal projection
- $f^* \in \mathrm{Range}(P) = \{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2) \mid X_2] = 0\}$

Collider regression in $\mathcal{F} = L^2(X)$



Take $P : L^2(X) \to L^2(X)$,

$$Pf(x_1, x_2) = f(x_1, x_2) - \mathbb{E}[f(X_1, X_2)|X_2 = x_2].$$

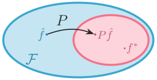
- P is an orthogonal projection
- $f^* \in \text{Range}(P) = \{f \in \mathcal{F} \mid \mathbb{E}[f(X_1, X_2) \mid X_2] = 0\}$

Proposition

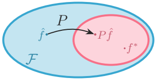
Let $h \in L^2(X)$, then we have

$$\Delta(h, Ph) = \mathbb{E}[(Y - h(X))^2] - \mathbb{E}[(Y - Ph(X))^2] = \|(\mathrm{Id} - P)h\|^2_{L^2(X)}.$$

Collider regression in a RKHS $\mathcal{F} = \mathcal{H}$

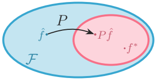


Collider regression in a RKHS $\mathcal{F} = \mathcal{H}$



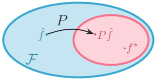
- Reproducing kernel Hilbert space with kernel $k(x, x')$

Collider regression in a RKHS $\mathcal{F} = \mathcal{H}$



- Reproducing kernel Hilbert space with kernel $k(x, x')$
- Convenient mathematical properties and dense in $L^2(X)$ under mild assumptions

Collider regression in a RKHS $\mathcal{F} = \mathcal{H}$



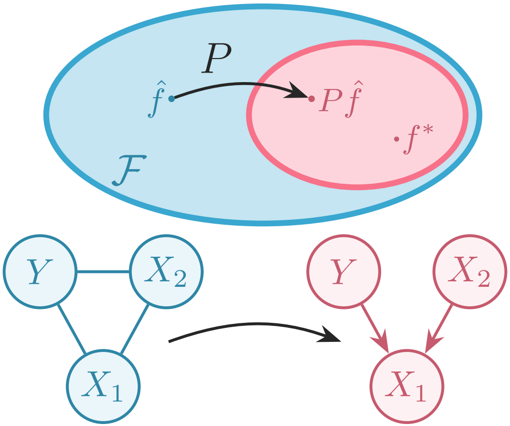
- Reproducing kernel Hilbert space with kernel $k(x, x')$
- Convenient mathematical properties and dense in $L^2(X)$ under mild assumptions

Theorem

Assume $M = \sup_{x \in \mathcal{X}} k(x, x) < \infty$ and $\mathrm{Var}(Y|X) \geq \eta > 0$. Then, the generalisation gap between $\hat{f}$ and $P\hat{f}$ satisfies

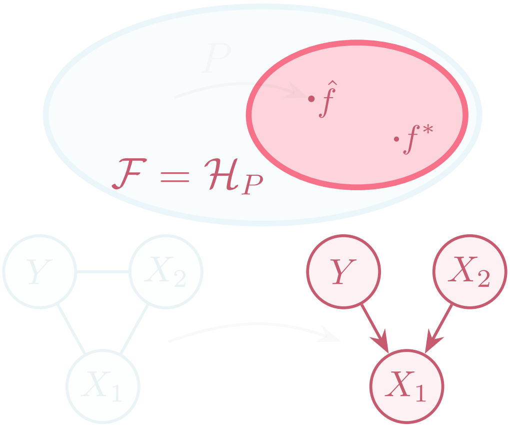
$$\mathbb{E}[\Delta(\hat{f}, P\hat{f})] \geq \frac{\eta \mathbb{E}\left[\|\mu_{X|X_2}(X)\|^2_{L^2(X)}\right]}{\left(\sqrt{n}M + \lambda/\sqrt{n}\right)^2} \tag{1}$$

where $\mu_{X|X_2} = \mathbb{E}[k(X, \cdot)|X_2]$ is a RKHS representation of $\mathbb{P}(X|X_2)$.



