Causal decision-making meets Gaussian Processes

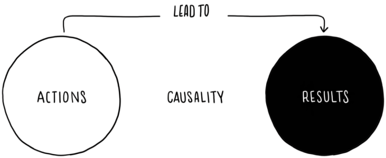
Virginia Aglietti

University of Warwick, The Alan Turing Institute

Causal decision-making

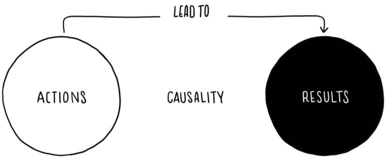
Integrate causal considerations into a choice process and take decisions based on causal knowledge [Hagmayer and Fernbach (2017)].

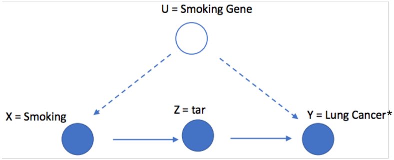
Why is this important?



Integrate causal considerations into a choice process and take decisions based on causal knowledge [Hagmayer and Fernbach (2017)].

Why is this important?





*Pearl and Mackenzie (2018).

Systems/processes decompose in sets of interconnected nodes

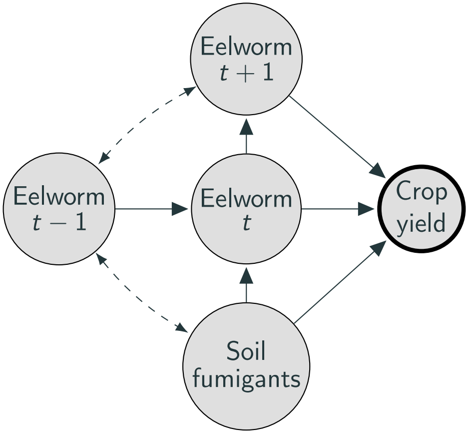


Figure 1: Causal Graph for Crop Yield. Nodes denote variables, arrows represent causal effects and dashed edges indicate unobserved confounders.

Systems/processes decompose in sets of interconnected nodes

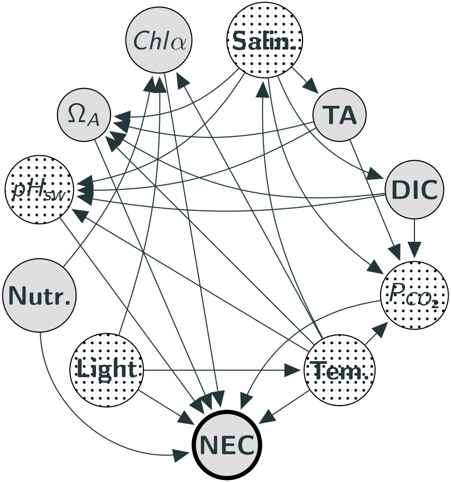


Figure 2: Causal Graph for Net Ecosystem Calcification (NEC). Dotted nodes represent non-manipulative variables.

Systems/processes decompose in sets of interconnected nodes

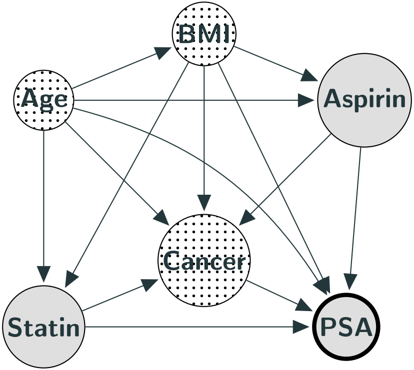


Figure 3: Causal Graph for Prostate Specific Antigen (PSA) level. Dotted nodes represent non-manipulative variables.

Common elements in these examples

- A causal graph (Directed Acyclic Graph - DAG).
- Observational data from all (non hidden) nodes.
- Ability of running experiments (in reality or in simulation).
- Cost of experiments depends on the number and type of nodes in which we intervene.

Research goal

Efficiently find the *system configuration* that optimises the *target node*.

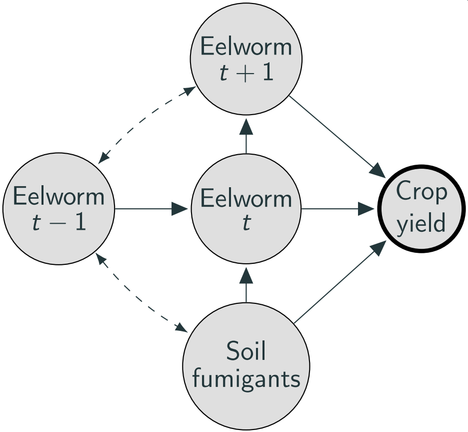
Research Goal

Research goal

Efficiently find the *system configuration* that optimises the *target node*.

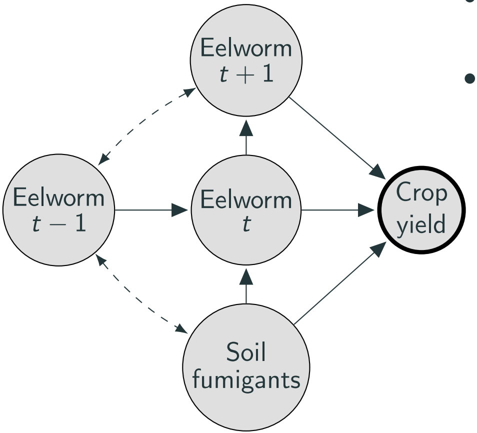
- *System configuration* $\rightarrow$ manipulative variables to be intervened on and their intervention levels e.g. Soil fumigants, CO_2, Statin.
- *Target node* $\rightarrow$ variable in the causal graph that we wish to optimize considering its causal relationships e.g. Crop yield, NEC, PSA.
- *Efficiently* $\rightarrow$ exploit all information, observational and interventional, that we collect when exploring the system.

