Causal Bayesian optimization

'Don't do everything, just do the right thing'

Javier González

July 8, 2020

Microsoft Research Cambridge

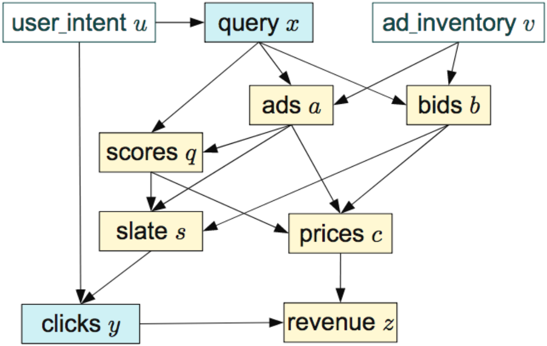
Virginia Aglietti

Xiaoyu Lu

Andrei Paleyes

Systems/processes decompose in sets of interconnected nodes

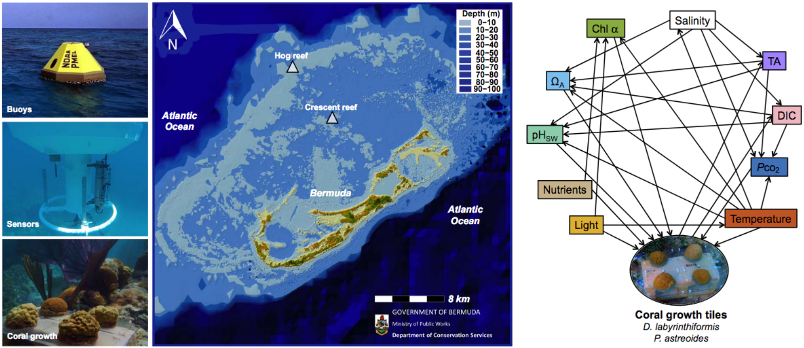
Example in advertising:



(Bottou et al, 2013) The goal is to design advertising campaigns to *maximize* revenue.

Systems/processes decompose in sets of interconnected nodes

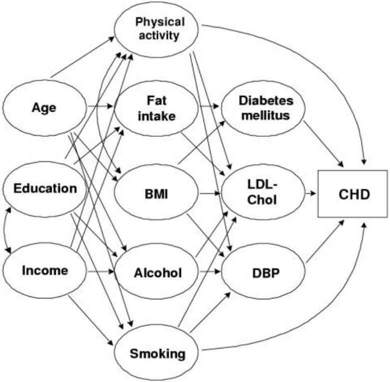
Example in ecology:



(Courtney et al, 2017) The goal is to apply policies to *improve/maximise* coral calcification.

Systems/processes decompose in sets of interconnected nodes

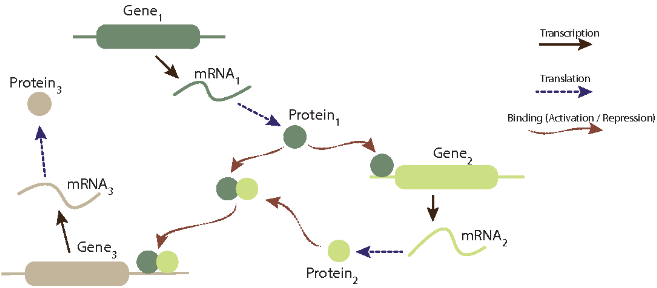
Example in healthcare:



(Murray et al, 2003) The goal is to define treatments to *minimize* the risk of coronary heart disease (CHD)

Systems/processes decompose in sets of interconnected nodes

Example in biology:



(González, 2015; Maksimov, 2015) The goal is to target is to *maximize* the synthetic production of a protein of pharmacological interest.

Common elements in these problems

- A causal graph.
- 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.

Common goal

Find the *system/process configuration* that optimises the *target node*.

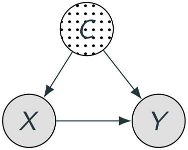
- *System/process configuration* $\rightarrow$ actionable variables.
- *Target node* $\rightarrow$ revenue, coral calcification, risk of desease, etc.

