Decision-Point Guided Safe Policy Improvement

HARVARD UNIVERSITY
Imperial College London
Abhishek Sharma, Leo Benac, Sonali Parbhoo, Finale Doshi-Velez

AISTATS
NSF NIH National Institutes of Health

INTRODUCTION

Motivation

- Batch RL is important in risky/expensive real-world applications (medicine, robotics, etc.)
- Safe policy improvement (SPI) is critical to ensure new policies perform at least as well as the existing ones.
- Existing SPI methods: density-based policy regularization, pessimism-based approaches are too conservative, SPIBB[1] needs π_b

Contributions

- Introduce Decision-Point RL (DPRL), limiting policy deviations to "decision points"
- DPRL does not require access to the behavior policy during training
- Tighter theoretical guarantees for SPI (dependent on visitation frequency)

Setup

- Assume MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, with discrete $\mathcal{A}$ and bounded $R \in [0, R_{\max}]$
- Given: Dataset $\mathcal{D} = \{S_0^n, A_0^n, R_0^n, \ldots, S_{T_n}^n, A_{T_n}^n, R_{T_n}^n\}_{n=1}^N$ with actions taking using Behavior Policy π_b

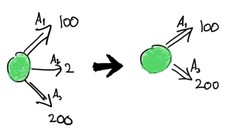
DECISION-POINT REINFORCEMENT LEARNING (DPRL)

Step 1

Restrict Policy Space

$$\Pi^{\text{DP}} = \{\pi : \pi(a|s) = 0 \quad \forall a \notin \mathcal{A}_s^{\text{DP}}\}$$
$$\mathcal{A}_s^{\text{DP}} = \{a \in \mathcal{A} : n(s,a) \geq N_\wedge \text{ and } \hat{Q}^{\pi_b}(s,a) \geq \hat{V}^{\pi_b}(s)\}$$

1. Exclude infrequent actions

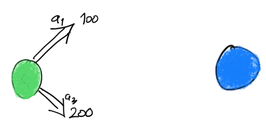


2. Exclude actions with lower Q-value than π_b

Step 2

Identify Decision Points

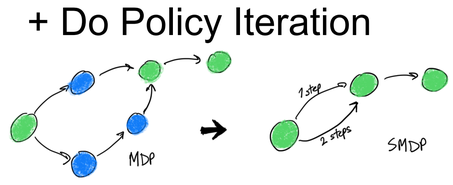
$$\mathcal{S}^{\text{DP}} = \{s \in \mathcal{S} : \mathcal{A}_s^{\text{DP}} \neq \emptyset\}$$



DP State Not a DP State

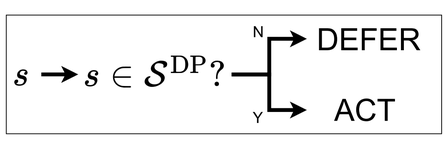
Step 3

Build Elevated SMDP* + Do Policy Iteration



Step 4

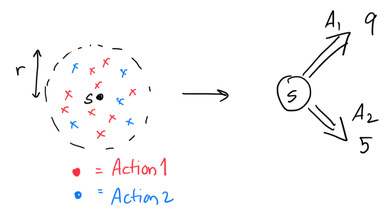
Build DP Policy with Deferral



* SMDP: Semi-Markov Decision Process

Continuous States

Compute neighbors within radius r



Takeaways

1. Functional form of π_b not required
2. We change π_b only where we have data

THEORETICAL RESULTS

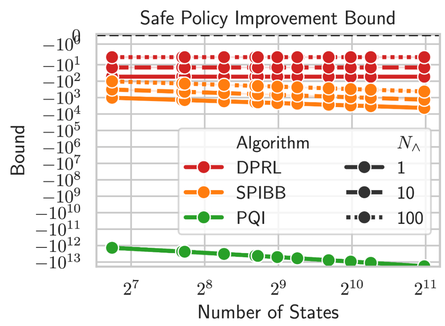
Discrete States

$$\rho(\pi_{\text{DP}}) - \rho(\pi_b) \geq -\frac{V_{\max}}{1-\gamma}\sqrt{\frac{1}{N_\wedge}\log\frac{C(N_\wedge)}{\delta}}$$

$C(N_\wedge)$ → $\#(s,a)$ with $\geq N_\wedge$ visits

Bound scales with visited pairs $C(N_\wedge)$, not $|\mathcal{S}||\mathcal{A}|$

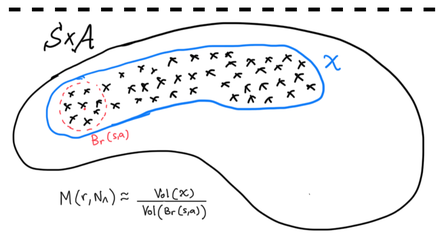
Raise $N_\wedge \rightarrow$ tighter bound but fewer decision points