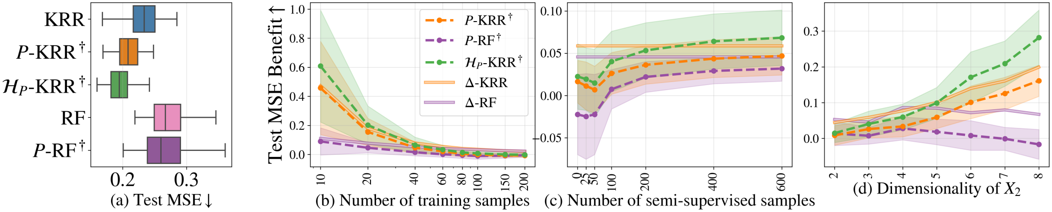
$\mathcal{F} = \mathcal{H}$ with kernel $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle$

Collider regression in a RKHS $\cancel{\mathcal{F} = \mathcal{H}}$ $\mathcal{F} = \mathcal{H}_P$



$\mathcal{F} = \mathcal{H}_P$ with kernel $k_P(x, x') = \langle P^* k(x, \cdot), P^* k(x', \cdot) \rangle$

Simulation experiment



(a) : Test MSEs for the simulation experiment ; dataset is generated using $d_1 = 3$, $d_2 = 3$, $n = 50$ and 100 semi-supervised samples ; experiments is run for 100 datasets generated with different seeds ; statistical significance is confirmed via Wilcoxon signed-rank ; (b, c, d) : Ablation study on the number of training samples, number of semi-supervised samples and dimensionality of X_2.

Table 1: MSE, signal-to-noise ratio (SNR) and correlation on test data for the aerosol radiative forcing experiment ; $n = 50$ and 200 semi-supervised samples ; statistical significance is confirmed via Wilcoxon signed-rank ; experiments are run for 100 datasets generated by FaIR with different seeds ; $\uparrow/\downarrow$ indicates higher/lower is better; we report 1 standard deviation; † indicates our proposed methods.

	MSE $\downarrow$	SNR $\uparrow$	Correlation $\uparrow$
RF	$0.90_{\pm 0.04}$	$0.44_{\pm 0.19}$	$0.32_{\pm 0.08}$
P-RF†	$\mathbf{0.89}_{\pm \mathbf{0.03}}$	$\mathbf{0.49}_{\pm \mathbf{0.15}}$	$\mathbf{0.34}_{\pm \mathbf{0.07}}$
KRR	$0.88_{\pm 0.04}$	$0.56_{\pm 0.21}$	$0.37_{\pm 0.05}$
P-KRR†	$\mathbf{0.86}_{\pm \mathbf{0.03}}$	$\mathbf{0.65}_{\pm \mathbf{0.14}}$	$\mathbf{0.40}_{\pm \mathbf{0.02}}$
$\mathcal{H}_P$-KRR†	$\mathbf{0.86}_{\pm \mathbf{0.03}}$	$\mathbf{0.64}_{\pm \mathbf{0.14}}$	$0.39_{\pm 0.03}$

Outlooks

- Collider structures within causal graphs consitute a useful form of inductive bias within supervised learning

To find out more come to our poster:
Poster Session 4, Wednesday 2:00-3:30, Exhibit hall 1.

Outlooks

- Collider structures within causal graphs consitute a useful form of inductive bias within supervised learning
- Provable benefit of semi-supervised learning arising from causal structure

To find out more come to our poster:
Poster Session 4, Wednesday 2:00-3:30, Exhibit hall 1.

Outlooks

- Collider structures within causal graphs consitute a useful form of inductive bias within supervised learning
- Provable benefit of semi-supervised learning arising from causal structure
- Collider regression for more general DAGs in the paper

To find out more come to our poster:
Poster Session 4, Wednesday 2:00-3:30, Exhibit hall 1.

Outlooks

- Collider structures within causal graphs consitute a useful form of inductive bias within supervised learning
- Provable benefit of semi-supervised learning arising from causal structure
- Collider regression for more general DAGs in the paper
- Similar reasoning can be applied to other forms of inference about $\mathbb{P}(Y|X)$, e.g. classification or quantile regression

To find out more come to our poster:
Poster Session 4, Wednesday 2:00-3:30, Exhibit hall 1.