NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2023

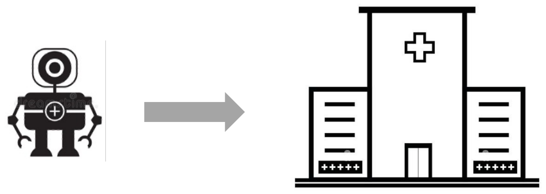
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Counterfactual-Augmented Importance Sampling for Semi-Offline Policy Evaluation

Shengpu Tang, Jenna Wiens
Computer Science & Engineering
University of Michigan

This work is funded by NIH-NLM.

Evaluating RL Policies in Healthcare

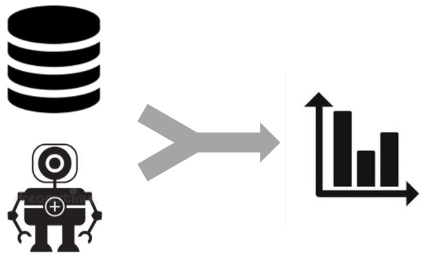


Online Evaluation

High-stakes environment
- Potentially unsafe to patients
- Disruptive to human users and clinical workflows

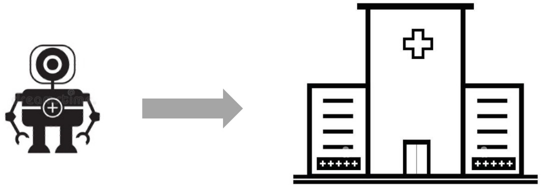
Wiens et al. "Do no harm: a roadmap for responsible machine learning for health care." *Nature Medicine* 2019.

Evaluating RL Policies in Healthcare



Offline Evaluation

?



Online Evaluation

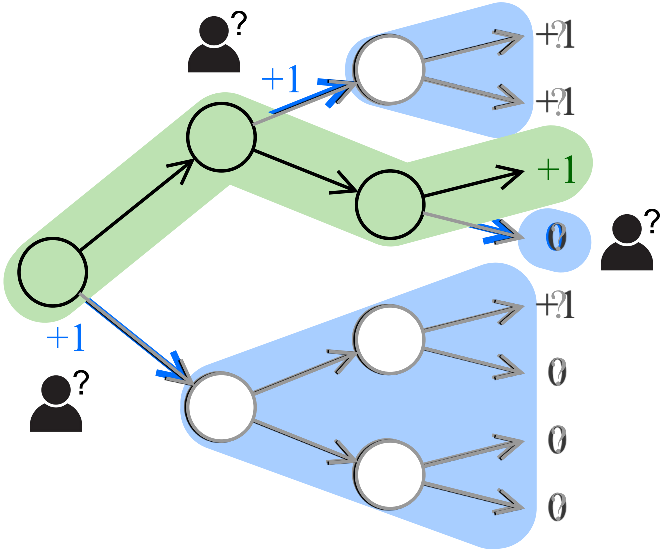
Observational dataset
- Limited by available data
- May not reflect distribution shift induced by new policies

High-stakes environment
- Potentially unsafe to patients
- Disruptive to human users and clinical workflows

Wiens et al. "Do no harm: a roadmap for responsible machine learning for health care." *Nature Medicine* 2019.
Gottesman et al. "Guidelines for reinforcement learning in healthcare." *Nature Medicine* 2019.

Our Contributions

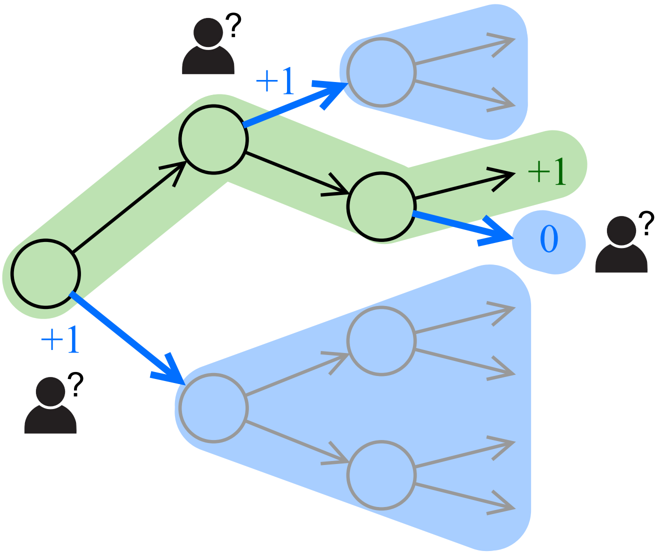
We propose a **semi-offline evaluation scheme** that combines observational data with **human annotations** of counterfactuals



