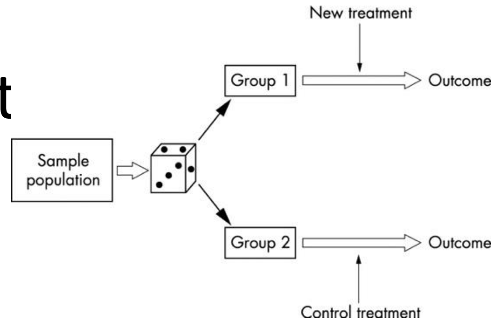
The Effects of Interventions

Mohammad Ali Javidian

The Difference Between Observational and Interventional Studies

- In an **observational** study, investigators collect information by <u>observing</u> or <u>measuring</u> how specific characteristics or outcomes change in study participants over time. However, **no** attempt is made throughout the trial/study to **interfere** or **change** any measured outcomes.
- An **interventional** study tests (or tries out) an intervention -- a potential drug, medical device, activity, or procedure -- in people. It is also commonly referred to as a <u>clinical trial</u>.
- The golden standard (**R**andomized **C**ontrolled **T**rial or **RCT**): In a properly randomized controlled experiment, all factors influencing the outcome variable are either static or vary at random, except for one. So, *any change in the outcome variable* must be due to that <u>one input variable</u>.



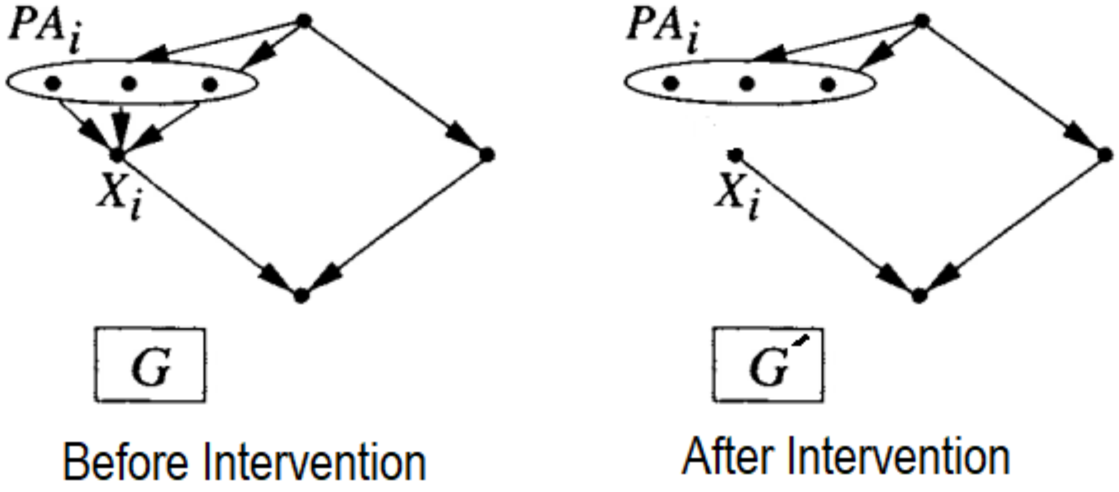
In many cases RCTs are not practical

1) **Feasibility**: We **cannot control (intervene)** the weather, so we can’t randomize the variables that affect wildfires.
2) **Cost**: RCTs are usually <u>time-consuming</u> and <u>expensive</u>.
3) **Ethical considerations**: How can a physician committed to doing what he thinks is best for each patient tell a woman with breast cancer that he is choosing her treatment by something like a coin toss? Or How can a physician force the members of a test group to smoke a box of cigarettes every day to investigate the effect of smoking on lung cancer?
4) **Credibility**: Even randomized drug trials can run into problems when participants *drop out*, *fail to take their medication*, or *misreport their usage*.

- In such cases, researchers instead perform <u>observational studies</u>, in which they merely record data rather than control it. However, *causal inference* from observational data is an *ambitious* and *difficult* task.

Atomic Intervention

- Formally, the **atomic intervention**, which we denote by $\boldsymbol{do}(X_i = x_i)$, or $\boldsymbol{do}(x_i)$ for short, amounts to removing the equation
$$x_i = f_i(pa_i, u_i)$$
- from the SCM model and substituting $X_i = x_i$ in the remaining equations.
- The graph corresponding to the reduced set of equations in an atomic intervention is a subgraph of DAG G from which <u>all arrows entering</u> X_i <u>have been pruned</u>.

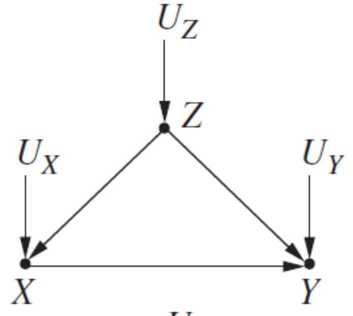


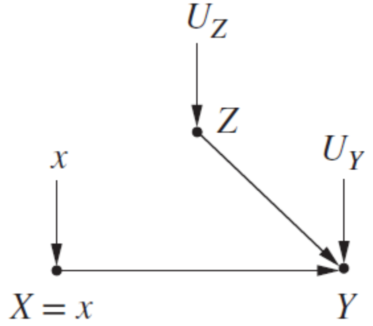
Intervening vs Conditioning

- In notation, we distinguish between cases where a variable X takes a value x naturally (**Conditioning**) and cases where we fix $X = x$ (**Intervening**) by denoting the latter $\boldsymbol{do}(X = x)$.
- So $P(Y = y|X = x)$ is the probability that $Y = y$ <u>conditional</u> on finding $X = x$, while $P(Y = y|\boldsymbol{do}(X = x))$ is the probability that $Y = y$ when we <u>intervene</u> to make $X = x$.
- In the distributional terminology, $P(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x. On the other hand, $P(Y = y|\boldsymbol{do}(X = x))$ represents the **population** distribution of Y if everyone in the population had their X value fixed at x.

