Causal Bayesian Optimization

COMP0081 Applied ML
12/03/2024
Virginia Aglietti (Research Scientist)

Contents

- Bayesian Optimization
- Causal Bayesian Optimization
- Extensions to constrained settings and functional interventions

Black-box optimization

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}} f(x)$$

$$f : \mathcal{X} \to \mathbb{R}^P$$

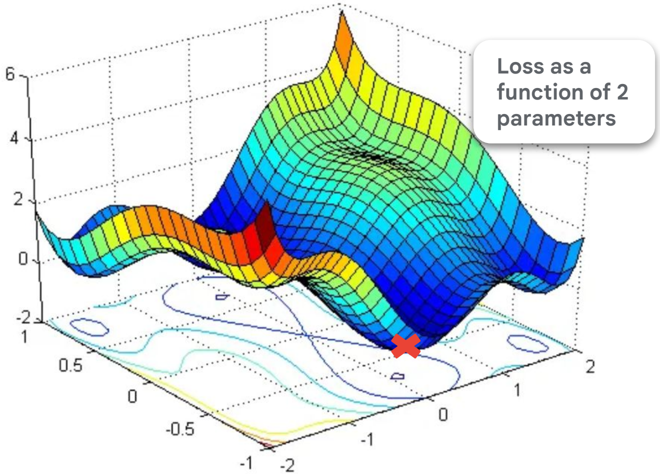
$$\mathcal{X} \subseteq \mathbb{R}^D$$

Setting:

- f is explicitly unknown and multimodal.
- Gradients are not available.
- We can query the function but evaluations of f are expensive.
- Evaluations of f may be perturbed by noise.

Goal: find the global optima x^* in the smallest number of queries

Applications: hyper-parameters optimization, LLMs data mixture, robotics, molecules design, drug design, identification of optimal policies in causal systems etc



Bayesian Optimization

- **Surrogate model:** model our belief about the function which gets updated as we sequentially observe function evaluations
    - Gaussian process/BNN/transformer

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

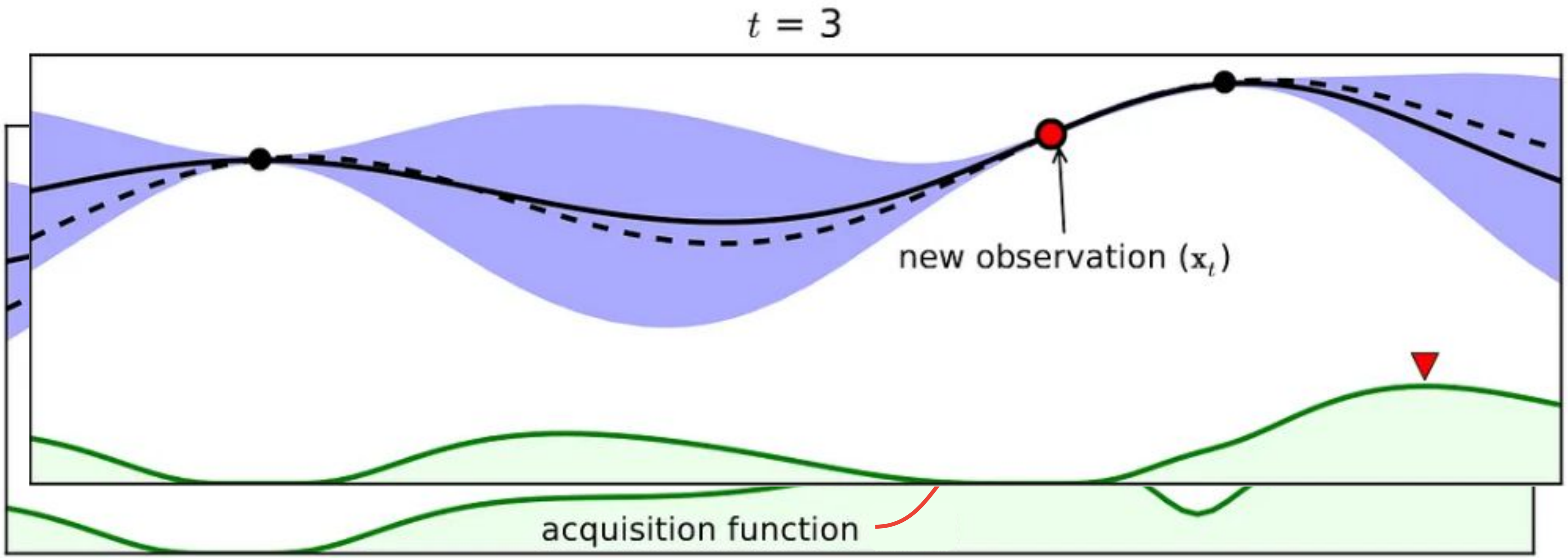
- **Acquisition function (AF):** determines the sequential acquisition of points thus balancing exploration and exploitation
    - Heuristic, ad-hoc choice, problem specific
    - Generally uses the mean and variance of the prediction to determine next function evaluation

