Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models

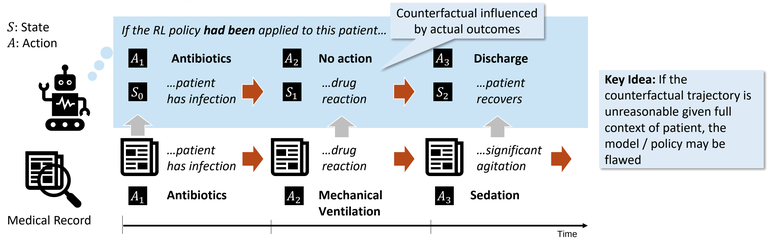
Michael Oberst, David Sontag
MIT

Motivation and background

For reinforcement learning (RL) in high-risk settings (e.g., healthcare), we propose to use *counterfactual trajectories* as an additional method of "sanity-checking" the resulting policy

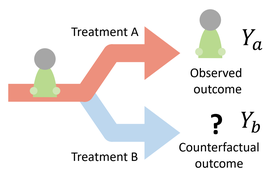
Counterfactuals as "sanity-checking":

- Inspired by interest in applying RL to healthcare, e.g., sepsis management[1]
  - Off-policy evaluation can give incorrect results, e.g., due to confounding, small sample sizes, poorly specified rewards, etc[2]
  - How can we "sanity-check" RL policies with domain experts, especially if we cannot visualize the policy directly?
- Idea: Streamline review of individual trajectories, by providing anticipated counterfactuals (e.g., per the MDP used to learn the policy)



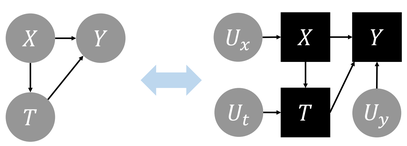
Estimating counterfactuals with Structural Causal Models (SCM)

- To generate individual-level counterfactuals:
  - *Infer* posterior over exogenous variables
  - *Intervene* to reflect alternative actions
  - *Predict* counterfactuals from the posterior
- **Identifiability Problem:** Multiple SCMs can replicate the same interventional distribution, but imply different counterfactuals



Example: Monotonicity assumption for binary outcomes

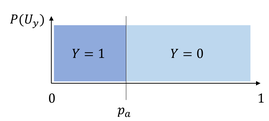
Treatment A was given, and we observed $Y_a = 1$. What **would have happened** if Treatment B had been given?



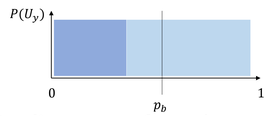
$U_y \sim Unif(0,1)$,
$Y_t = 1\{U_y \leq p_t\}$
where $p_t := E[Y \mid do(T = t), X]$

This SCM has the **monotonicity** property (Pearl 2000[3]), which identifies counterfactuals in the binary case

1 **Infer** the posterior of U_y given $X, Y_a = 1$



2 **Intervene** to set $T = b$



3 **Predict** counterfactual outcome

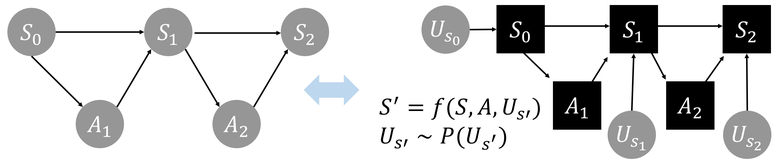
$P(U_y \leq p_b \mid U_y \leq p_a) = 1$
implies $Y_b = 1$

Generating counterfactual trajectories

Given a discrete (PO)MDP, policy, and real trajectories, we generate counterfactuals using an additional assumption, the *Gumbel-Max SCM*, based on a notion of *counterfactual stability*

Discrete (PO)MDP can be reformulated as an SCM

To generate counterfactuals in a discrete (PO)MDP, we will need a categorical SCM to represent the transition and reward distributions, e.g.,



Benefit: Decompose expected reward across real episodes

Lemma 1 (Simplified): Decomposition of expected reward

Let trajectories τ be drawn from $p(\tau)$ under the behavior policy and a given SCM. Let $\tau_\pi(u)$ be a counterfactual trajectory, as a deterministic function of exogenous U terms in the SCM and new policy π. Then:

$$E_\pi[R(\tau)] = \int p(\tau) E_{u \sim p(u|\tau)}\left[R\left(\tau_\pi(u)\right)\right] d\tau$$

Issue: Non-identifiability of categorical SCMs from data

Counterfactual Question

Given $S' = 2,\ S = s,\ A = a$; What would have happened if $A = a'$?

Interventional Distribution

S'	$p(S' \mid s, a)$	$p(S' \mid s, a')$
1	0.25	↓ 0
2	0.25	0.25
3	0.3	↓ 0.25
4	0.2	↑ 0.5

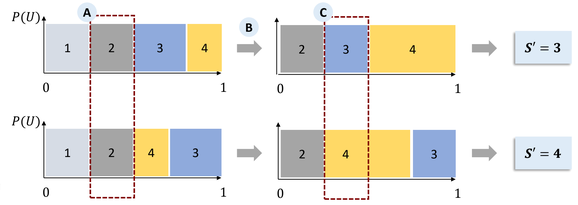
Example SCM(s): Order outcomes on (0, 1), and sample from a uniform distribution – but **order matters!**

Procedure

A **Infer** the posterior of U_y **B** **Intervene** to set $A = a'$ **C** **Predict** counterfactual

Result

Different based on the ordering!