The Adjustment Formula

- A graphical model representing the **effects** of a *new drug*, with Z representing <u>gender</u>, X standing for <u>drug usage</u>, Y standing for <u>recovery</u>.
- A **modified** graphical model representing an intervention on the model that sets drug usage in the *population*, and results in the manipulated probability P_m.
- The marginal probability $P(Z = z)$ is **invariant** under the *intervention*, because the process determining Z is <u>not</u> affected by removing the arrow from Z to X (In our example, this means that the proportions of males and females remain the same, before and after the intervention): $P_m(Z = z) = P(Z = z)$



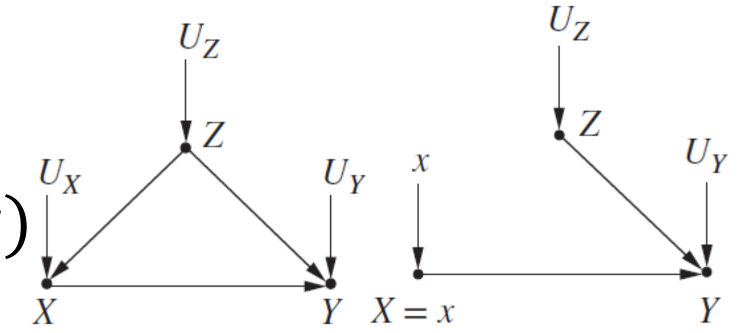


- The conditional probability $P(Y = y|Z = z, X = x)$ is **invariant**, because the process by which Y responds to X and Z, $Y = f\ (x, z, u_Y)$, remains the same, regardless of whether X changes spontaneously or by deliberate manipulation:

$$P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x)$$

- Z and X are <u>*d*-separated</u> in the *modified* model and are, therefore, **independent** under the intervention distribution, i.e., $P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z)$

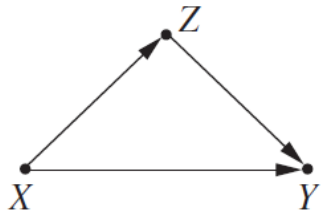
The Adjustment Formula: $P(Y=y|\boldsymbol{do}(X=x))$



- $P_m(Z=z) = P(Z=z)$
- $P_m(Y=y|Z=z, X=x) = P(Y=y|Z=z, X=x)$
- $P_m(Z=z|X=x) = P_m(Z=z) = P(Z=z)$
- Putting these considerations together, we have:
- $P(Y=y|\boldsymbol{do}(X=x)) = P_m(Y=y|X=x)$ by definition
- $= \sum_z P_m(Y=y|X=x, Z=z) P_m(Z=z|X=x)$ by the law of total probability, conditioning on and summing over all values of $Z=z$.
- $= \sum_z P_m(Y=y|X=x, Z=z) P_m(Z=z)$ by the independence of Z and X in the modified model.
- $P(Y=y|\boldsymbol{do}(X=x)) = \sum_z P(Y=y|X=x, Z=z) P(Z=z)$

Adjust or not to Adjust?

- A graphical model representing a *new drug*'s **effects**, X representing <u>drug usage</u>, Y representing <u>recovery</u>, and Z representing <u>blood pressure</u> (*measured at the end of the study*). <u>Exogenous</u> variables are **not** shown in the graph, implying that they are mutually independent.
- The intervention graph is equal to the original graph—no arrow need be removed—and the adjustment formula reduces to:
$$P(Y = y|\boldsymbol{do}(X = x)) = P(Y = y|X = x)$$
- Obviously, if we were to adjust for blood pressure, we would obtain an incorrect assessment—one corresponding to a model in which blood pressure causes people to seek treatment.



Simpson's Paradox: Second Look

- A graphical model representing the **effects** of a *new drug*, with Z representing gender, X standing for drug usage, Y standing for recovery. Given the results of this study in **Table 1.1**, then, should a doctor prescribe the drug for a woman? A man? A patient of unknown gender?
$$P(Y=y|\boldsymbol{do}(X=x)) = \sum_z P(Y=y|X=x,Z=z)P(Z=z)$$
- A graphical model representing a *new drug*'s **effects**, X representing drug usage, Y representing recovery, and Z representing blood pressure (*measured at the end of the study*). Given the results of this study in **Table 1.2**, would you recommend the drug to a patient?
$$P(Y=y|\boldsymbol{do}(X=x)) = P(Y=y|X=x)$$