Take home messages:

1. **Many systems/processes decompose in interconnected nodes.**
2. **Optimization requires 'intervening' in the actionable nodes.**

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

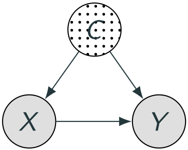
- $\mathbf{U}$: independent *exogenous* background variables.
- $P(\mathbf{U})$ distribution of $\mathbf{U}$.
- $\mathbf{V}$: *endogenous* variables (non-manipulative $\mathbf{C}$, treatment $\mathbf{X}$).
- $F = \{f_1, ..., f_{|\mathbf{V}|}\}$: functions $v_i = f_i(pa_i, u_i)$, pa_i are the parents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



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

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

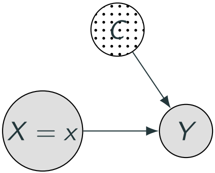
Observed universe



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

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

Post-interventional universe



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

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

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

Key question: *How to do inference in the post-interventional universe with data from the observed universe.*

Observing vs. doing:

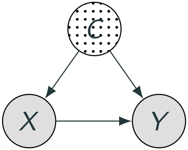
- $P(Y|\text{do}\,(X = x)\,, C)$ requires change the way the universe works.
- $P(Y|X = x, C)$ only requires 'observing' the universe.

do-calculus: algebra to emulate the post-intervention universe in terms of conditionals in the observed universe (experiments emulation).

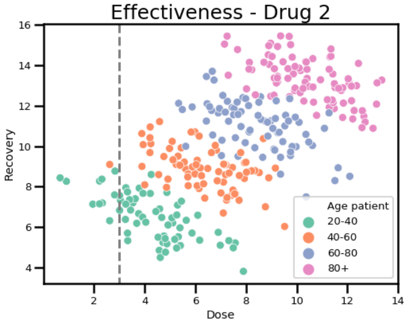
$$P(\mathbf{V}) = \prod_{i=1}^{|\mathbf{V}|} p(V_i|pa_i) \quad \text{(Markov condition)}$$

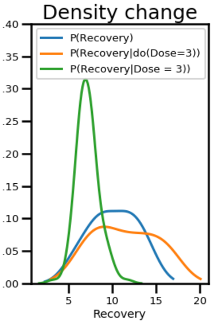
.

Do-calculus: back-door adjustment:



$$p(Y|do(X=x)) = \int P(Y|C,X)P(C)dC$$



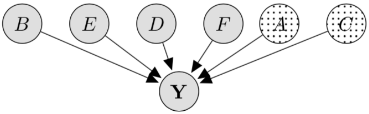


Take home messages:

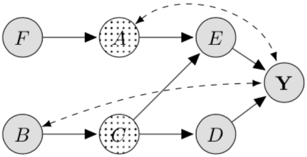
1. Many real systems decompose in interconnected nodes.
2. Optimization requires ‘intervening’ in the actionable nodes.
3. Do-calculus: ‘emulating’ experiments with observational data.

Gobal optimization vs. Causal optimization

Global optimization



Causal optimization



$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in D(\mathbf{X})} \mathbb{E}_{P(\mathbf{Y}|\mathrm{do}(\mathbf{X}=\mathbf{x}),\mathbf{C})}[\mathbf{Y}]$$

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

Assumptions

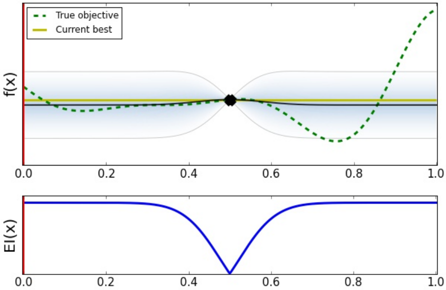
Global optimization

- f, the objective function, is explicitly unknown and multimodal.
- Evaluations of f may be perturbed by noise.
- Evaluations of f are expensive.