Resolution: Counterfactual Stability and Gumbel-Max SCM

We introduce the property of "counterfactual stability" for categorical SCMs, inspired by the monotonicity property for binary variables, and show that our proposed Gumbel-Max SCM satisfies this property

Counterfactual Stability (Informal):

Given $S' = i$ *under* $A = a$, *then for all* $j \neq i$, $\frac{p_i'}{p_i} \geq \frac{p_j'}{p_j} \rightarrow S' \neq j$ *under* $A = a'$

Theorem 1: Counterfactual stability implies monotonicity (Pearl, 2000[3]) when there are only two categories

Gumbel-Max SCM:

Sample via the *Gumbel-Max trick*[4], with Gumbel variables as exogenous terms:

$g_j \sim Gumbel$
$S' = argmax_j \{\log P(S' = j \mid S, A) + g_j\}$

Theorem 2: Gumbel-Max SCM satisfies counterfactual stability

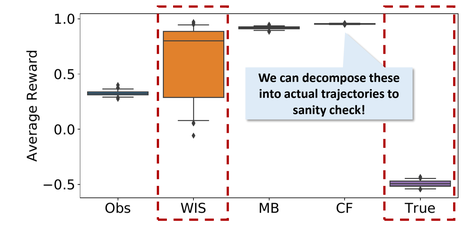
Illustrative application: Sepsis management

Using a synthetic example, we demonstrate how our approach *can highlight flaws in the policy / MDP*, even when other quantitative off-policy methods are overly optimistic

Overview of synthetic case study

- **Simulator of sepsis management** as an MDP, including discrete states (heart rate, BP, glucose, etc.) and actions (on/off antibiotics, ventilator, etc.). Time-independent state of diabetes. Reward is -1 for death, +1 for discharge.
- **Behavior policy is excellent:** Data generated using illustrative "optimal" physician policy, learned using policy iteration on true MDP
- **RL policy (small sample size + confounding):** Transition / reward distribution learned based on 1000 observed trajectories, and used to train RL policy using policy iteration, while **diabetes / glucose not observed**

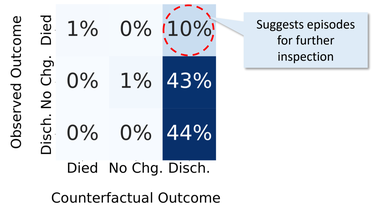
Off-policy estimates of reward are misleading in this case



Obs: Observed Reward of behavior policy
WIS: Weighted Importance Sampling
MB: Model-Based Rollouts
CF: Counterfactual Rollouts
True: Actual RL reward, not known

Counterfactuals help identify episodes to examine

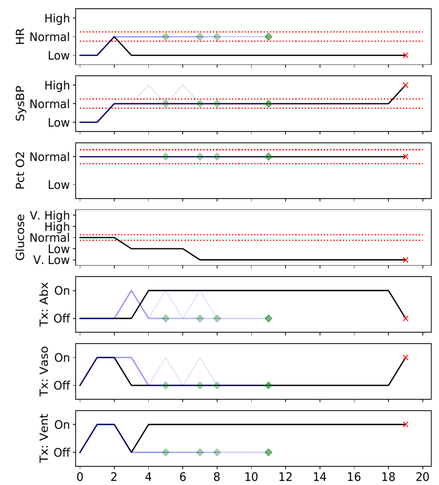
To identify episodes for further inspection, we choose those where the SCM implies that "**this patient, who died, would have lived if the RL policy were applied**"



Examination of individual trajectories reveals flaws

For one of these patients, we observe:

- Policy would have **stopped treatment** in all counterfactual trajectories and **expected a speedy discharge** from the hospital
- **This reveals a "bug" in our model!** Given access to full medical record, we note that glucose was out of range, and stopping treatment would have been dangerous



Red dotted lines = normal range
Black line = observed trajectory (ends in death at time 19)
Blue lines = counterfactual trajectories (end in discharge)

Green diamonds = "Discharge" event
Red cross = "Death" event

Recall, glucose is excluded from model, so there is no counterfactual trajectory, just the actual trajectory for this variable

1. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24(11): 1716–1720, 2018
2. Gottesman, O., Johansson, F., Komorowski, M., Faisal, A.,Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. Nature Medicine, 25(1):16–18, 2019
3. Pearl, J. Probabilities of Causation: Three counterfactual interpretations and their identification. Synthese, 121(1):93–149, 2000
4. Maddison, C. J., Mnih, A., and Teh, Y. W. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In International Conference on Learning Representations (ICLR), 2017

Code / notebooks to replicate synthetic experiment: www.github.com/clinicalml/gumbel-max-scm