Can Active Sampling Reduce Causal Confusion in Offline Reinforcement Learning?

TOYOTA RESEARCH INSTITUTE

OATML

Gunshi Gupta, Tim G. J. Rudner, Rowan Thomas McAllister, Adrien Gaidon, Yarin Gal

Motivation

- An agent is said to be **causally-confused** when it learns a policy reflecting spurious correlations in the data. Such a policy may appear to be optimal during training but fail catastrophically at deployment.
- We investigate whether actively sampling points from the dataset may enable offline RL agents to alleviate causal confusion in offline reinforcement learning, and produce a safer model for deployment.

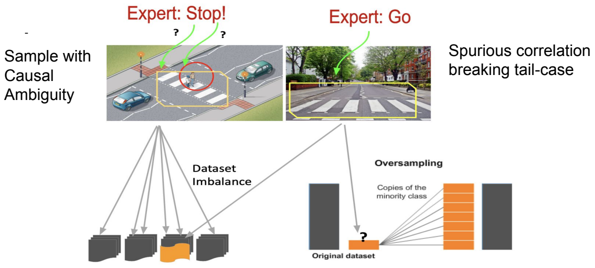

Expert: Stop!
Expert: Go
Sample with Causal Ambiguity
Spurious correlation breaking tail-case
Dataset Imbalance
Oversampling
Copies of the minority class
Original dataset

Figure depicting the causal confusion problem.

Background

Causal Confusion happens most often in the offline learning setup, since the agent can't do interventions to correct it's beliefs.
We consider causal confusion the offline RL setup.

Offline RL: Learn an optimal policy by learning the Q-function from state-action pairs in offline dataset.

$$\mathcal{L}_{\text{critic}}^{\text{CQL}}(\theta) = \frac{1}{2} \mathop{\mathbb{E}}_{(s,a,s')\sim\mathcal{D}} \left[\left(Q_\theta(s,a) - \mathcal{B} Q_{\bar{\theta}}(s,a) \right)^2 \right] \quad \text{TD-Error}$$

$$\text{Conservatism penalty} \quad + \alpha_0 \mathop{\mathbb{E}}_{s\sim\mathcal{D}} \left[\log \sum_a \exp Q_\theta(s,a) - \mathop{\mathbb{E}}_{a\sim\pi_\beta} [Q_\theta(s,a)] \right]$$

Causal Inference: Treatment effect estimation for a covariate x subjected to a treatment (t=1). Measured through CATE:

$$\tau(X) \equiv \mathbb{E}[Y \mid x, t=1] - \mathbb{E}[Y \mid x, t=0].$$

Treating the outcome function as the Q-function, the advantage function can be connected to CATE, since it is estimating **the relative effect of action *a* on state *s*.** Defining the CATE estimator similar to the previous equation with the outcome function being the *expected return* in this case we get:

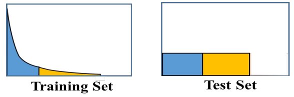
$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\left(\mathbf{s}_t, \mathbf{a}_t\right) \mid s_t = s, a_t = a\right] - \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\left(\mathbf{s}_t, \mathbf{a}_t\right) \mid s_t = s\right]$$
$$= Q(s,a) - V(s) = A(s,a),$$

Active Sampling: Aims to acquire (labelled) datapoints from the dataset where the model is most 'uncertain'. We acquire points with

- high TD-Error, or
- high variance of the advantage estimates.

Experiments

- We design vision-based long-tailed datasets where most expert-like transitions can be explained by spurious correlations, and a small minority that can not and require the model to learn a sensible policy.
- We construct offline datasets by collecting expert-like trajectories in the following environments:
  - ○ Traffic-World (gridworld with traffic and a traffic light)
  - ○ Maze from Procgen: navigation to randomly-sampled goals
  - ○ Atari car-racing game: Enduro

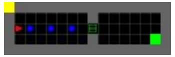


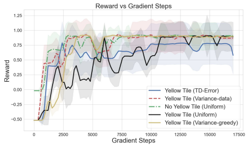
Policies learnt on long-tailed data are evaluated on a uniform sample of environment from each scenario (test-set is not long-tailed). The datasets' composition can be described as follows:

TRAFFIC-WORLD

🟦 90% cases where the agent is stuck behind the leading vehicle and can imitate it by following the tail-light (simulated by a yellow-tile)

🟨 Cases where the yellow-tile isn't predictive of the optimal action



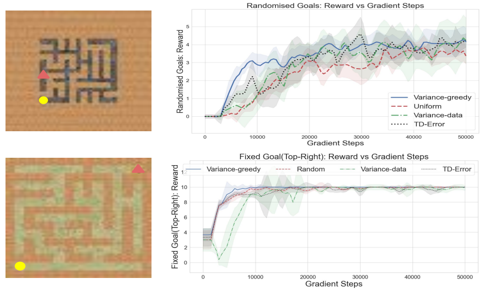


Left: Sample input for Traffic-world, Right: Figures showing reward curves for uniform and active sampling agents in **Traffic-World.**

MAZE

🟦 90% Goals sampled at the top-right

🟨 10% Goals sampled randomly in the maze

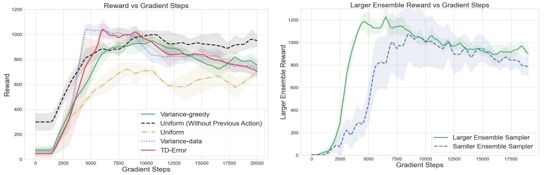


Figures showing sample inputs and reward curves for uniform and active sampling agents on **Maze** in environments with random-sampled goals (top) and fixed goals (bottom)

ENDURO

🟦 (100-x)% cases where action is predictable from the previous action

🟨 x% cases with complex actions not predictable from previous action



Left: Sample inputs for **Enduro**.
Top left: Figures showing reward curves for uniform and active sampling agents in **Enduro. Top Right:** Effect of uncertainty quantification (bigger v/s smaller ensemble) on sampling quality.

We provide empirical evidence that uniform and active sampling techniques are able to consistently reduce causal confusion as training progresses and that active sampling is able to do so significantly more efficiently than uniform sampling.