Observational data contains **factual trajectories**

Query domain experts for **annotations** of the **counterfactual trajectories**

Augmenting Factual Data with Counterfactuals



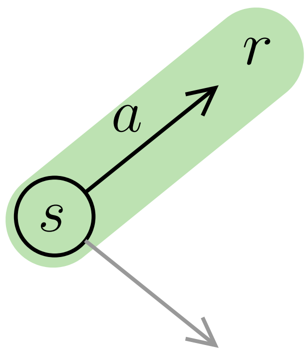
Intuition: as if we collected **more data**

How do we use both **counterfactual annotations** and **observational data** to evaluate policies?

> *"Simply adding annotations as new data"*
> … is not theoretically valid.

Key Idea: Augmenting Standard IS

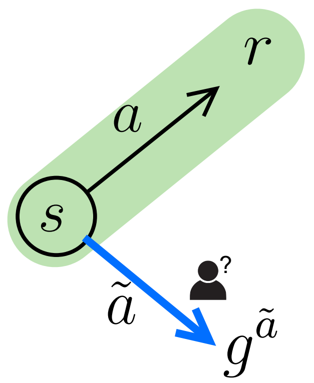
where $\rho = \frac{\pi_e(a|s)}{\pi_b(a|s)}$



$$\hat{v}^{\mathrm{IS}} = \rho\, r$$

Key Idea: Reweighted IS with Counterfactuals

where $\rho^{\tilde{a}} = \frac{\pi_e(\tilde{a}|s)}{\pi_b(\tilde{a}|s)}$



$$\hat{v}^{\text{C-IS}} = w^a \underbrace{\boldsymbol{\rho^a r}}_{\text{factual IS estimate from observed reward}} + \sum_{\tilde{a}\in\mathcal{A}\setminus\{a\}} w^{\tilde{a}} \underbrace{\boldsymbol{\rho^{\tilde{a}} g^{\tilde{a}}}}_{\text{counterfactual IS estimates based on annotations}}$$

where $w^a + \sum_{\tilde{a}\in\mathcal{A}\setminus\{a\}} w^{\tilde{a}} = 1$

Theoretical Insights

See paper for details

$$\hat{v}^{\text{C-IS}} = w^a \rho^a r + \sum_{\tilde{a} \in \mathcal{A} \setminus \{a\}} w^{\tilde{a}} \rho^{\tilde{a}} g^{\tilde{a}}$$

Intuition: as if we collected **more data**

- More data for regions that lack support → reduce bias
- Even more data for regions with support → reduce variance

C-IS can achieve **lower bias** and **lower variance** than IS

Experimental Results

Experiments conducted on the sepsis simulator

Based on the sepsis simulator introduced by Oberst & Sontag, ICML 2019.

Simulate collection of
- Factual dataset
- Counterfactual annotations

to evaluate multiple treatment policies.

Compare
- Standard approach (PDIS)
- Proposed approach (C-PDIS)

Metrics
- ↓ Evaluation error (RMSE)
- ↑ Ranking ability (Spearman correlation)

with respect to ground-truth policy performance

Experimental Results

Estimator	↓ Evaluation Error	↑ Ranking Ability
Baseline	0.113	0.596
Proposed	0.013	0.995

Our proposed approach **outperforms** the baseline method (without annotations) in terms of all metrics.

(under the assumption that annotations are "good")

Experimental Results

Estimator	↓ Evaluation Error	↑ Ranking Ability
Baseline	0.113	0.596
Proposed	0.013	0.995
Proposed (biased)	0.028	0.979
Proposed (noisy)	0.029	0.977
Proposed (missing)	0.067	0.823

Our proposed approach **remains competitive** to the baseline method even with imperfect annotations (biased, noisy, missing).

Takeaways

We propose a **new estimator** for **semi-offline evaluation** that combines observational data with **human annotations** of counterfactuals

$$\hat{v}^{\text{C-IS}} = w^a \rho^a r + \sum_{\tilde{a} \in \mathcal{A} \setminus \{a\}} w^{\tilde{a}} \rho^{\tilde{a}} g^{\tilde{a}}$$

- Theoretical insights show potential to ***reduce both bias and variance***
- Experiments demonstrate robustness to ***bias, noise, and missingness*** of annotations

🔗 https://github.com/MLD3/CounterfactualAnnot-SemiOPE