Standard method for this scenario $\rightarrow$ ***Bayesian optimization***

Bayesian optimization

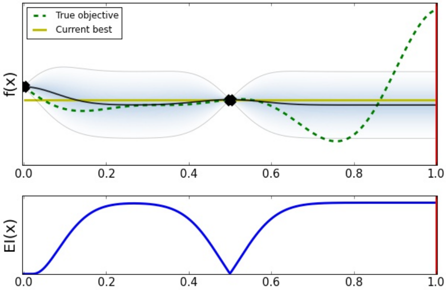
- **Goal**: Collect data $x_1, \ldots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$



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

Bayesian optimization

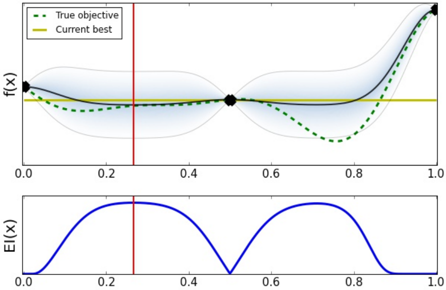
- **Goal**: Collect data $x_1, \ldots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$



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

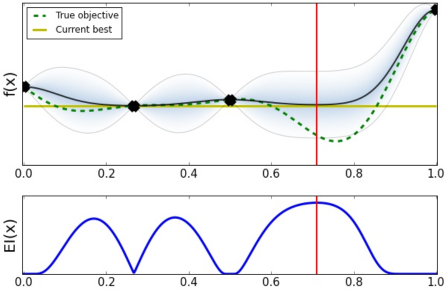
Bayesian optimization

- **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 $x_{n+1} = arg \max \alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}_n)$

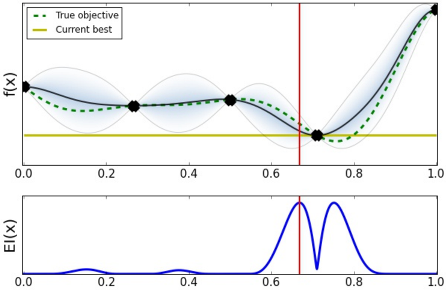
- **Goal**: Collect data $x_1, \ldots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$



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

Bayesian optimization

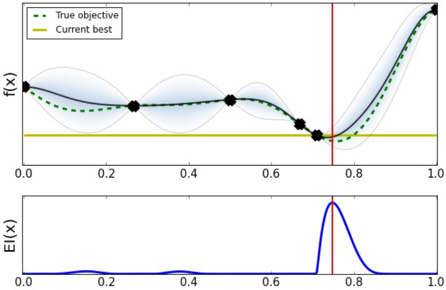
- **Goal**: Collect data $x_1, \ldots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y)p(y|\mathbf{x}; \theta, \mathcal{D})dy$



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

Bayesian optimization

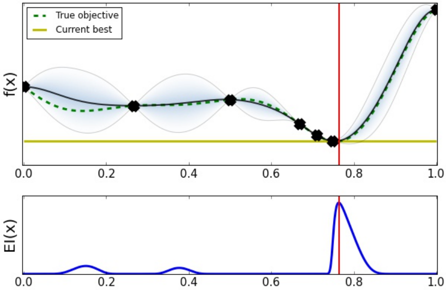
- **Goal**: Collect data $x_1, \dots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$



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

Bayesian optimization

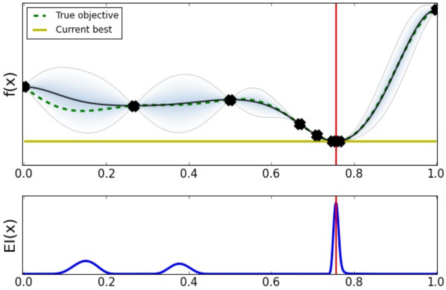
- **Goal**: Collect data $x_1, \ldots, x_n$ to find the optimum as fast as possible.
- **Model**: Gaussian process $f(x) \sim \mathcal{GP}(\mu(x), k_\theta(x, x'))$.
- **Acquisition**: $\alpha_{EI}(\mathbf{x}; \theta, \mathcal{D}) = \int_y \max(0, y_{best} - y) p(y|\mathbf{x}; \theta, \mathcal{D}) dy$



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

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

Assumptions

Causal optimization

- f is explicitly unknown and multimodal.
- Evaluations of f may be perturbed by noise.
- Evaluations of f are expensive.