Questions to answer



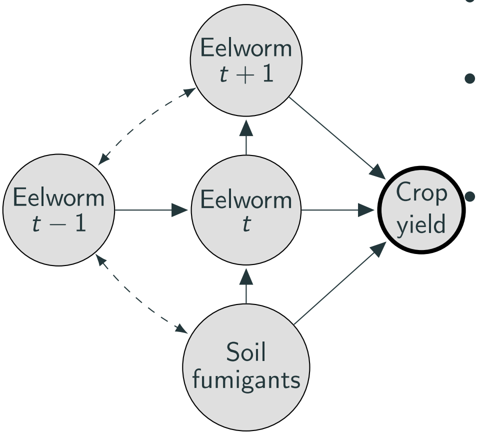
- How can I learn the expected crop yield given different interventions?

Questions to answer



- How can I learn the expected crop yield given different interventions?
- How can I learn the optimal intervention that is the intervention maximizing the the crop yield?

Questions to answer



- How can I learn the expected crop yield given different interventions?
- How can I learn the optimal intervention that is the intervention maximizing the the crop yield?
- I have observed the crop yield over seasons and have previously intervened on soil fumigants. How can I integrate this information to infer the crop yield I would get by intervening on soil fumigants and eelworm t?

Research Goal

Research goal

Efficiently find the *system configuration* that optimises the *target node*.

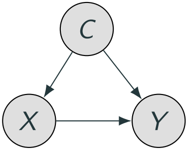
1. Perform experiments
2. Integrate interventional and observational data
3. Transfer interventional information.

Causal models and *do*-calculus

Causal models and *do*-calculus (1 of 3)

Causal model: DAG $\mathcal{G}$ + four-tuple $\langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) \rangle$

- $\mathbf{U}$: independent *exogenous* background variables.
- $P(\mathbf{U})$ distribution of $\mathbf{U}$.
- $\mathbf{V}$: *endogenous* variables (non-manipulative, manipulative, target).
- $F = \{f_1, ..., f_{|\mathbf{V}|}\}$: functions $v_i = f_i(pa_i, u_i)$, pa_i are the parents of V_i.

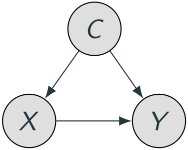


$$
\begin{aligned}
C &= f_c(U_c),\ U_c \sim \mathcal{N}(0, \sigma_c^2) \\
X &= f_x(C, U_x),\ U_x \sim \mathcal{N}(0, \sigma_x^2) \\
Y &= f_y(X, C, U_y),\ Y_c \sim \mathcal{N}(0, \sigma_y^2)
\end{aligned}
$$

Causal models and *do*-calculus (2 of 3)

Intervention: Setting a manipulative variable X to a value x, $do(X = x)$.

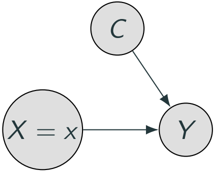
Observed universe



$$\begin{aligned} C &= f_c(U_c) \\ X &= f_x(C, U_x) \\ Y &= f_y(X, C, U_y) \end{aligned}$$

$$P(X, C, Y)$$

Post-intervention universe



$$\begin{aligned} C &= f_c(U_c) \\ X &= x \\ Y &= f_y(x, C, U_y) \end{aligned}$$

$$P^{do(X=x)}(C, Y)$$

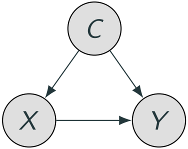
$$\boxed{P(Y|do(X = x)) := P^{do(X=x)}(Y|X = x)}$$

Causal models and *do*-calculus (3 of 3)

Key question: *How to do inference in the post-intervention universe?*

- Intervene $\rightarrow$ Interventional data $\rightarrow$ $P(Y|\mathrm{do}\,(X=x))$
- Observe $\rightarrow$ Observational data $\rightarrow$ *do*-calculus $\rightarrow$ $\hat{P}(Y|\mathrm{do}\,(X=x))$

do-calculus: algebra to emulate the post-intervention universe in terms of conditionals $P(Y|X=x)$ in the observed universe.



Back-door adjustment: $p(Y|do(X=x)) = \int P(Y|c, X=x)P(c)dc$.

Causal Optimization

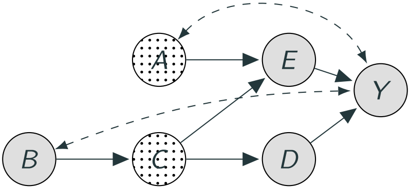
$$\boxed{\mathbf{X}_s^\star, \mathbf{x}_s^\star = \underset{\substack{\mathbf{X}_s \in \mathcal{P}(\mathbf{X}) \\ \mathbf{x}_s \in D(\mathbf{X}_s)}}{\arg\min}\, \mathbb{E}[Y | \mathrm{do}\,(\mathbf{X}_s = \mathbf{x}_s)]} \tag{1}$$

- $\mathcal{P}(\mathbf{X})$ gives all possible interventions.
- $D(\mathbf{X}_s)$ fixed interventional domain for each $\mathbf{X}_s$.
- $\mathbf{X}_s$, $\mathbf{x}_s$ one possible intervention set and value.
- $\mathbf{X}_s^\star$, $\mathbf{x}_s^\star$, optimal intervention set and value.