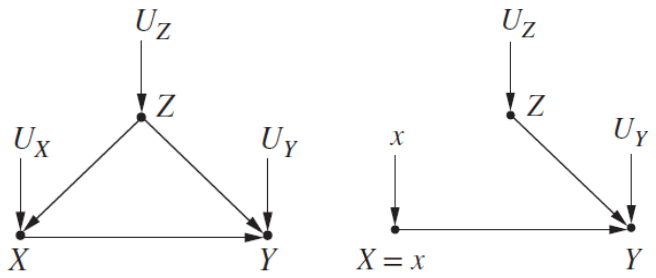


Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

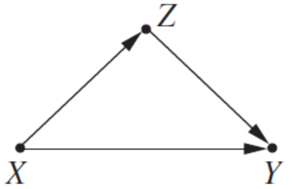
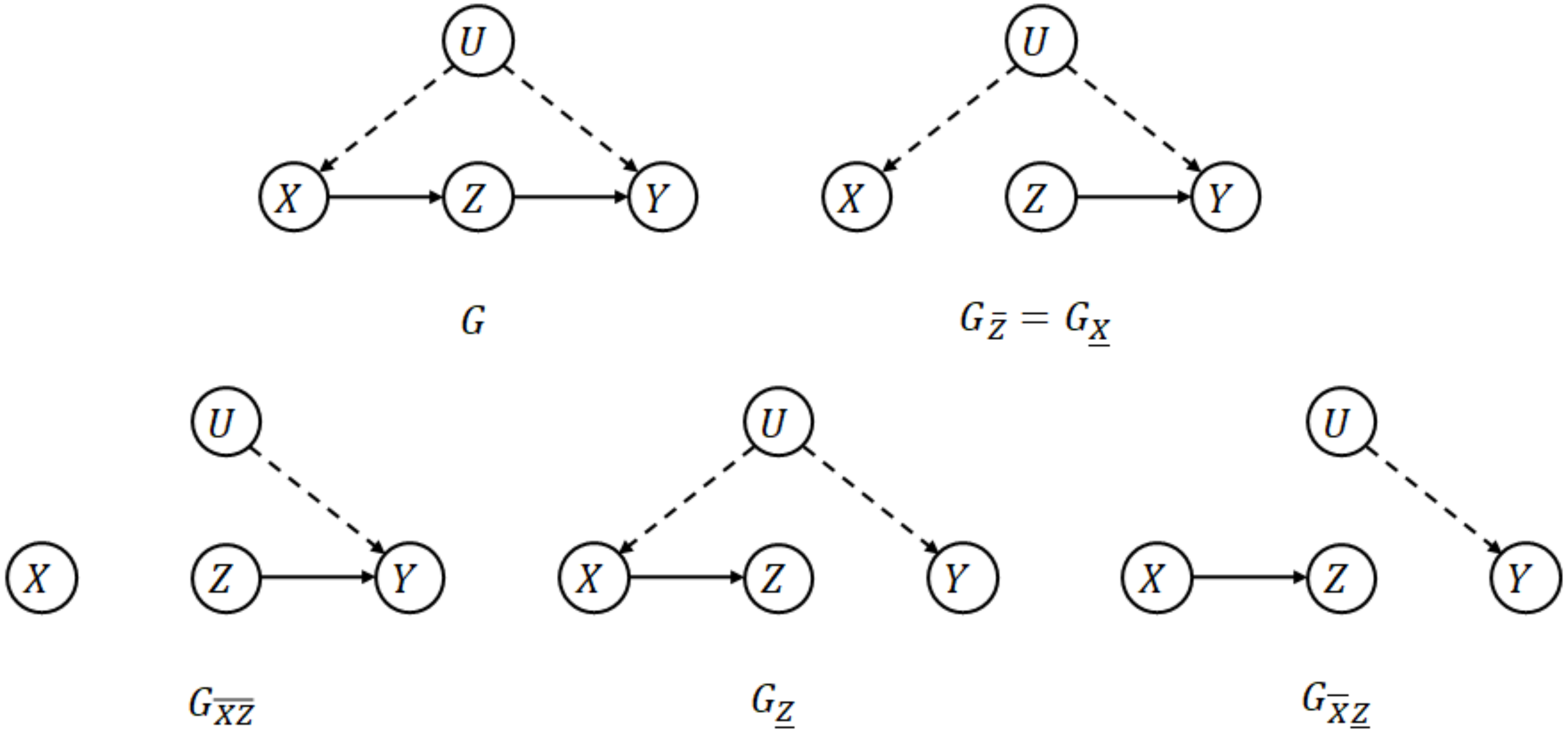


Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Symbolic Derivation of Causal Effects: Graphical Notation

- Subgraphs of G used in the derivation of causal effects



do-Calculus

- **Rule 1** (Insertion/deletion of observations):

$$P(y|\boldsymbol{do}(x), z, w) = P(y|\boldsymbol{do}(x), w) \qquad if \qquad (Y \perp\!\!\!\perp_d Z|X, W)_{G_{\overline{X}}}$$

- **Rule 2** (Action/observation exchange):

$$P(y|\boldsymbol{do}(x), \boldsymbol{do}(z), w) = P(y|\boldsymbol{do}(x), z, w) \quad if \quad (Y \perp\!\!\!\perp_d Z|X, W)_{G_{\overline{X}\underline{Z}}}$$

- **Rule 3** (Insertion/deletion of actions):

$$P(y|\boldsymbol{do}(x), \boldsymbol{do}(z), w) = P(y|\boldsymbol{do}(x), w) \quad if \quad (Y \perp\!\!\!\perp_d Z|X, W)_{G_{\overline{X}\overline{Z(W)}}}$$

  where $Z(W)$ is the set of Z -nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