$+$

- Causal graph
- Cost associate to experiment with each variable.

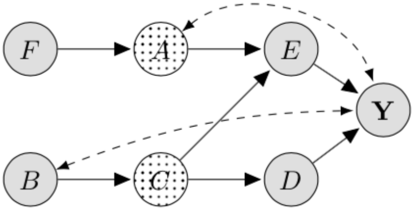
New method for this scenario $\rightarrow$ ***Causal Bayesian optimization***

Take home messages:

1. **Many systems/processes decompose in interconnected nodes.**
2. **Optimization requires 'intervening' in the actionable nodes.**
3. **Do-calculus: 'emulating' experiments with observational data.**
4. **Standard Bayesian Optimization ignores causal assumptions.**
5. **Causal Optimization requires a new approach.**

Goal for the new setup

- $\mathbf{X} = \{B, E, D, F\}$: treatment variables
- $\mathcal{P}(\mathbf{X})$, all possible combinations of interventions.
- $\mathbf{X}_s$, $\mathbf{x}_s$, intervention set and its value.
- $\mathbf{X}_s^\star$, $\mathbf{x}_s^\star$, optimal intervention set and value.



Goal: 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 find $\mathbf{X}_s^\star$ in the $2^{|\mathbf{X}|}$ sets in $\mathcal{P}(\mathbf{X})$? NO!

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

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 $\mathbb{E}[Y|\mathrm{do}\,(\mathbf{X}_s = \mathbf{x}_s)\,, \mathbf{C}] = \mathbb{E}[Y|\mathrm{do}\,(\mathbf{X}_s' = \mathbf{x}_s')\,, \mathbf{C}]$.

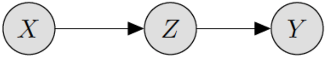
Possibly-Optimal Minimal Intervention set (POMIS, $\mathbb{P}_{\mathcal{G},Y}^{\mathbf{C}}$)

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

- BO, $\mathbb{B}_{\mathcal{G},Y}^{\mathbf{C}}$: all treatment variables
- MIS, $\mathbb{M}_{\mathcal{G},Y}^{\mathbf{C}}$: set of variables 'worth' intervening.
- POMIS, $\mathbb{P}_{\mathcal{G},Y}^{\mathbf{C}}$: set of variables in MIS that always improve Y.

Take home messages:

1. Many systems/processes decompose in interconnected nodes.
2. Optimization requires ‘intervening’ in the actionable nodes.
3. Do-calculus: ‘emulating’ experiments with observational data.
4. Standard Bayesian Optimization ignores causal assumptions.
5. Causal Optimization requires a new approach.
6. No need to explore $\mathcal{P}(\mathbf{X})$ to solve Causal Optimization problems.

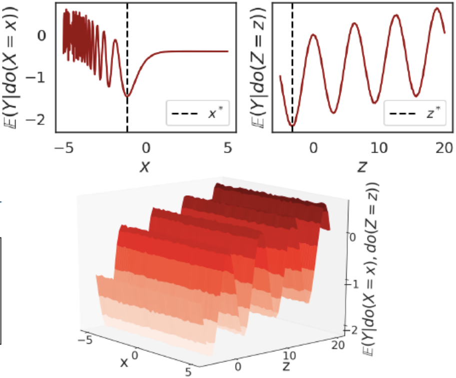


$$X = \epsilon_X$$

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

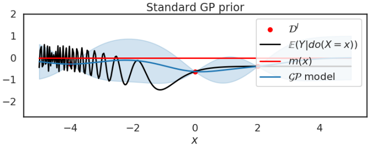
$$Y = \cos(Z) - \exp(-\frac{Z}{20}) + \epsilon_Y$$