Causal optimization

$$\boxed{\mathbf{X}_s^\star, \mathbf{x}_s^\star = \operatorname*{arg\,min}_{\substack{\mathbf{X}_s \in \mathcal{P}(\mathbf{X}) \\ \mathbf{x}_s \in D(\mathbf{X}_s)}} \mathbb{E}[Y | \mathrm{do}\,(\mathbf{X}_s = \mathbf{x}_s)]}$$

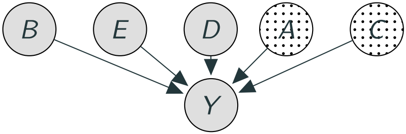
- Explore $\mathcal{P}(\mathbf{X})$
- Find the intervention set and the intervention level



Global optimization

$$\boxed{\mathbf{x}^\star = \operatorname*{arg\,min}_{\mathbf{x} \in D(\mathbf{X})} \mathbb{E}[Y | \mathrm{do}\,(\mathbf{X} = \mathbf{x})]}$$

- Set the intervention set to $\mathbf{X}$
- Find the intervention level



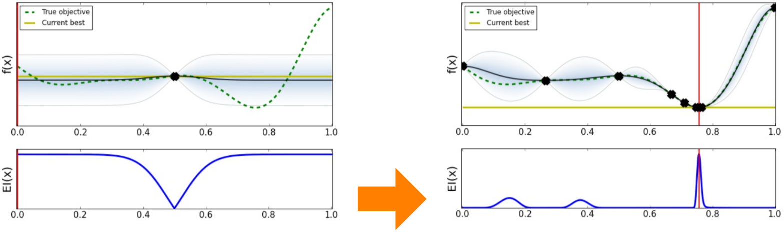
Solving Global Optimization

- Target function f is explicitly unknown and multimodal.
- Evaluations of f are perturbed by noise.
- Evaluations of f are expensive.

$\Rightarrow$ Bayesian Optimization

Bayesian optimization

- **Goal**: Collect data $x_1, \ldots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$



Each point x_{n+1} is collected as $x_{n+1} = arg\max \alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}_n)$

Solving Causal Optimization

- Target function f is explicitly unknown and multimodal.
- Evaluations of f are perturbed by noise.
- Evaluations of f are expensive.

$+$

- **Causal graph**

$\Rightarrow$ $\boxed{\text{Causal Bayesian Optimization (CBO)}}$

Idea: Run interventions $(\mathbf{X}_{s_1}, \mathbf{x}_{s_1}), \ldots, (\mathbf{X}_{s_n}, \mathbf{x}_{s_n})$ to find the optimum as fast as possible.

Do we need to explore all $2^{|\mathbf{X}|}$ sets in $\mathcal{P}(\mathbf{X})$? NO!

$$\boxed{\mathbf{X}_s^\star, \mathbf{x}_s^\star = \underset{\substack{\mathbf{X}_s \in \mathcal{P}(\mathbf{X}) \\ \mathbf{x}_s \in D(\mathbf{X}_s)}}{\arg\min}\, \mathbb{E}[Y | \mathrm{do}\,(\mathbf{X}_s = \mathbf{x}_s)]}$$

Do we need to explore all $2^{|\mathbf{X}|}$ sets in $\mathcal{P}(\mathbf{X})$? NO!

$$\boxed{\mathbf{X}_s^\star, \mathbf{x}_s^\star = \mathop{\arg\min}_{\substack{\mathbf{X}_s \in \mathcal{P}(\mathbf{X}) \\ \mathbf{x}_s \in D(\mathbf{X}_s)}} \mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}_s = \mathbf{x}_s)]}$$

Minimal Intervention Set (MIS, $\mathbb{M}^{\mathbf{C}}_{\mathcal{G},Y}$)

Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X}, \mathbf{C} \rangle$, a set $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ is said to be a MIS if there is no $\mathbf{X}_s' \subset \mathbf{X}_s$ such that $\mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}_s = \mathbf{x}_s), \mathbf{C}] = \mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}_s' = \mathbf{x}_s'), \mathbf{C}]$.

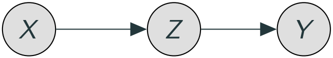
Possibly-Optimal Minimal Intervention set (POMIS, $\mathbb{P}^{\mathbf{C}}_{\mathcal{G},Y}$)

Let $\mathbf{X}_s \in \mathbb{M}^{\mathbf{C}}_{\mathcal{G},Y}$. $\mathbf{X}_s$ is a POMIS if there exists a sem conforming to $\mathcal{G}$ such that $\mathbb{E}[Y \mid \mathrm{do}(\mathbf{X}_s = \mathbf{x}_s^*), \mathbf{C}] > \forall_{\mathbf{W} \in \mathbb{M}^{\mathbf{C}}_{\mathcal{G},Y} \setminus \mathbf{X}_s} \mathbb{E}[Y \mid \mathrm{do}(\mathbf{W} = \mathbf{w}^*), \mathbf{C}]$ where $\mathbf{x}^*$ and $\mathbf{w}^*$ denote the optimal intervention values.