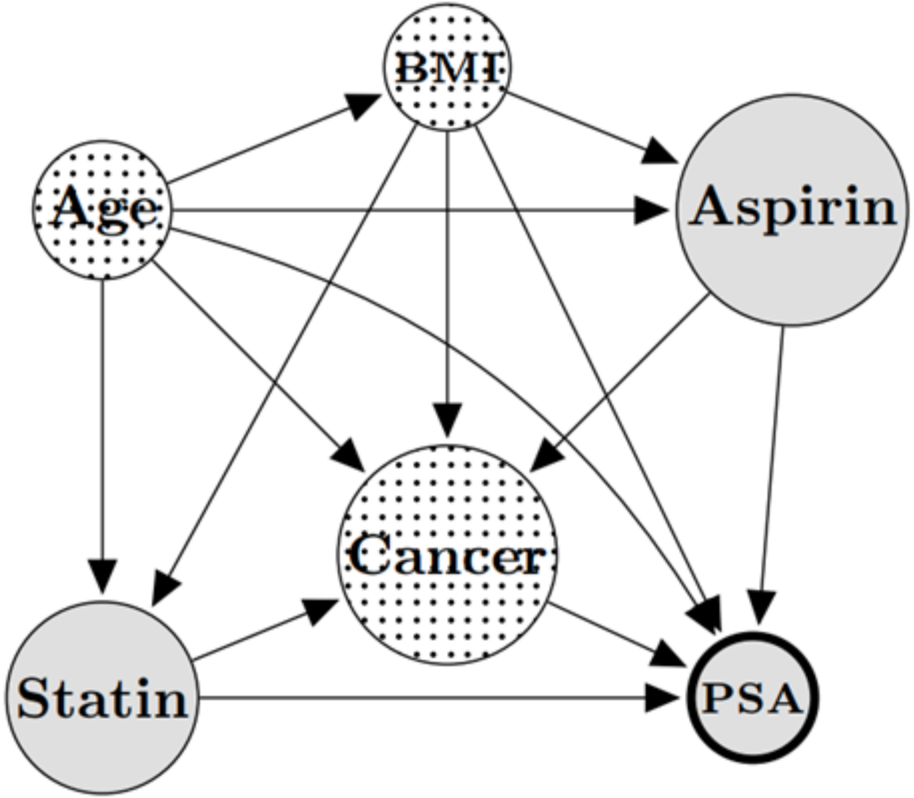
- Note that, in all the derivations, the graph G provides both the license for applying the inference rules and the guidance for choosing the right rule to apply.

CausalBO: A Python Package for Causal Bayesian Optimization

Jeremy Roberts
Dr. Mohammad Ali Javidian
Appalachian State University
Department of Computer Science

What is the goal?

- Doctors want to estimate PSA levels in unseen patients based on accumulated data.
- Seems to be a classic regression problem.

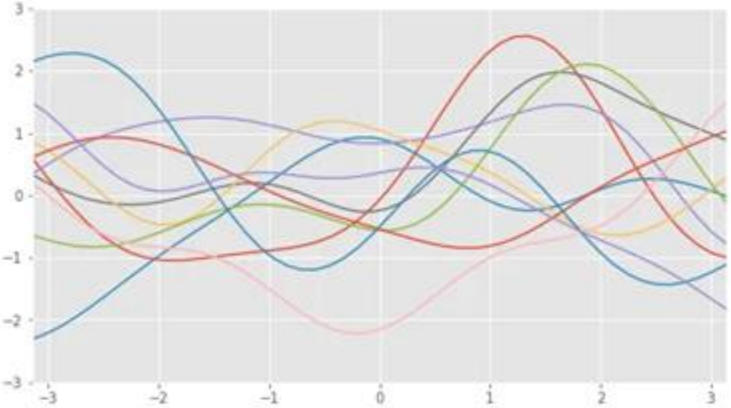


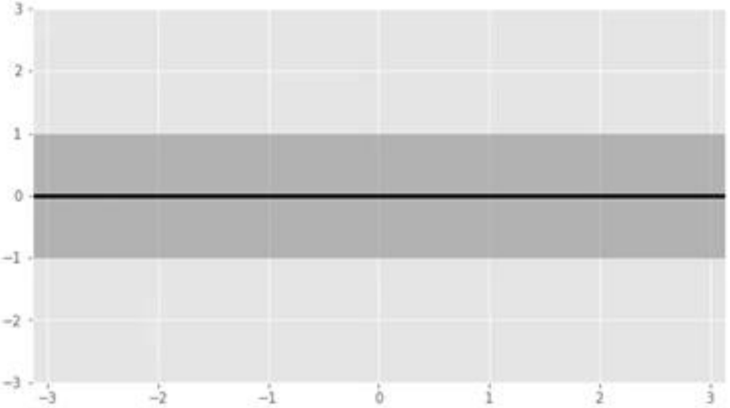
Bayesian Optimization - Overview

- Goal is to optimize (maximize or minimize) some unknown function f for which we have observed some values.
  - Model f as a probability distribution
  - Given observed values $f(x_1)$, $f(x_2)$, . . . , $f(x_n)$, compute the conditional probability distribution $\mathcal{P}(f(x) \mid f(x_1), f(x_2), \ldots, f(x_n))$.
  - Use conditional probability distribution to estimate $f(x)$ for unobserved values of x to find optimal value of unknown function.