$$x^* = \cancel{\underset{x \in \mathcal{X}}{\operatorname{argmin}} f}(x)$$

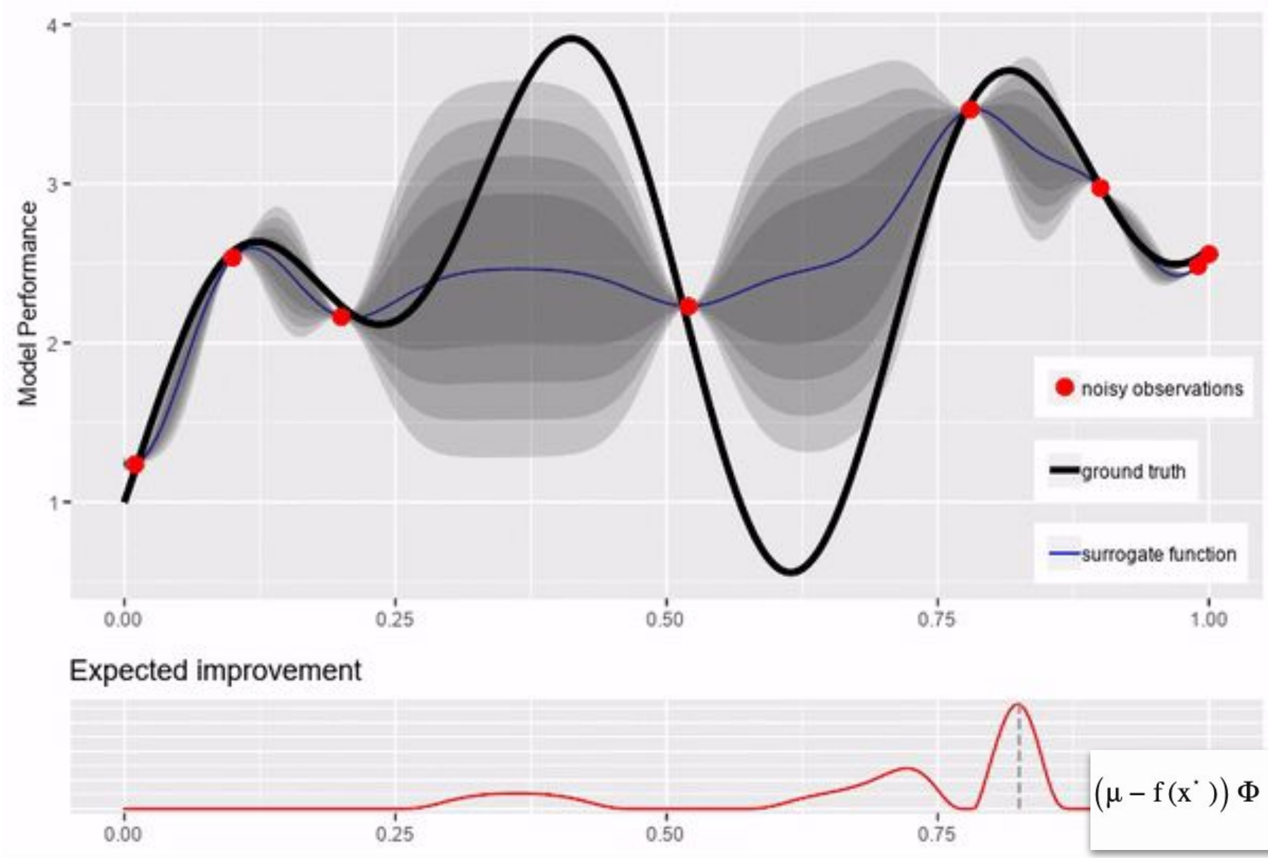
At every trial *t*, solve:

$$x_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\ \alpha_{t,\theta}(x)$$

Bayesian Optimization

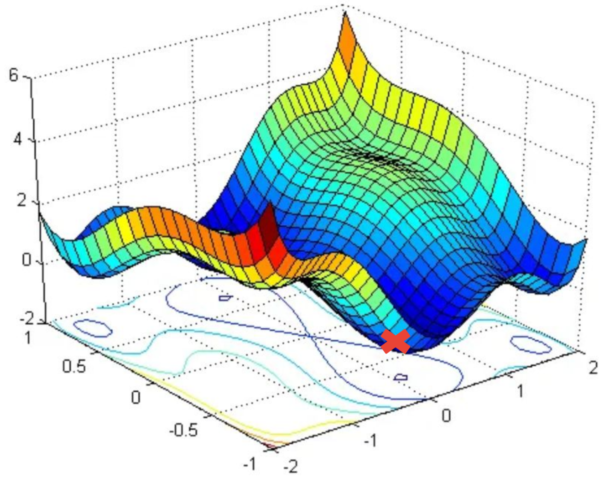


Bayesian Optimization

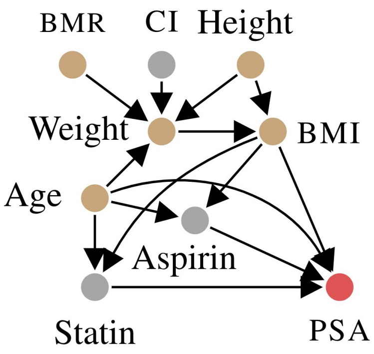


$$(\mu - f(x^*))\,\Phi\left(\frac{\mu - f(x^*)}{\sigma}\right) + \sigma\phi\left(\frac{\mu - f(x^*)}{\sigma}\right)$$