MIS and POMIS are sets of variables 'worth' intervening on.

Causal Bayesian Optimization

Toy example

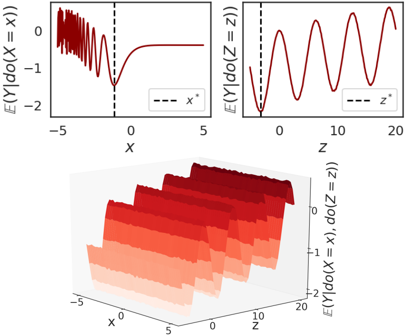


$$X = \epsilon_X$$

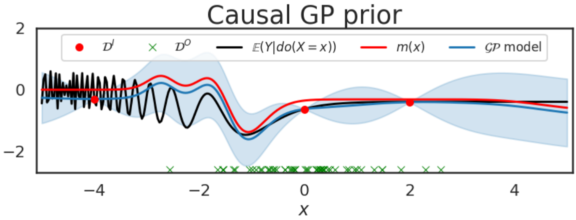
$$Z = \exp(-X) + \epsilon_Z$$

$$Y = \cos(Z) - \exp(-\frac{Z}{20}) + \epsilon_Y$$

$$\boxed{\begin{array}{l} \mathbb{M}_{\mathcal{G},Y} = \{\emptyset, \{X\}, \{Z\}\} \\ \mathbb{P}_{\mathcal{G},Y} = \{\{Z\}\} \\ \mathbb{B}_{\mathcal{G},Y} = \{\{X, Z\}\} \end{array}}$$



Modelling $\mathbb{E}[Y|do(X_s = x_s)]$ for each X_s with Causal GP prior



$$f(\mathbf{x}_s) \sim \mathcal{GP}(m(\mathbf{x}_s), k(\mathbf{x}_s, \mathbf{x}_s'))$$

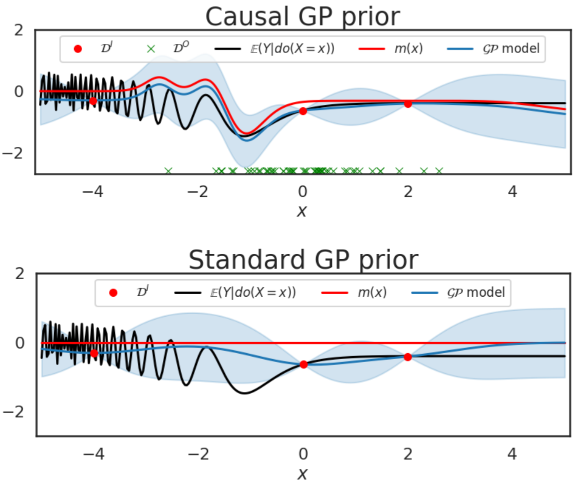
$$m(\mathbf{x}_s) = \hat{\mathbb{E}}[Y|\text{do}\,(\mathbf{X}_s = \mathbf{x}_s)]$$

$$k(\mathbf{x}_s, \mathbf{x}_s') = k_{RBF}(\mathbf{x}_s, \mathbf{x}_s') + \sigma(\mathbf{x}_s)\sigma(\mathbf{x}_s')$$

where

- $k_{RBF}(\mathbf{x}_s, \mathbf{x}_s') := \exp(-\frac{||\mathbf{x}_s - \mathbf{x}_s'||^2}{2l^2})$
- $\sigma(\mathbf{x}_s) = \sqrt{\hat{\mathbb{V}}(Y|\text{do}\,(\mathbf{X}_s = \mathbf{x}_s))}$ with $\hat{\mathbb{V}}$ is the variance of the causal effects estimated from observational data.

Modelling $\mathbb{E}[Y|do(X_s = x_s)]$ for each X_s with Causal GP prior



Standard GP prior: $f(\mathbf{x}_s) \sim \mathcal{GP}(m(\mathbf{x}_s), k(\mathbf{x}_s, \mathbf{x}'_s))$, $m(\mathbf{x}_s) = \mathbf{0}$ and $k(\mathbf{x}_s, \mathbf{x}'_s) = k_{RBF}(\mathbf{x}_s, \mathbf{x}'_s)$ with $k_{RBF}(\mathbf{x}_s, \mathbf{x}'_s) := \exp(-\frac{||\mathbf{x}_s - \mathbf{x}'_s||^2}{2l^2})$.