Bayesian Optimization - Steps

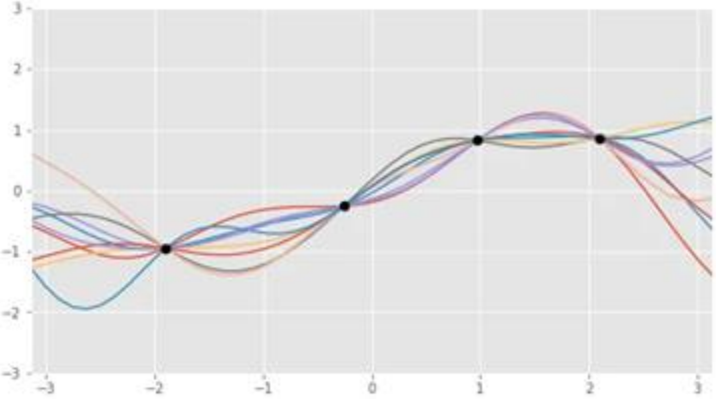
1. Observe set of **prior** data points and build Gaussian Process from **prior**.
    a. Gaussian process - probability distribution over infinite number of possible functions that fit prior data.
    b. Generally selected through random sampling at first iteration.

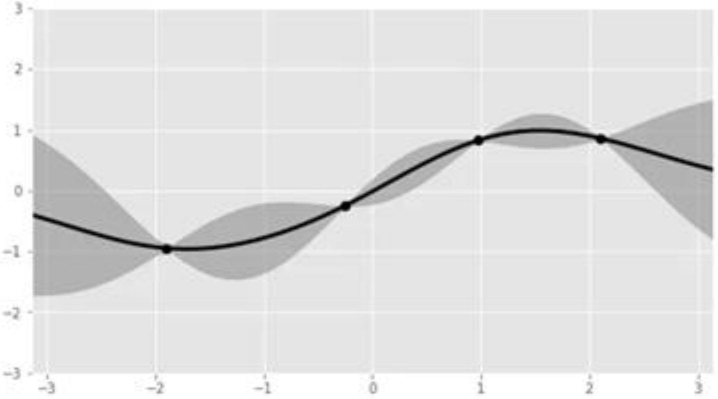




Bayesian Optimization - Steps

2. Sample an observation chosen via **acquisition function**.
    a. Acquisition function - Metric that examines current **prior** and determines which point to observe next.
    b. Attempts to maximize the amount of information gain at each step.
3. Generate **posterior** by adding observation to **prior**.

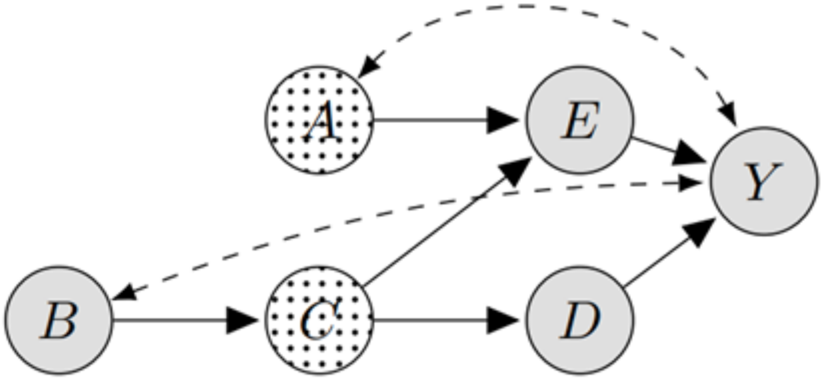




Bayesian Optimization - Steps

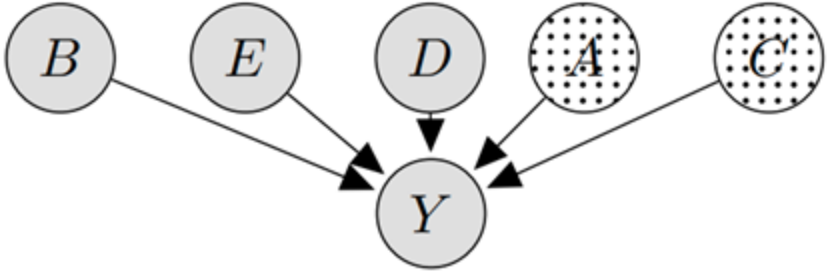
4. Until budget for iterations is exhausted **posterior** becomes the new **prior**, then goto 1.
5. Return maximum/minimum value and argmax/argmin of GP surrogate function.

Bayesian Optimization - Limitations



1. True DAG for Optimization Problem

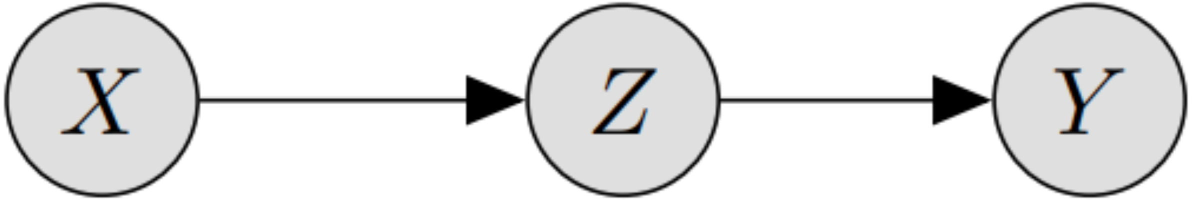
2. DAG Considered by Bayesian Optimization



Causal Bayesian Optimization - Exploration Sets

- Contain a list of possible sets upon which intervention can be performed.
- **M**inimal **I**ntervention **S**et **(MIS)**
  - Defined as a set of variables X_s where no subset $X_s' \subset X_s$ exists such that $\mathbb{E}[Y \mid \mathrm{do}(X_s = x_s)] = \mathbb{E}[Y \mid \mathrm{do}(X_s' = x_s')]$.
- **P**ossibly **O**ptimal **M**inimal **I**ntervention **S**et **(POMIS)**
  - Optimized subset of **MIS** that removes redundancy by removing sets that have the same causal effect as another set that has a lower cardinality.
- **E**xploration **S**et **(ES)**
  - Either the **MIS** or **POMIS** can be chosen as the exploration set for a given problem, this is left up to the agent. The remainder of the algorithm is agnostic to the choice of **ES**.