Causal black-box optimization



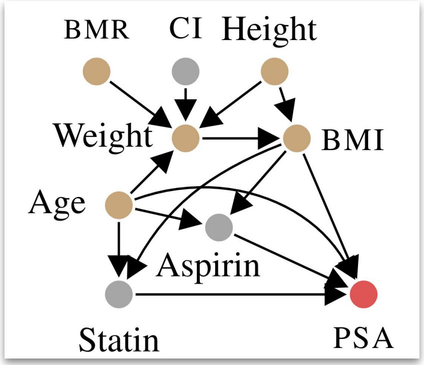
- Loss of a ML model as a function of 2 hyperparameters
- The average causal effect on a target variable given the intervention on 2 different values at different continuous levels



- Target variable Y = PSA
- Intervenable variables $\boldsymbol{I}$ = {CI, Statin, Aspirin}
- $\mathcal{P}_{\boldsymbol{I}}$ = {∅ ,{CI}, {Statin}, {Aspirin}, {CI, Statin}, {CI, Aspirin}, {Statin, Aspirin}, {CI, Statin, Aspirin}}

Causal black-box optimization

- A causal graph $\mathcal{G}$ with nodes $\boldsymbol{V}$
- Target variable $Y \in \boldsymbol{V}$
- Intervenable variables $\boldsymbol{I} \subseteq \boldsymbol{V} \backslash Y$
- Interventional domain $D(\boldsymbol{X}) = \times_{X \in \boldsymbol{X}} D(X)$



E.g. $\boldsymbol{X}^*$ = {CI, Statin, Aspirin}
$\boldsymbol{x}^*$ = (CI=1, Statin=1, Aspirin=0)

$$\underbrace{\boldsymbol{X}^*}_{\text{Optimal intervention set}}, \underbrace{\boldsymbol{x}^*}_{\text{Optimal intervention values}} = \arg\min_{\boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}}, \boldsymbol{x} \in D(\boldsymbol{X})} \mu^{Y}_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}$$

Optimal intervention **set**

Optimal intervention **values**

Power set of intervenable variables ($\mathcal{P}_{\boldsymbol{I}}$)

Domain of possible intervention values ($D(\boldsymbol{X})$)

Target effect $\mathbb{E}_{p_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}}[Y]$

Non-causal vs Causal Bayesian Optimization

Non-causal

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in D(\boldsymbol{I})} \mu^Y_{\mathrm{do}(\boldsymbol{I}=\boldsymbol{x})}$$

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

Bayesian Optimization

Causal

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \arg\min_{\boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x} \in D(\boldsymbol{X})} \mu^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}$$

...

+ Causal Graph

Causal Bayesian Optimization

Causal Bayesian Optimization

1 **Restrict the search space** by exploiting redundancies

2 **Model the target effects using GPs** and exploiting observational and interventional data

3 **Define an acquisition function** that allows to explore the interventions space

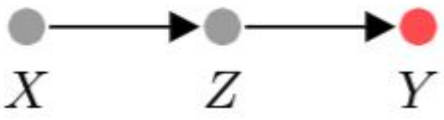
→ *[Causal Bayesian Optimization.](#)*
[V. Aglietti, X. Lu, A. Paleyes, & J. González](#)

Causal Bayesian Optimization: search space

1

Restrict the search space
by exploiting redundancies in $\mathcal{G}$

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \underset{\boxed{\boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x} \in D(\boldsymbol{X})}}{\arg\min}\ \mu^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}$$



$$\mu^Y_{\mathrm{do}(X=x, Z=z)} = \mu^Y_{\mathrm{do}(Z=z)}$$

Assumption: preference for sets of smaller cardinality (e.g. due the intervention cost)

Minimal intervention sets [1]

$$\mathcal{P}_{\boldsymbol{I}} = \{\emptyset, \{X\}, \{Z\}, \{X, Z\}\} \rightarrow \mathbb{M}_{Y,\mathcal{G}} = \{\emptyset, \{X\}, \{Z\}\}$$

[1] Lee Sanghack and Elias Bareinboim. "Structural causal bandits: where to intervene?" *Advances in Neural Information Processing Systems* (2018).

Causal Bayesian Optimization: surrogate models

2 **Model the target effects using GPs** and exploiting observational and interventional data

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \underset{\boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x} \in D(\boldsymbol{X})}{\arg\min} \boxed{\mu^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}}$$