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



Modelling $\mathbb{E}_{P(\mathbf{Y}|\mathbf{do}(\mathbf{X}_s=\mathbf{x}_s))}[\mathbf{Y}]$ for each $\mathbf{X}_s$

Standard Gaussian process



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

$$m(\mathbf{x}_s) = \mathbf{0}$$

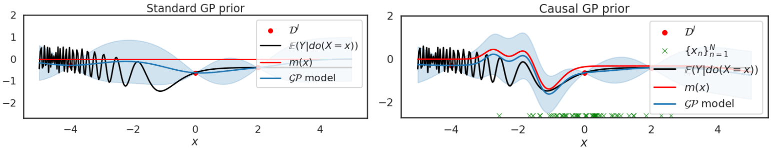
$$k(\mathbf{x}_s, \mathbf{x}'_s) = k_{RBF}(\mathbf{x}_s, \mathbf{x}'_s)$$

where

- $k_{RBF}(\mathbf{x}_s, \mathbf{x}'_s) := \exp(-\frac{||\mathbf{x}_s - \mathbf{x}'_s||^2}{2l^2})$

Modelling $\mathbb{E}_{P(\mathbf{Y}|\mathbf{do}(\mathbf{X}_s=\mathbf{x}_s))}[\mathbf{Y}]$ for each $\mathbf{X}_s$

Causal Gaussian process



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

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

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

where

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

Take home messages:

1. Many systems/processes decompose in interconnected nodes.
2. Optimization requires ‘intervening’ in the actionable nodes.
3. Do-calculus: ‘emulating’ experiments with observational data.
4. Standard Bayesian Optimization ignores causal assumptions.
5. Causal Optimization requires a new approach.
6. No need to explore $\mathcal{P}(\mathbf{X})$ to solve Causal Optimization problems.
7. Causal GPs to merge observational and interventional data.

Causal Expected Improvement (CEI)

Expected improvement in the intervention set:

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

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

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

New intervention set and value

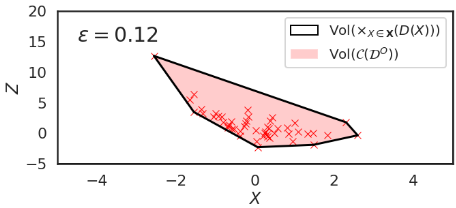
$$\alpha^\star := \max\{\alpha_1, \ldots, \alpha_{|\mathbf{es}|}\}$$

.

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

Intervention-observation trade off

ϵ-greedy criteria to balance interventions and observations



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

- $\mathrm{Vol}(\mathcal{C}(\mathcal{D}^O))$: 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)

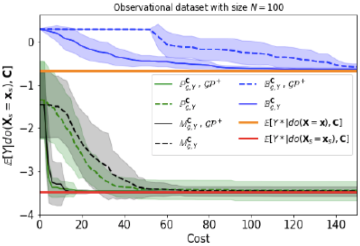
Algorithm 1: Causal Bayesian Optimization-CBO

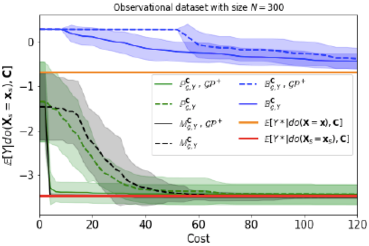
Data: $\mathcal{D}^{\mathrm{O}}$, $\mathcal{D}^{\mathrm{I}}$, $\mathcal{G}$, ES, number of steps T
Result: $\mathbf{X}_s^\star, \mathbf{x}_s^\star, \hat{\mathbb{E}}[\mathbf{Y}^\star|\mathrm{do}\left(\mathbf{X}_s^* = \mathbf{x}_s^\star\right), \mathbf{C}]$
Initialise: Set $\mathcal{D}_0^{\mathrm{I}} = \mathcal{D}^{\mathrm{I}}$ 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 (Eq. (2)).
  end
  else
    (Intervene)
    1. Compute $EI^s(\mathbf{x})/Co(\mathbf{x})$ for each element $s \in$ ES (Eq. (5)).
    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
Return the optimal value $\hat{\mathbb{E}}[\mathbf{Y}^\star|\mathrm{do}\left(\mathbf{X}_s^\star = \mathbf{x}_s^\star\right), \mathbf{C}]$ in $\mathcal{D}_T^{\mathrm{I}}$ and the corresponding $\mathbf{X}_s^\star, \mathbf{x}_s^\star$.