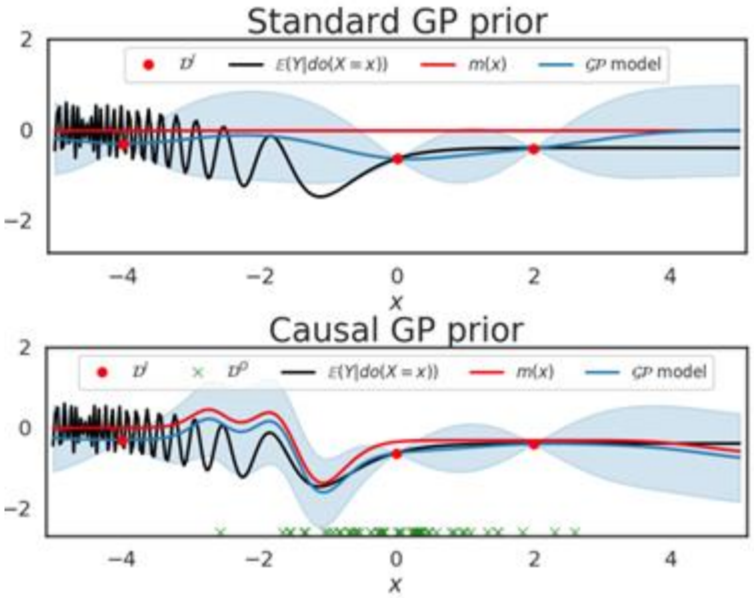
Causal Bayesian Optimization - Exploration Sets (cont.)



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

Causal Bayesian Optimization - Causal GP

- CBO begins by initializing a Gaussian Process on $\mathbf{f(x_s) = E[Y \mid do\ (X_s = x_s)]}$ for every set $\mathbf{X_s \in ES}$.



Causal Bayesian Optimization - Acquisition

- Standard acquisition functions aim to find the next best area of the observational data to observe.
- Causal acquisition function aims to find the next best area in the DAG to intervene.
  - $EI^s(x) = E_{p(y_s)}[\max(y_s - y^\star, 0)]/Co(x)$
- Where $y_s - y^\star$ represents the difference in performance between the proposed interventional setting and the current best observed interventional setting across all sets in **ES**, and $Co(x)$ represents the cost of performing the proposed intervention.

Causal Bayesian Optimization - ε

- Standard Bayesian Optimization aims to balance an exploration-exploitation tradeoff.
  - Should the algorithm continue to explore areas where it has already found promising results (exploitation) or begin observing unknown areas (exploration)?
- Causal Bayesian Optimization aims to balance the observation-intervention tradeoff.
  - Observing new datapoints allows for reliable causal estimation using *do*-calculus, while predicting causal effects for areas outside of observational data requires intervention.
- Parameter ε represents the probability of observing a datapoint rather than intervening.

Causal Bayesian Optimization (Virginia Aglietti et. Al. 2020)

Algorithm 1: Causal Bayesian Optimization - CBO

Data: $\mathcal{D}^{\mathrm{O}}$, $\mathcal{D}^{\mathrm{I}}$, $\mathcal{G}$, ES, number of steps T
Result: $\mathbf{X}_s^\star, \mathbf{x}_s^\star, \hat{\mathbb{E}}[\mathbf{Y}^\star|\mathrm{do}\,(\mathbf{X}_s^\star = \mathbf{x}_s^\star)]$
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 (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 causal GP.
  end
end
Return the optimal value $\hat{\mathbb{E}}[\mathbf{Y}^\star|\mathrm{do}\,(\mathbf{X}_s^\star = \mathbf{x}_s^\star)]$ in $\mathcal{D}_T^{\mathrm{I}}$ and the corresponding $\mathbf{X}_s^\star, \mathbf{x}_s^\star$.

Note: Observing updates the GP prior for each $\mathbf{X_s}$∈ **ES**, while intervening updates the GP posterior for only set **s*.**

CausalBO - Causal Modules

- To integrate experimental and observational data, for each $X_s \in ES$, we place a GP prior on $f(x_s) = E(Y|\mathbf{do}(X_s = x_s))$
- CausalMean
    - BoTorch Mean object that includes information about the causal relationships between variables to predict mean using do-calculus.
- CausalRBF
    - BoTorch Kernel object that includes information about causal relationships to calculate variances required to determine covariances.
- Modules can easily replace Mean and Kernel modules in existing BoTorch implementations.