Solving the exploration-exploitation trade-off

Causal Expected Improvement

- $y_s = \mathbb{E}[Y|\mathrm{do}\,(\mathbf{X}_s = \mathbf{x}_s)]$
- $y^\star = \max_{\mathbf{X}_s \in \mathbf{es}, \mathbf{x} \in D(\mathbf{X}_s)} \mathbb{E}[Y|\mathrm{do}\,(\mathbf{X}_s = \mathbf{x}_s)]$

$$EI^s(\mathbf{x}) = \mathbb{E}_{p(y_s)}[\max(y_s - y^\star, 0)]/Co(\mathbf{X}_s, \mathbf{x}_s)$$

- $\alpha_1, \ldots, \alpha_{|\mathbf{es}|}$: solutions of optimizing $EI^s(\mathbf{x})$ for each set in **es** and

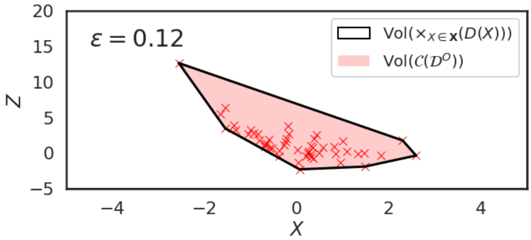
New intervention set and value

$$\alpha^\star := \max\{\alpha_1, \ldots, \alpha_{|\mathbf{es}|}\}$$

$$s^\star = \underset{s \in \{1, \cdots, |\mathbf{es}|\}}{\operatorname{argmax}} \alpha_s$$

Solving the intervention-observation trade-off

ϵ-greedy criteria



$$\epsilon = \frac{\mathrm{Vol}(\mathcal{C}(\mathcal{D}^O))}{\mathrm{Vol}(\times_{X\in\mathbf{X}}(D(X)))} \times \frac{N}{N_{\mathrm{max}}},$$

- $\mathrm{Vol}(\mathcal{C}(\mathcal{D}^O))$: volume of the convex hull for observational data.
- $\mathrm{Vol}(\times_{X\in\mathbf{X}}(D(X)))$: volume of the interventional domain.

Causal Bayesian Optimization - CBO

Algorithm: Causal Bayesian Optimization

Data: $\mathcal{D}^{\mathrm{O}}$, $\mathcal{D}^{\mathrm{I}}$, $\mathcal{G}$, **es**, number of steps T

Result: $\mathbf{X}_s^\star, \mathbf{x}_s^\star, \hat{\mathbb{E}}[\mathbf{Y}^\star|\mathrm{do}\,(\mathbf{X}_s^* = \mathbf{x}_s^\star)]$

Initialise: Set $\mathcal{D}_0^{\mathrm{I}} = \mathcal{D}^{\mathrm{I}}$ and $\mathcal{D}_0^{\mathrm{O}} = \mathcal{D}^{\mathrm{O}}$

for *t=1, ..., T* **do**

  Compute ϵ and sample $u \sim \mathcal{U}(0,1)$

  if $\epsilon > u$ **then**

    (Observe)

    1. Observe new observations $(\mathbf{x}_t, c_t, \mathbf{y}_t)$.
    2. Augment $\mathcal{D}^{\mathrm{O}} = \mathcal{D}^{\mathrm{O}} \cup \{(\mathbf{x}_t, c_t, \mathbf{y}_t,)\}$.
    3. Update prior of the causal GP.

  end

  else

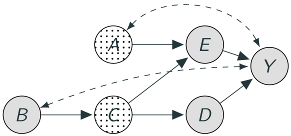
    (Intervene)

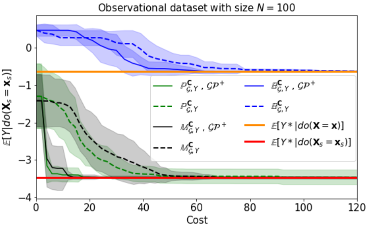
    1. Compute $EI^s(\mathbf{x})$ for each element $s \in$ **es**.
    2. Obtain the optimal interventional set-value pair $(s^\star, \alpha^\star)$.
    3. Intervene on the system.
    4. Update posterior of the interventional GP.

  end

end

Simulation analysis





- Results are consistent with what is expected.
- Better results that BO: propagation of effect beyond default domain.

Take home messages:

1. Many real systems decompose in interconnected nodes.

2. Optimization requires ‘intervening’ in the manipulative nodes and solving a Causal Optimization problem.

3. Standard Bayesian Optimization ignores causal assumptions.

4. CBO solves Causal Optimization problems and improves BO when causal information is available.

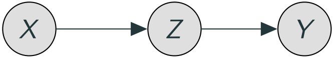
5. CBO efficiently explores ‘worthy’ interventions.

6. Causal GPs prior merges observational and interventional data.

Limitations of CBO

- The number of GPs we require is determined by $|\mathcal{P}(\mathbf{X})|$ which is potentially huge.
- We don't transfer interventional information across GPs e.g. we don't account for the fact that intervening on $\mathbf{X}$ might give us some information about an intervention on $\mathbf{X}$ and $\mathbf{Z}$.

Limitation of CBO



$$\boxed{\mathbb{M}_{\mathcal{G},Y} = \{\varnothing, \{X\}, \{Z\}\}}$$

- $f_X(x)$
- $f_Z(z)$
- $f_{X,Z}(x, z)$

Limitation of CBO

- The number of GPs we require is determined by $|\mathcal{P}(\mathbf{X})|$ which is potentially huge.
- We don't transfer interventional information across GPs e.g. we don't account for the fact that intervening on $\mathbf{X}$ might give us some information about an intervention on $\mathbf{X}$ and $\mathbf{Z}$.

Research goal

Efficiently find the *system configuration* that optimises the *target node*.