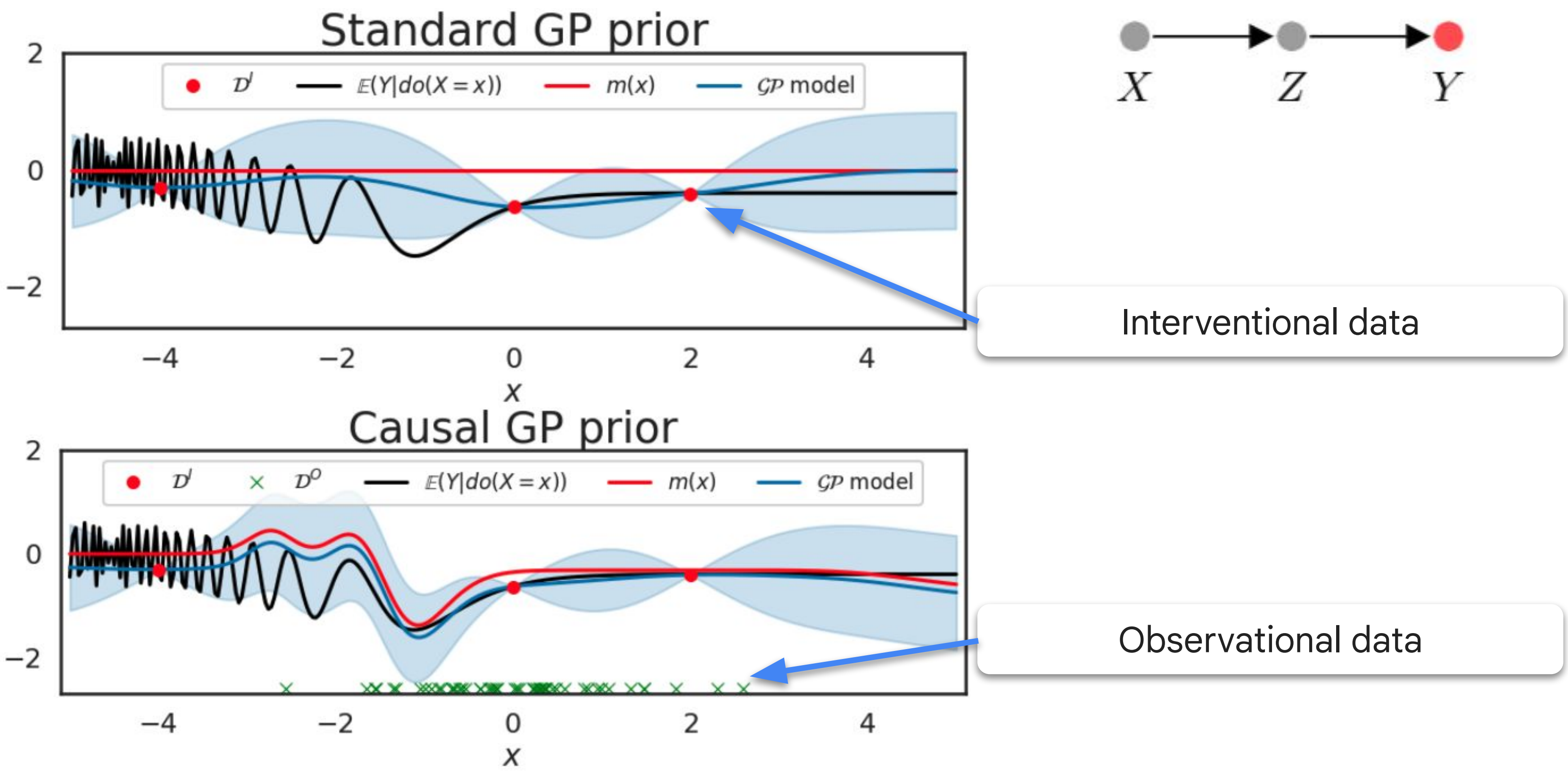
$$g^Y_{\mathbf{X}} \sim \mathcal{GP}(m^Y_{\mathbf{X}}(\boldsymbol{x}'), S^Y_{\mathbf{X}}(\boldsymbol{x}, \boldsymbol{x}'))$$

$$m^Y_{\mathbf{X}}(\boldsymbol{x}') = \hat{\mu}^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}$$

Estimated using observational data and the do-calculus

$$S^Y_{\mathbf{X}}(\boldsymbol{x}, \boldsymbol{x}')) = \sigma_f^2 \exp(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{2l^2}) + \hat{\sigma}^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})} \times \hat{\sigma}^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}')}$$

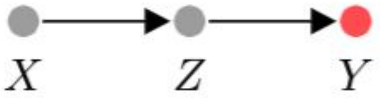
Causal Bayesian Optimization: surrogate models



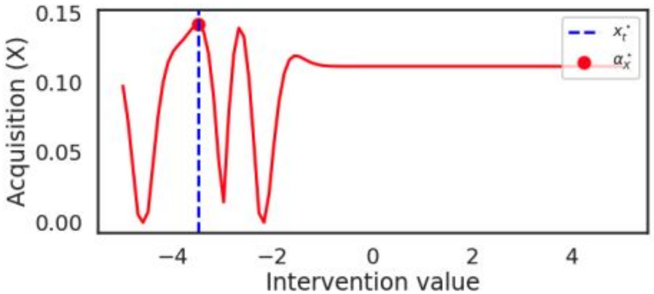
Causal Bayesian Optimization: acquisition function

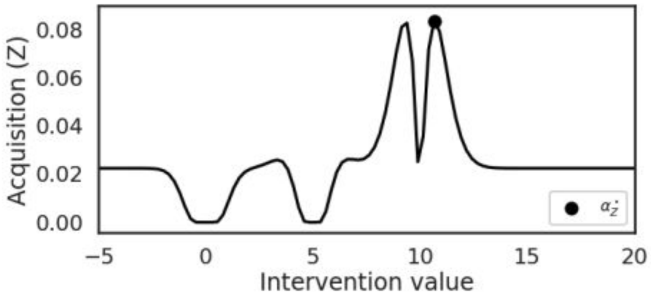
3 **Define an acquisition function** that allows to explore the interventions space

In CBO we optimize the **expected improvement per unit of cost** for every set in $\mathbb{M}_{Y,\mathcal{G}}$ and select the intervention set and intervention values giving the highest expected improvement.