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

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

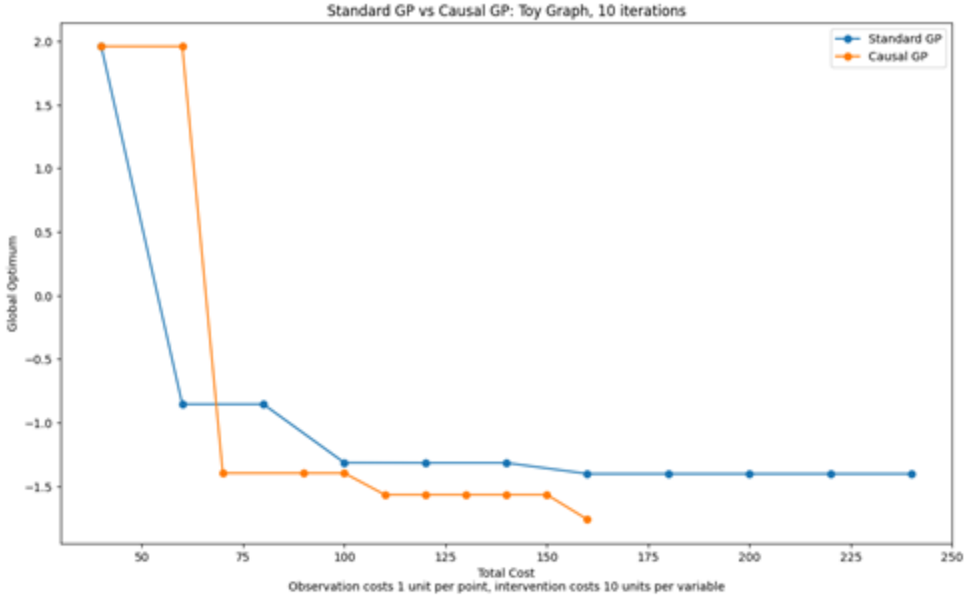
$$k_C(\mathbf{x}_s, \mathbf{x}'_s) = k_{RBF}(\mathbf{x}_s, \mathbf{x}'_s) + \sigma(\mathbf{x}_s)\sigma(\mathbf{x}'_s) \text{ where } \sigma(\mathbf{x}_s) = \sqrt{\hat{\mathbb{V}}(Y|\mathrm{do}(\mathbf{X}_s = \mathbf{x}_s))}$$

CausalBO - Results

$$X = \epsilon_X$$

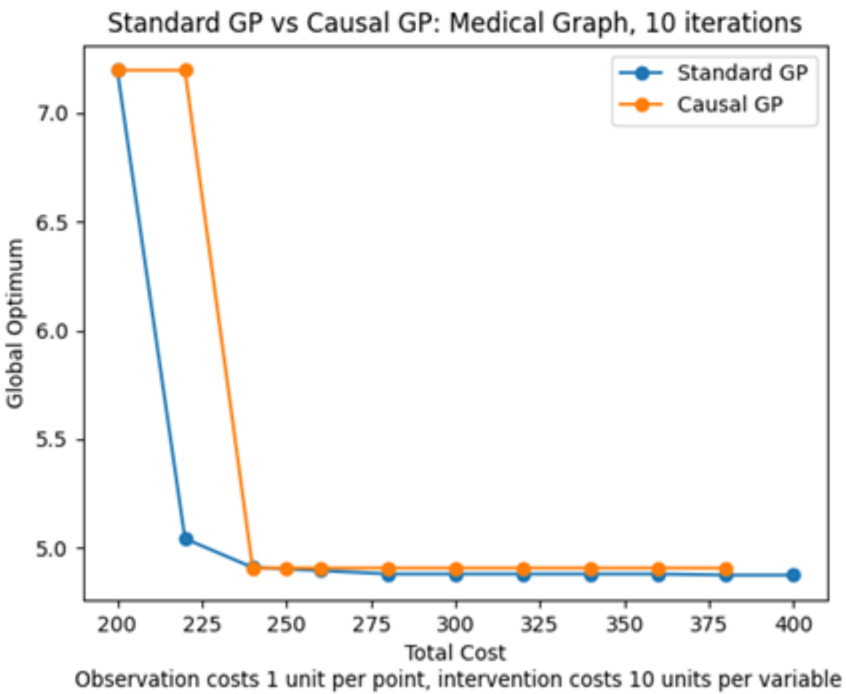
$$Z = \exp(-X) + \epsilon_Z$$

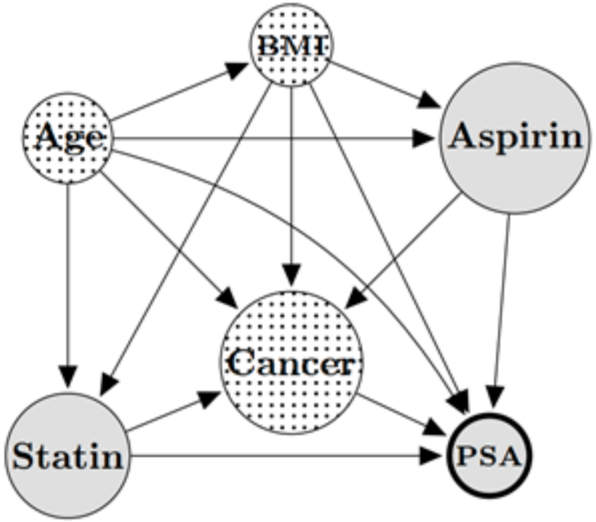
$$Y = \cos(Z) - \exp(-\frac{Z}{20}) + \epsilon_Y$$



Optimal set-value pair in paper: ({Z}, [-3.2])
Optimal set-value pair using CausalBO: ({Z}, [-2.7])