$\Rightarrow$ $\boxed{\textbf{Methodology to transfer interventional information.}}$

Research goal

We aim at learning the set of ***intervention functions*** for Y in $\mathcal{G}$:

$$\mathbf{T} = \{t_s(\mathbf{x})\}_{s=1}^{|\mathcal{P}(\mathbf{X})|} \qquad t_s(\mathbf{x}) = f_s(\mathbf{x}) = \mathbb{E}_{p(Y|\mathrm{do}(\mathbf{X}_s=\mathbf{x}))}[Y] = \mathbb{E}[Y|\mathrm{do}\,(\mathbf{X}_s = \mathbf{x})].$$

given $\mathcal{D}^O = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$ and $\mathcal{D}^I = (\mathbf{X}^I, \mathbf{Y}^I)$ with $\mathbf{X}^I = \bigcup_s \{\mathbf{x}_{si}^I\}_{i=1}^{N_s^I}$ and $\mathbf{Y}^I = \bigcup_s \{y_{si}^I\}_{i=1}^{N_s^I}$.

$$\Downarrow$$

Goal

Define $p(\mathbf{T})$ and compute $p(\mathbf{T}|\mathcal{D}^I)$ so as to make probabilistic predictions for $\mathbf{T}$ at some unobserved intervention sets and levels.

Steps in the methodology

1. Study the correlation among functions in $\mathbf{T} = \{f_s(\mathbf{x})\}_{s=1}^{|\mathcal{P}(\mathbf{X})|}$ which varies with the topology of $\mathcal{G}$.
2. Define a joint prior distribution $p(\mathbf{T})$.
3. Develop a multi-task model based on $p(\mathbf{T})$ so as to compute the posterior $p(\mathbf{T}|\mathcal{D}^I)$.

1. Study the correlation among intervention functions

Any function in $\mathbf{T}$ can be written as an integral transformation of some *base function* f, via some *causal operator* L_s such that $t_s(\mathbf{x}) = L_s(f)(\mathbf{x}), \forall \mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ (Theorem 3.1):

$$L_s(f)(\mathbf{x}) = \int \cdots \int \pi_s(\mathbf{x}, (\mathbf{v}_s^N, \mathbf{c})) f(\mathbf{v}, \mathbf{c}) \mathrm{d}\mathbf{v}_s^N \mathrm{d}\mathbf{c},$$

1. Study the correlation among intervention functions

Any function in $\mathbf{T}$ can be written as an integral transformation of some *base function* f, via some *causal operator* L_s such that $t_s(\mathbf{x}) = L_s(f)(\mathbf{x}), \forall \mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ (Theorem 3.1):

$$L_s(f)(\mathbf{x}) = \int \cdots \int \pi_s(\mathbf{x}, (\mathbf{v}_s^N, \mathbf{c}))f(\mathbf{v}, \mathbf{c})\mathrm{d}\mathbf{v}_s^N \mathrm{d}\mathbf{c},$$

- $f(\mathbf{v}, \mathbf{c}) = \mathbb{E}\left[Y|\mathrm{do}\left(\mathbf{I} = \mathbf{v}\right), \mathbf{C}^N = \mathbf{c}\right]$
- $\mathbf{C}^N$ is the set of variables in $\mathcal{G}$ that are *directly* confounded with Y and are not colliders.
- $\mathbf{I}$ is the set $\mathrm{Pa}(Y)$.
- $\pi_s(\mathbf{x}, (\mathbf{v}_s^N, \mathbf{c})) = p(\mathbf{c}_s^I|\mathbf{c}_s^N)p(\mathbf{v}_s^N, \mathbf{c}_s^N|\mathrm{do}\left(\mathbf{X}_s = \mathbf{x}\right))$ is the integrating measure for the set $\mathbf{X}_s$.

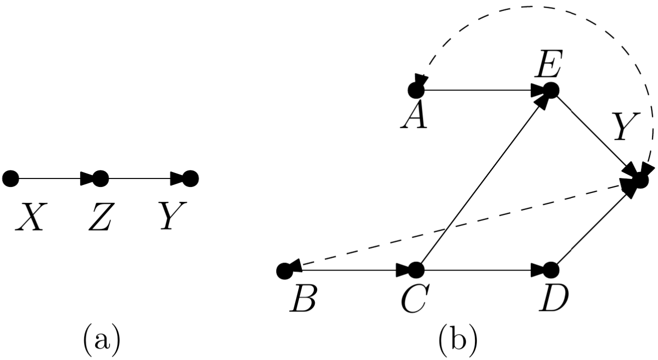
1. Study the correlation among intervention functions

$$\begin{aligned}&\mathbf{T} = \{t_X(x), t_Z(z), t_{X,Z}(x,z)\}\\&\mathbf{I} = \{Z\}\\&\mathbf{C} = \varnothing\end{aligned}$$

$$f(z) = \mathbb{E}[Y|\mathrm{do}\,(Z = z)]$$

$$t_X(x) = \int f(z)p(z|X = x)dz$$

1. Study the correlation among intervention functions



- (a) $\mathbf{I} = \{Z\}$, $\mathbf{C}^N = \emptyset$
- (b) $\mathbf{I} = \{E, D\}$, $\mathbf{C}^N = \{A, B\}$

2. Define a joint prior $p(\mathbf{T})$

$$t_s(\mathbf{x}) = L_s(f)(\mathbf{x}), \forall \mathbf{X}_s \in \mathcal{P}(\mathbf{X})$$

We can define a prior for $\mathbf{T}$ by:

- Step (1): placing a *causal* prior on f
- Step (2): propagating this prior through $L_s(\cdot)$ for all $t_s(\mathbf{x})$