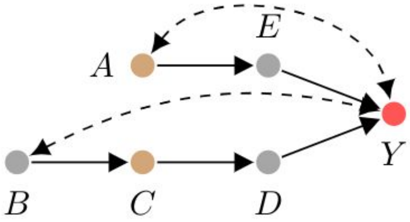


$$\mathrm{EI}_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{E}_{p(g^Y_{\boldsymbol{X}} | \mathcal{D}^I)}[\max(g^Y_{\boldsymbol{X}}(\boldsymbol{x}) - y^*, 0)] \backslash \mathrm{Co}(\boldsymbol{X}, \boldsymbol{x})$$

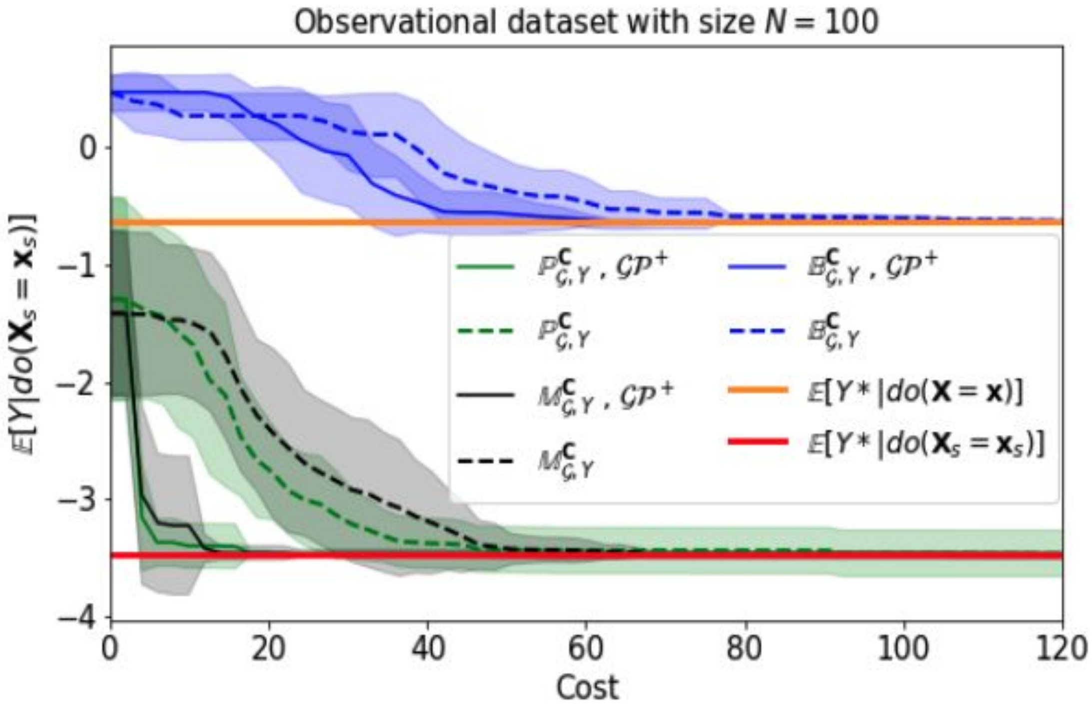




Causal Bayesian Optimization: experimental results



- BO is slower and identifies a suboptimal intervention
- CBO achieves the best result when using the Causal GP model



Causal Bayesian Optimization: limitations

- The number of models GPs we require is determined by the number of sets to explore which is potentially large.
- We don’t transfer interventional information across GPs e.g. we don’t account for the fact that intervening on e.g. X might give us some information about an intervention on X and Z.
- We do not account for time and dynamic changes in the causal effects.
- We assume the causal graph to be known.
- We do not account for the existence of constrained variables.
- We only consider hard interventions.
- …

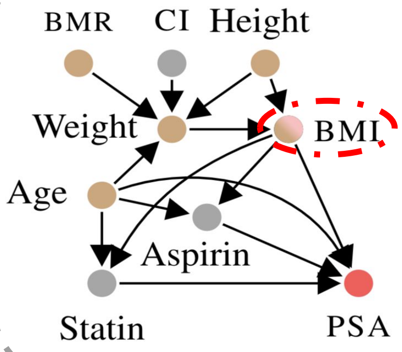
Causal Bayesian Optimization: limitations

- The number of models GPs we require is determined by the number of sets to explore which is potentially large.
- We don’t transfer interventional information across GPs e.g. we don’t account for the fact that intervening on e.g. X might give us some information about an intervention on X and Z.
- We do not account for time and dynamic changes in the causal effects.
- We assume the causal graph to be known.
- **We do not account for the existence of constrained variables.**
- **We only consider hard interventions.**
- …

Constrained Causal Global Optimization

- A causal graph $\mathcal{G}$ with nodes $\boldsymbol{V}$
- Target variable $Y \in \boldsymbol{V}$
- Intervenable variables $\boldsymbol{I} \subseteq \boldsymbol{V} \backslash Y$
- Interventional domain $D(\boldsymbol{X}) = \times_{X \in \boldsymbol{X}} D(X)$

A set of protected variables $\boldsymbol{C} \subseteq \boldsymbol{V} \backslash Y$



Goal: minimize PSA intervening on Statin, Aspirin and CI while keeping BMI lower than 25.