CausalBO - Results

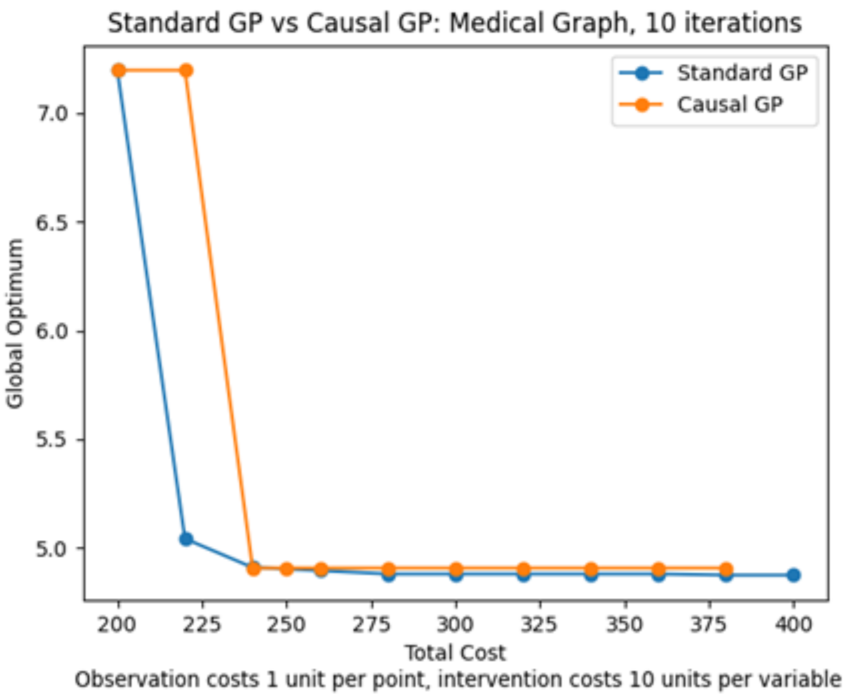




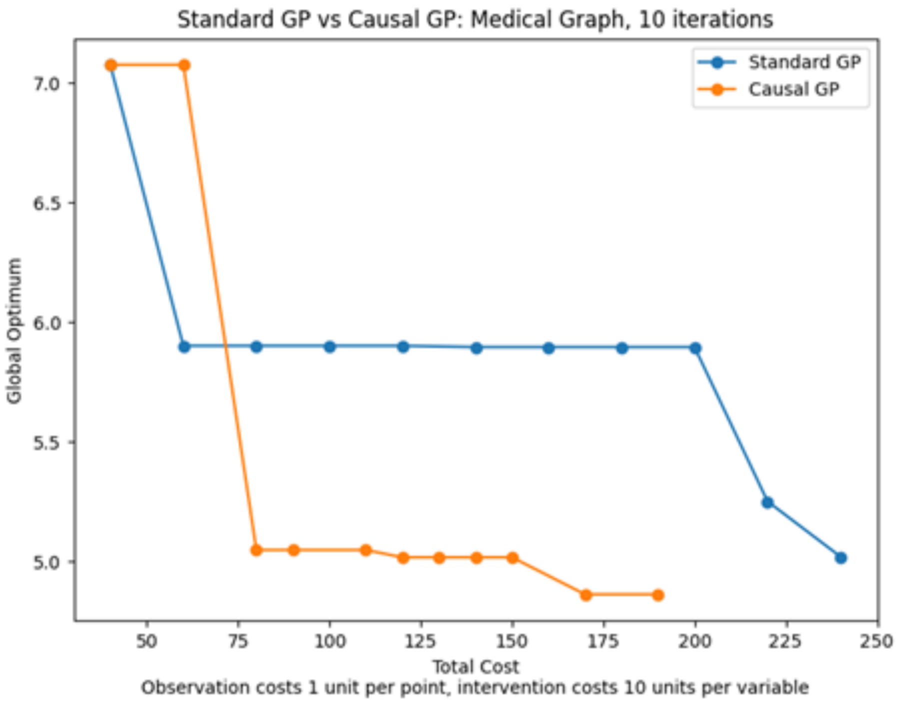
$$
\begin{aligned}
age &= \mathcal{U}(55, 75) \\
bmi &= \mathcal{N}(27.0 - 0.01 \times age, 0.7) \\
aspirin &= \sigma(-8.0 + 0.10 \times age + 0.03 \times bmi) \\
statin &= \sigma(-13.0 + 0.10 \times age + 0.20 \times bmi) \\
cancer &= \sigma(2.2 - 0.05 \times age + 0.01 \times bmi - 0.04 \times statin + 0.02 \times aspirin) \\
Y &= \mathcal{N}(6.8 + 0.04 \times age - 0.15 \times bmi - 0.60 \times statin + 0.55 \times aspirin + 1.00 \times cancer, 0.4)
\end{aligned}
$$

Optimal set-value pair in paper: ({aspirin, statin}, [0.0, 1.0])
Optimal set-value pair using CausalBO: ({aspirin, statin}, [0.02, 0.97])

CausalBO - Results



N = 200



N = 40

CausalBO - Future Work

- Add multithreading support for faster calculations.
- Switch causality backend from DoWhy to Ananke.
    - Ananke employs more general causal effect estimation methods - should fix the convergence issue.
- Add support for multiple information sources.
    - Will allow for estimation using information from multiple sources which models the same process, but obtained using different procedures.

References

V. Aglietti et. al., “Causal Bayesian Optimization,” *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, vol. 108, 2020.

V. Aglietti et. al., “Multi-task Causal Learning with Gaussian Processes,” *34th Conference on Neural Information Processing Systems*, 2020.