2. Define a joint prior $p(\mathsf{T})$

- Step (1): Placing a *causal* prior on $f(\mathbf{v}, \mathbf{c}) = f(\mathbf{b})$

$$f(\mathbf{b}) \sim \mathcal{GP}(m(\mathbf{b}), K(\mathbf{b}, \mathbf{b}'))$$

with $m(\mathbf{b}) = \hat{f}(\mathbf{b})$ and $K(\mathbf{b}, \mathbf{b}') = k_{\text{rbf}}(\mathbf{b}, \mathbf{b}') + \hat{\sigma}(\mathbf{b})\hat{\sigma}(\mathbf{b}')$ where

$$\hat{f}(\mathbf{b}) = \hat{f}(\mathbf{v}, \mathbf{c}) = \hat{\mathbb{E}}[Y|\text{do}\left(\mathbf{I} = \mathbf{v}\right), \mathbf{c}]$$
$$\hat{\sigma}(\mathbf{b}) = \hat{\sigma}(\mathbf{v}, \mathbf{c}) = \hat{\mathbb{V}}[Y|\text{do}\left(\mathbf{I} = \mathbf{v}\right), \mathbf{c}]^{1/2}$$

where $\hat{\mathbb{V}}$ and $\hat{\mathbb{E}}$ represent the variance and expectation of the causal effects estimated from $\mathcal{D}^O$ using *do-calculus*.

2. Define a joint prior $p(\mathsf{T})$

- Step (2): propagating this prior through $L_s(\cdot)$ for all $t_s(\mathbf{x})$

For each $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$, we have $t_s(\mathbf{x}) \sim \mathcal{GP}(m_s(\mathbf{x}), k_s(\mathbf{x}, \mathbf{x}'))$ with:

$$m_s(\mathbf{x}) = \int \cdots \int m(\mathbf{b}) \pi_s(\mathbf{x}, \mathbf{b}_s) \, \mathrm{d}\mathbf{b}_s$$

$$k_s(\mathbf{x}, \mathbf{x}') = \int \cdots \int K(\mathbf{b}, \mathbf{b}') \pi_s(\mathbf{x}, \mathbf{b}_s) \, \pi_s(\mathbf{x}', \mathbf{b}'_s) \, \mathrm{d}\mathbf{b}_s \mathrm{d}\mathbf{b}'_s.$$

where $\mathbf{b}_s = (\mathbf{v}_s^N, \mathbf{c})$ is the subset of $\mathbf{b}$ including only the $\mathbf{v}$ values corresponding to the set $\mathbf{I}_s^N$.

The DAG-GP model

Joint prior distribution: $T^D \sim \mathcal{N}(m_{\mathbf{T}}(D), K_{\mathbf{T}}(D, D))$

Likelihood function: $p(\mathbf{Y}^I | \mathbf{T}^I, \sigma^2) = \mathcal{N}(\mathbf{T}^I, \sigma^2 \mathbf{I})$.

Joint posterior distribution: $\mathbf{T}^D | \mathcal{D}^I \sim \mathcal{N}(m_{\mathbf{T}|\mathcal{D}^I}(D), K_{\mathbf{T}|\mathcal{D}^I}(D, D))$

with $m_{\mathbf{T}|\mathcal{D}^I}(D) = m_{\mathbf{T}}(D) + K_{\mathbf{T}}(D, \mathbf{X}^I)[K_{\mathbf{T}}(\mathbf{X}^I, \mathbf{X}^I) + \sigma^2 \mathbf{I}](\mathbf{T}^I - m_{\mathbf{T}}(\mathbf{X}^I))$
and $K_{\mathbf{T}|\mathcal{D}^I}(D, D) = K_{\mathbf{T}}(D, D) - K_{\mathbf{T}}(D, \mathbf{X}^I)[K_{\mathbf{T}}(\mathbf{X}^I, \mathbf{X}^I) + \sigma^2 \mathbf{I}] K_{\mathbf{T}}(\mathbf{X}^I, D)$.

	Interventional data: No	Interventional data: Yes	
		Single-task	Multi-task
Observational data: No	Mechanistic model	**GP** $p(T) = \prod_s p(t_s(\mathbf{x}))$ $t_s(\mathbf{x}) \sim \mathcal{GP}(0, K_{RBF}(\mathbf{x}, \mathbf{x}'))$	**DAG-GP** $p(\mathbf{T}) = \prod_s p(t_s(\mathbf{x})\|f)$ $t_s(\mathbf{x}) = \int f(\mathbf{b})\pi_s(\mathbf{x}, \mathbf{b}_s)\mathrm{d}\mathbf{b}_s$ $f(\mathbf{b}) \sim \mathcal{GP}(0, K_{RBF}(\mathbf{b}, \mathbf{b}'))$
Observational data: Yes	*do*-calculus	**GP$^+$** $p(T) = \prod_s p(t_s(\mathbf{x}))$ $t_s(\mathbf{x}) \sim \mathcal{GP}(m^+(\mathbf{x}), K^+(\mathbf{x}, \mathbf{x}'))$	**DAG-GP$^+$** $p(\mathbf{T}) = \prod_s p(t_s(\mathbf{x})\|f)$ $t_s(\mathbf{x}) = \int f(\mathbf{b})\pi_s(\mathbf{x}, \mathbf{b}_s)\mathrm{d}\mathbf{b}_s$ $f(\mathbf{b}) \sim \mathcal{GP}(m^+(\mathbf{b}), K^+(\mathbf{b}, \mathbf{b}'))$

Figure 6: Models for learning the intervention functions $\mathbf{T}$ defined on a dag.