Take home messages:

1. Many systems/processes decompose in interconnected nodes.
2. Optimization requires ‘intervening’ in the actionable nodes.
3. Do-calculus: ‘emulating’ experiments with observational data.
4. Standard Bayesian Optimization ignores causal assumptions.
5. Causal Optimization requires a new approach.
6. No need to explore $\mathcal{P}(\mathbf{X})$ to solve Causal Optimization problems.
7. Causal GPs to merge observational and interventional data.
8. CBO optimizes systems/processes with interconnected nodes.

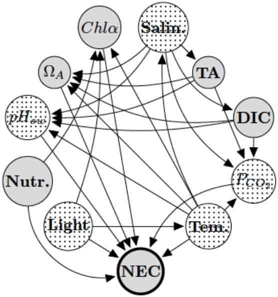
Toy example - simulation analysis

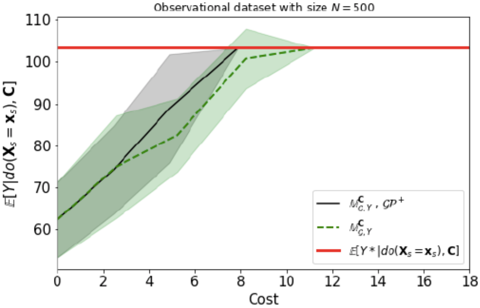




- Equal cost of intervening each variable.
- Results are consistent with what is expected.
- Better results that BO: propagation of effect beyond default domain.

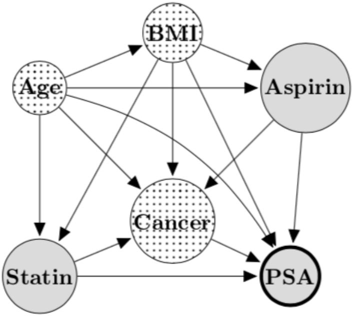
CBO in Ecology





- Goal: maximising the net coral ecosystem calcification (NEC).
- Five manipulative variables: cardinality of MIS is 25.
- Fast convergence to the optimum.

CBO in Healthcare



- Goal is minimize prostate specific antigen (PSA) providing aspiring and/or statin (domain [0,1]).
- Optimal found solution is to only provide statin.
- This agrees with the general practice in medicine.

Take home messages:

1. **Many systems/processes decompose in interconnected nodes.**
2. **Optimization requires ‘intervening’ in the actionable nodes.**
3. **Do-calculus: ‘emulating’ experiments with observational data.**
4. **Standard Bayesian Optimization ignores causal assumptions.**
5. **Causal Optimization requires a new approach.**
6. **No need to explore $\mathcal{P}(\mathbf{X})$ to solve Causal Optimization problems.**
7. **Causal GPs to merge observational and interventional data.**
8. **CBO optimizes systems/processes with interconnected nodes.**
9. **CBO improves BO when causal information is available.**

Discussion

'Causal decision making' takes the best of two worlds:

<table>
<tr><td></td><td></td><td colspan="2">Obs. data + Causal assumptions</td></tr>
<tr><td></td><td></td><td>No</td><td>Yes</td></tr>
<tr><td rowspan="2">Int. data</td><td>No</td><td>Mechanical models
(OR, control, etc.)</td><td>Causal inference
(PO, do-calculus etc.)</td></tr>
<tr><td>Yes</td><td>Sequential decision making
(AL, BayesOpt, etc.)</td><td>Causal Decision Making
(This work!)</td></tr>
</table>

Causal Bayesian Optimization. Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes and Javier González. AISTATS 2020.