- $Y = \text{PSA}$
- $\boldsymbol{I} = \{\text{Statin}, \text{Aspirin}, \text{CI}\}$
- $\boldsymbol{C} = \{\text{BMI}\}$ $(\boldsymbol{C}_{\boldsymbol{X}} = \boldsymbol{C}\ \forall \boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}})$
- $\lambda^C = 25$

Protected/ Constrained variables

Constrained Causal Global Optimization

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \underset{\boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x} \in D(\boldsymbol{X})}{\arg\min} \mu^{Y}_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}, \quad \text{s.t. } \mu^{\boldsymbol{C_X}}_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})} \geq \lambda^{\boldsymbol{C_X}}$$

Set of **constraint effects** with

$\boldsymbol{C_X} := \boldsymbol{C} \backslash (\boldsymbol{C} \cap \boldsymbol{X})$

Threshold values

Constrained vs Unconstrained vs non-causal Bayesian Optimization

Non-causal

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in D(\boldsymbol{I})}{\arg\min}\, \mu^Y_{\mathrm{do}(\boldsymbol{I}=\boldsymbol{x})}$$

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

Causal unconstrained

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \underset{\boldsymbol{X}\in\mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x}\in D(\boldsymbol{X})}{\arg\min}\ \mu^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}$$

...

Causal constrained

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \underset{\boldsymbol{X}\in\mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x}\in D(\boldsymbol{X})}{\arg\min}\ \mu^Y_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})},\ \text{s.t. } \mu^{\boldsymbol{C_X}}_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})} \geq \lambda^{\boldsymbol{C_X}}$$

...

+ Causal Graph

+ Causal Graph
+ Unknown constraints

Bayesian Optimization

Causal Bayesian Optimization

Constrained Causal Bayesian Optimization

Constrained Causal Bayesian Optimization (cCBO)

1

Restrict the search space
by exploiting redundancies for target *and* constraint effects

2

Model the target effects using GPs
and exploiting observational and interventional data and **capturing the correlation** between target and constraint effects.

3

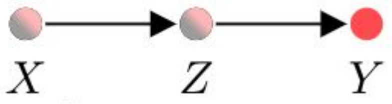
Define an acquisition function
that allows to explore the interventions space accounts for both the target and constraint effects.

→ [*Constrained Causal Bayesian Optimization. V. Aglietti, A. Malek, S. Chiappa.*]

Constrained Causal Bayesian Optimization: surrogate models

2

Model the target effects using GPs and exploiting observational and interventional data and **capturing the correlation** between target and constraint effects.



$$X = \alpha U_X$$
$$Z = \gamma X + U_Z$$
$$Y = \beta Z + U_Y$$
$$U_X, U_Z, U_Y \sim \mathcal{N}(0, 1)$$

When intervening on X:

- Target effect $\mu^Y_{\mathrm{do}(X=x)}$
- Constraint effect $\mu^Z_{\mathrm{do}(X=x)}$

$$\mu^Y_{\mathrm{do}(X=x)} = \beta \mu^Z_{\mathrm{do}(X=x)}$$

Constrained Causal Bayesian Optimization: surrogate models

$$\boldsymbol{X}^*, \boldsymbol{x}^* = \underset{\boldsymbol{X} \in \mathcal{P}_{\boldsymbol{I}},\ \boldsymbol{x} \in D(\boldsymbol{X})}{\arg\min} \boxed{\mu^{Y}_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}}, \quad \text{s.t.} \boxed{\mu^{\boldsymbol{C}_{\boldsymbol{X}}}_{\mathrm{do}(\boldsymbol{X}=\boldsymbol{x})}} \geq \lambda^{\boldsymbol{C}_{\boldsymbol{X}}}$$

We model each effect $\mu^{V_k}_{\mathrm{do}(\boldsymbol{X})}\ \forall V_k \in \boldsymbol{C}_{\boldsymbol{X}} \cup Y$ with a GP:

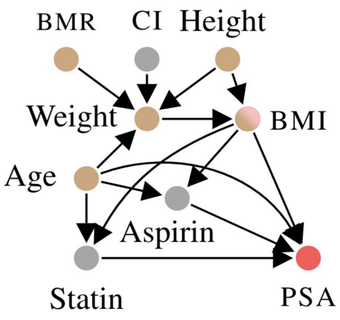
$$g^{V_k}_{\boldsymbol{X}}(\boldsymbol{x}) \sim \mathcal{GP}(\boxed{m^{V_k}_{\boldsymbol{X}}(\boldsymbol{x})}, \boxed{S^{V_k}_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{x}')})$$

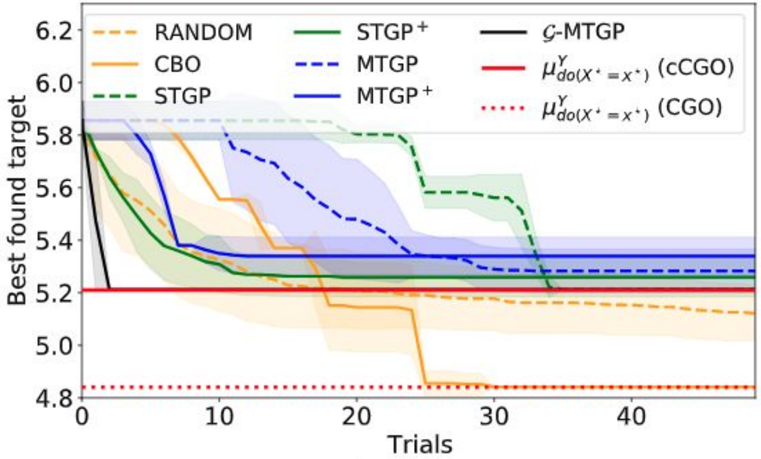
hyperparameters construction that accounts for the correlation induced by the structure of the graph (**multi-task** GP models).