Learning intervention function from data

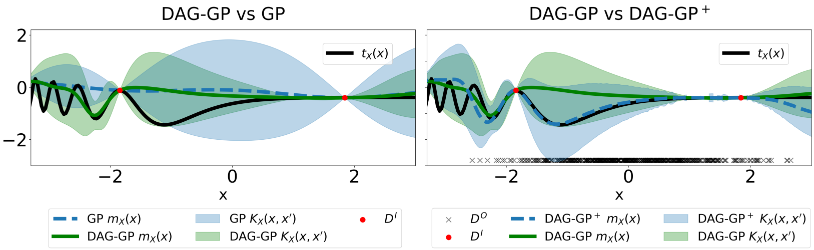


Figure 7: Posterior mean and variance for $t_X(\mathbf{x}) = f_X(\mathbf{x})$.

DAG-GP as surrogate model in CBO

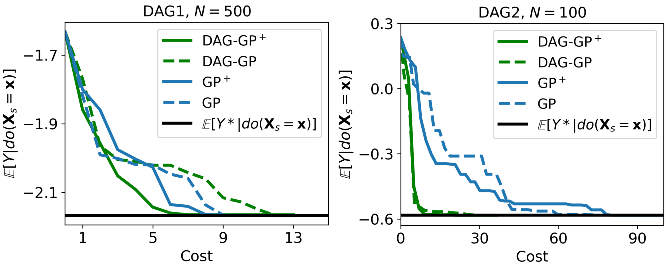
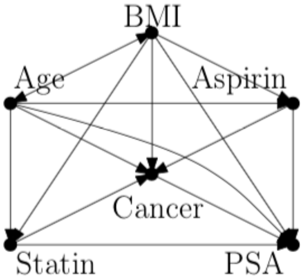


Figure 8: Convergence of the CBO algorithm to the global optimum ($\mathbb{E}[Y^\star|\mathrm{do}\,(\mathbf{X}_s = \mathbf{x})]$) when a single-task or a multi-task GP model are used as surrogate models.

DAG-GP as surrogate model in CBO



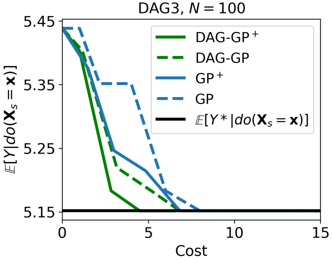


Figure 9: Convergence of the CBO algorithm to the global optimum for PSA ($\mathbb{E}[PSA^\star|\mathrm{do}\,(\mathbf{X}_s = \mathbf{x})]$) when a single-task or a multi-task GP model are used as surrogate models.

Take home messages:

1. The DAG-GP allows to efficiently learn the causal effects in a graph and identify the optimal intervention to perform
2. It captures the non-trivial correlation structure across different experimental outputs.
3. It enables proper uncertainty quantification and can be used within decision-making algorithm to choose experiments to perform.

Thanks for your attention!

Multi-task Causal Learning with Gaussian Processes

Virginia Aglietti
University of Warwick
The Alan Turing Institute
V.Aglietti@warwick.ac.uk

Theodoros Damoulas
University of Warwick
The Alan Turing Institute
T.Damoulas@warwick.ac.uk

Mauricio Álvarez
University of Sheffield
Mauricio.Alvarez@sheffield.ac.uk

Javier González
Microsoft Research Cambridge
Gonzalez.Javier@microsoft.com

Abstract

This paper studies the problem of learning the correlation structure of a set of intervention functions defined on the directed acyclic graph (DAG) of a causal model. This is useful when we are interested in jointly learning the causal effects of interventions on different subsets of variables in a DAG, which is common in field

Causal Bayesian Optimization

Virginia Aglietti
University of Warwick
The Alan Turing Institute
V.Aglietti@warwick.ac.uk

Xiaoyu Lu
Amazon
Cambridge, UK
luxiaoyu@amazon.com

Andrei Paleyes
Amazon
Cambridge, UK
paleyes@amazon.com

Javier González
Amazon
Cambridge, UK
gojav@amazon.com

Abstract

This paper studies the problem of globally optimizing a variable of interest that is part of a causal model in which a sequence of interventions can be performed. This problem arises in biology, operational research, communications and, more generally, in all fields where the goal is to optimize an output metric of a system of interconnected nodes. Our approach combines ideas from causal inference, uncertainty quantification and sequential decision making. In particular, it generalizes Bayesian optimization, which treats

manipulating variables in order to optimize an outcome of interest. For instance, in strategic planning, companies need to decide how to allocate scarce resources across different projects or business units in order to achieve performance goals. In biology, it is common to change the phenotype of organisms by acting on individual components of complex gene networks. This paper describes how to find such interventions or policies.

Focusing on a specific example, consider a setting in which $\mathbf{Y}$ denotes the crop yields for different agricultural products, X denotes soil fumigants and $\mathbf{Z} = \{Z_1, Z_3, Z_4\}$ represents the eel-worm population at different times (Cochran and Cox, 1957). Given a causal graph (Pearl, 1995) representing the investigator's un-