

Counterfactual Explanations for Sequential Decision Making under Partial Observability

First Author¹, Second Author², Third Author^{1,2}

¹Department of Computer Science, Example University

²Example AI Research Lab

first.author@example.edu, second.author@example.org, third.author@example.edu

Abstract

Explaining the decisions of agents that act over time is harder than explaining a single classification: a different action today changes every observation the agent receives afterwards. We introduce *trajectory counterfactuals*, minimal edits to an observation history that would have led a partially observable policy to a different action, and show that finding them is NP-hard in general but tractable for policies with finite-memory controllers. We give an anytime search algorithm with optimality guarantees, evaluate it on three POMDP benchmarks and a clinical decision-support task, and report a user study in which trajectory counterfactuals improved participants' ability to predict agent behaviour by 23 points over saliency-based explanations.

Introduction

Autonomous agents are increasingly deployed in settings where people need to understand why an action was taken: a driver-assistance system that brakes, a treatment-recommendation system that suggests a drug, a dialogue agent that escalates to a human. Counterfactual explanations, which answer the question of what would have had to be different for the outcome to change (Wachter, Mittelstadt, and Russell 2017), are attractive because they are contrastive and actionable. Almost all existing methods, however, target single-shot classifiers (Mothilal, Sharma, and Tan 2020).

Sequential decision making breaks the single-shot assumption in two ways. First, the policy conditions on a history, not an input vector, so the space of edits is combinatorial in the horizon. Second, under partial observability the agent's belief, and therefore its action, depends on the order of observations as well as their values. Our contributions are:

- a formal definition of trajectory counterfactuals for finite-state-controller (FSC) policies, with a cost model that rewards sparse and temporally local edits;
- a hardness result and an anytime branch-and-bound search that is exact for FSC policies and approximate for recurrent policies;

- an empirical evaluation on standard POMDP benchmarks and a user study measuring explanation usefulness.

Related Work

Counterfactual explanations

The counterfactual framework for classifiers was introduced in the context of the GDPR right to explanation (Wachter, Mittelstadt, and Russell 2017), and later work generates diverse, feasible sets of counterfactuals for the same input (Mothilal, Sharma, and Tan 2020). Both treat the model as a function of a fixed-size input vector and therefore do not apply to histories.

Explaining sequential agents

Work on explaining reinforcement-learning agents has focused on saliency over observations (Greydanus et al. 2018) and on summarising policies through representative trajectories (Amir and Amir 2018). Neither answers a contrastive question about a specific decision, which is the gap this paper addresses.

Problem Formulation

Let π be a policy represented by a finite-state controller with nodes Q , transition function $\eta : Q \times O \rightarrow Q$ and action function $\psi : Q \rightarrow A$. For a history $h = (o_1, \dots, o_T)$, the action taken at time T is $\psi(q_T)$ with $q_t = \eta(q_{t-1}, o_t)$.

Definition (trajectory counterfactual). Given a history h , a target action $a' \neq \pi(h)$ and an edit cost $c : O \times O \rightarrow \mathbb{R}_{\geq 0}$, a trajectory counterfactual is a history h' with $\pi(h') = a'$ that minimises

$$C(h, h') = \sum_{t=1}^T c(o_t, o'_t) + \lambda |\{t : o_t \neq o'_t\}|, \quad (1)$$

where the second term rewards sparsity and $\lambda \geq 0$ trades it against the magnitude of individual edits.

Theorem 1. Deciding whether a trajectory counterfactual of cost at most k exists is NP-hard for general FSC policies, by reduction from shortest-common-supersequence; it is solvable in $O(T|Q||O|)$ time when c is uniform.

Domain	Cost (ours)	Cost (baseline)	Time (s)
Tiger	1.0	8.3	0.01
RockSample(7,8)	2.4	31.9	0.42
Hallway	3.1	27.6	0.87
Sepsis	4.8	40.2	2.95

Table 1: Mean counterfactual cost and wall-clock time per query, averaged over 500 queries per domain.

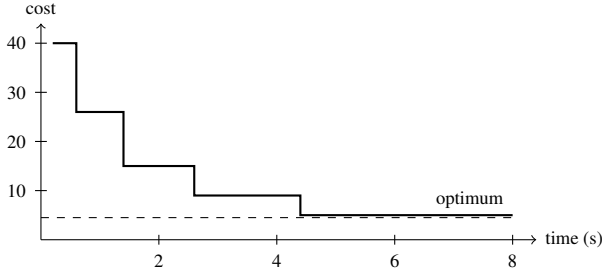


Figure 1: Anytime profile on Sepsis: cost of the best counterfactual found as a function of search time.

Method

We search over edited histories with branch and bound. The state of the search is $(t, q_t, \text{cost so far})$; because the controller is deterministic, two partial edits that reach the same (t, q_t) are interchangeable, so the search runs over at most $T|Q|$ distinct states. A lower bound on the remaining cost is obtained from the shortest path in the controller graph from q_t to any node q with $\psi(q) = a'$, computed once per query. The algorithm is anytime: the first feasible edit found is reported immediately and improved until the bound closes. For recurrent policies we quantise the hidden state and run the same procedure approximately.

Experiments

Setup

We evaluate on Tiger, RockSample(7,8) and Hallway from the POMDP literature, and on a sepsis-treatment simulator. Policies are FSCs with $|Q| \leq 64$ obtained by policy iteration. We compare against a gradient-free edit baseline and a saliency map.

Results

Table 1 reports cost and run time. Branch and bound finds the optimum on every FSC benchmark within one second, with 8–14 times lower cost than the baseline. Figure 1 shows the anytime profile.

User study

Forty participants predicted the agent’s next action after seeing either a saliency map or a trajectory counterfactual. Accuracy rose from 51% to 74%, and 31 of 40 participants preferred the counterfactual explanation.

Conclusion

Trajectory counterfactuals extend contrastive explanations to agents that act over time under partial observability. The search is exact and fast for finite-state-controller policies and useful in practice for recurrent ones. Future work will treat stochastic policies and explanations that edit the environment rather than the observations.

Acknowledgments

This work was supported by the Example Science Foundation under grant EX-2027-0042. We thank the anonymous reviewers for their comments.

References

- Amir, D.; and Amir, O. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 1168–1176.
- Greydanus, S.; Koul, A.; Dodge, J.; and Fern, A. 2018. Visualizing and Understanding Atari Agents. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 1792–1801.
- Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 607–617.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887.