Continuous States

$$\rho(\pi_{\text{DP}}) - \rho(\pi_b) \geq -\frac{V_{\max}}{1-\gamma}\sqrt{\frac{1}{2N_\wedge}\log\frac{M(r,N_\wedge)}{\delta}} - 3\epsilon_r$$

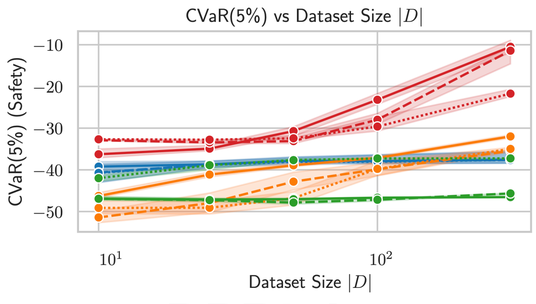
Number of balls $B_r(s,a)$ to cover the neighborhood of all "dense" s-a pairs



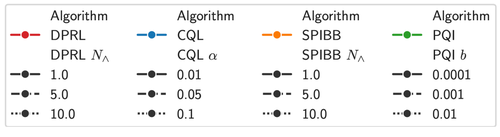
Both bounds are tighter when π_b only visits a slice of the $(s-a)$-space

EXPERIMENTAL RESULTS & CONCLUSION

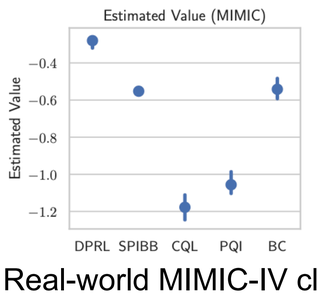
DPRL learns safer policies



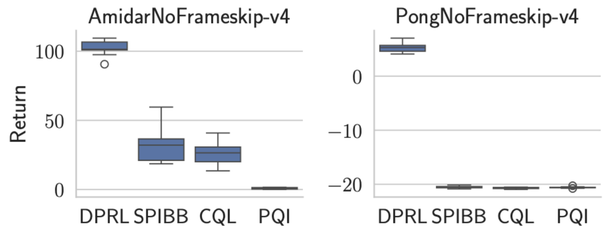
DPRL consistently safer (higher CVaR) than SPIBB[1], CQL[2], PQI[3]



DPRL learns performant policies

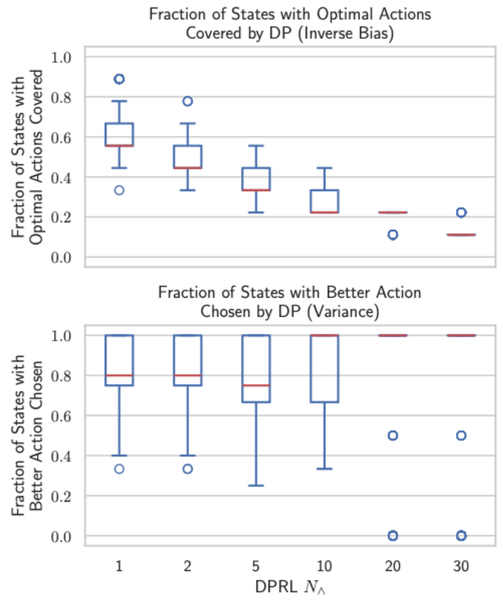


Real-world MIMIC-IV clinical dataset with hypotensive patients



Atari Environments with suboptimal behavior (see paper for details)

$N_\wedge$ controls the bias-variance trade-off



Fraction of states where the optimal action is allowed decreases, leading to an increase in bias.

Fraction of states where an action with higher value is chosen increases.
This suggests reduced variance in value estimation of valid state-actions because we have more observations to estimate the value

References
[1] Laroche, Romain, Paul Trichelair, and Remi Tachet Des Combes. "Safe policy improvement with baseline bootstrapping." International conference on machine learning. PMLR, 2019.
[2] Kumar, Aviral, et al. "Conservative q-learning for offline reinforcement learning." Advances in neural information processing systems 33 (2020): 1179-1191.
[3] Liu, Yao, et al. "Provably good batch off-policy reinforcement learning without great exploration." Advances in neural information processing systems 33 (2020): 1264-1274.