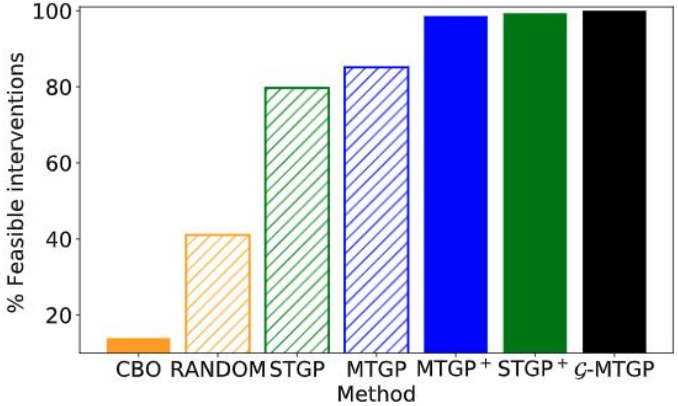
Constrained Causal Bayesian Optimization: Experimental results

- High level of noise in the observational data leads to less accurate prior formulation thus penalizing STGP⁺, MTGP⁺ and $\mathcal{G}$-MTGP.

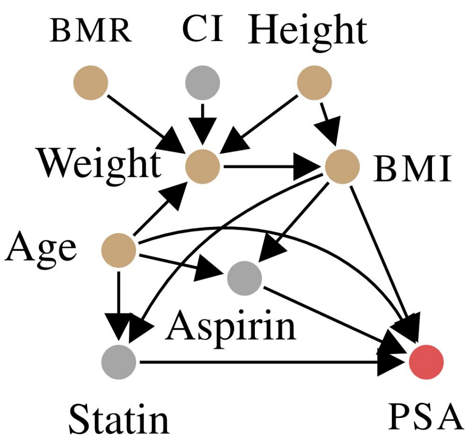
- Capturing the correlation is very important in this setting and leads G-MTGP to outperform all other methods according to both metrics.







Why only Hard Interventions?



Soft/Contextual/Functional Intervention

Often the decision maker has the ability to perform a **conditional/contextual** replacement of the existing causal mechanism, i.e. replace $p(\boldsymbol{X} \mid pa_{\mathcal{G}}(\boldsymbol{X}))$ with another conditional distribution $\pi_{\boldsymbol{X} \mid \boldsymbol{C}_X}$

New parents called **contexts**

E.g.

When finding an optimal value for Statin, we would likely want to take Age and BMI levels into account, as those hold information about the outcome node

Replace $p(\text{Statin} \mid \text{Age}, \text{BMI})$ with $\pi_{\text{Statin} \mid \text{Age,BMI}}$

- Targeted, more personalized treatment
- Subgroup optimality
- Lower treatment cost
- Hard interventions are special cases of soft, so no loss by considering soft

Causal Bayesian Optimization with functional interventions

Mixed Policy Scope (MPS) $\mathcal{S}$

Collection of tuples $\langle X, \mathbf{C}_X \rangle$ where

- ❖ X is an intervenable node $X \in \boldsymbol{I}$
- ❖ $\mathbf{C}_X$ is associated set of contexts for intervention $\pi_{X \mid \mathbf{C}_X}$
- ❖ $\langle X, \mathbf{C}_X \rangle$ does not introduce cycles in the graph

$$\boxed{\mathcal{S}^*}, \boxed{\pi^*_{\mathcal{S}^*}} = \arg \min_{\mathcal{S} \in \boxed{\Sigma}, \boxed{\pi_{\mathcal{S}}} \in \Pi_{\mathcal{S}}} \mu^Y_{\pi_{\mathcal{S}}}$$

Optimal **mixed policy scope (MPS)**

Optimal **MPS realization** (collection of functions)

Set of **all MPSs**

Set of **all possible MPS realizations** (interventions)

➔ [*Functional Causal Bayesian Optimization.*](#)
[*L. Gultchin, V. Aglietti, A. Bellot, I. Ktena, S. Chiappa*](#)

Causal Bayesian Optimization with functional interventions

$$\mathcal{S}^*, \pi^*_{\mathcal{S}^*} = \arg\min_{\mathcal{S}\in\Sigma, \pi_{\mathcal{S}}\in\Pi_{\mathcal{S}}} \mu^{Y}_{\pi_{\mathcal{S}}}$$

Optimal **mixed policy scope (MPS)**

Optimal **MPS realization** (collection of functions)

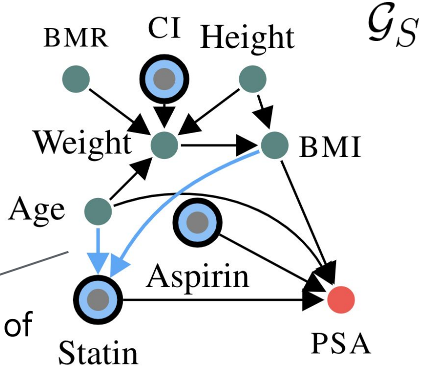
Set of **all MPSs**

Set of **all possible MPS realizations** (interventions)

$$\mathcal{S} = \{\langle \text{Statin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{Aspirin}, \{\text{Age}, \text{BMI}\}\rangle, \langle \text{CI}, \emptyset\rangle\}$$

$$\pi_{\mathcal{S}} = \{\pi_{\text{Statin}\,|\,\text{Age,BMI}}, \pi_{\text{Aspirin}\,|\,\text{Age,BMI}}, \pi_{\text{CI}}\}$$

$$\boxed{\pi_{\text{Statin}|\text{Age, BMI}} = \delta_{\text{Statin}}(\alpha * \text{Age} + \beta * \text{BMI})}$$



Retain the dependency of Statin on Age and BMI

Functional Causal Bayesian Optimization: Surrogate models

2 **Model the target effects using GPs**

$$g_{\mathcal{S}}(\pi) \sim \mathcal{GP}(m_{\mathcal{S}}(\pi), K^{\theta}_{\mathcal{S}}(\pi, \pi'))$$

Surrogate model for the **target effect** $\mu^{Y}_{\mathcal{S}}$ under possible interventions on MPS $\mathcal{S}$

- Prior mean functional $m_{\mathcal{S}}(\pi)$, initialized at 0
- Prior covariance functional $K^{\theta}_{\mathcal{S}}(\pi, \pi')$, RBF kernel with hyperparameters θ
- Functional objective from the space $\Pi_{\mathcal{S}}$ of all bounded (vector-valued) functions on to the reals $\boldsymbol{C}_{\mathcal{S}} = \bigcup_{\langle X, \boldsymbol{C}_X \rangle \in \mathcal{S}} \boldsymbol{C}_X$

$$K^{\theta}_{\mathcal{S}} : \Pi_{\mathcal{S}} \times \Pi_{\mathcal{S}} \to \mathbb{R}$$

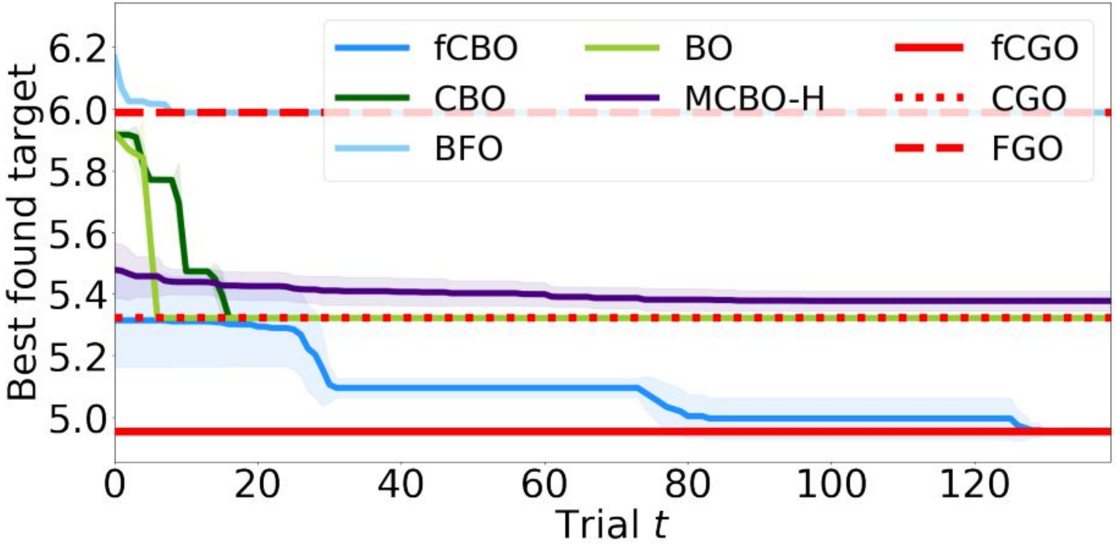
$$\Downarrow$$

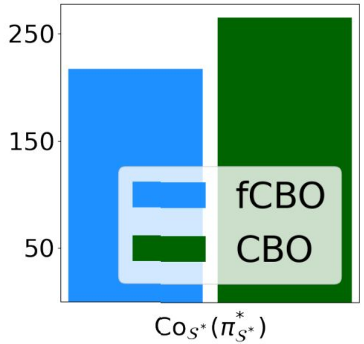
$$\sigma_f^2 \exp(\boxed{-||\pi - \pi'||^2}/2\ell^2) \quad \Rightarrow \quad \langle \pi_{\text{func}} - \pi'_{\text{func}}, \pi_{\text{func}} - \pi'_{\text{func}} \rangle_{H^{\mathcal{S}}_{\kappa}}$$

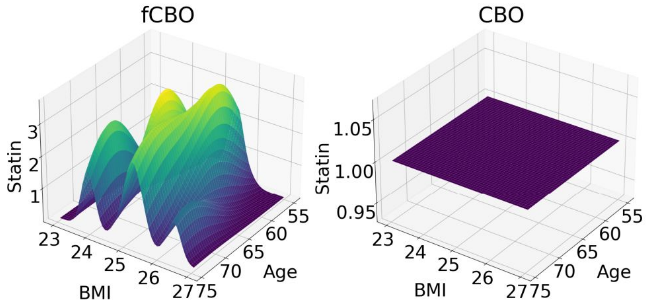
Functional Causal Bayesian Optimization: Healthcare experiment

CBO $\boldsymbol{X}^* = \{\mathrm{CI}, \mathrm{Statin}, \mathrm{Aspirin}\}$ $\boldsymbol{x}^* = (1, 1, 0)$

fCBO $\boldsymbol{S}^* = \{\langle \mathrm{CI}, \emptyset \rangle, \langle \mathrm{Statin}, \{\mathrm{Age}, \mathrm{BMI}\} \rangle, \langle \mathrm{Aspirin}, \emptyset \rangle\}$







Thank you.

Get in touch at aglietti@google.com