Safe Active Exploration

Iason Chalas Gian Hess Alexander Spiridonov

ETH zürich

Motivation

- Measuring an unknown quantity over a Markov Chain is a typical problem.
- Access to **safe** state-action space

Goal: Use correlation to learn about **unsafe** state-action space

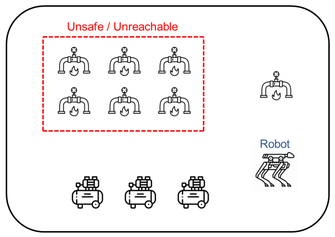


Figure 1. A robot measuring gas concentration at an industrial site. The gas concentration is correlated between pipes and compressors, respectively. Part of the site is restricted.

- **Example** in Figure 1: State-of-the-art methods do not consider correlation between unsafe and safe pipes.
- Resulting policy visits compressors more frequently than pipes (more safe compressors than pipes)
- This is sub-optimal for estimating gas concentration over the whole site.

Learning Problem

- **Known and deterministic MDP** $M = (\mathcal{S}, \mathcal{A}, P, H, f, s_0)$ with episode length H
- **Unknown quantity** f, part of a RKHS $f \in \mathcal{H}_k$ with known kernel $k((s,a),(s',a')) = \Phi(s,a)^T\Phi(s',a')$ and bound $\|f\|_{\mathcal{H}_k} \leq \frac{1}{\lambda}$
- **Fixed and known unsafe set** $\mathcal{B}_{unsafe} \subset \mathcal{S} \times \mathcal{A}$.
- **Noisy observations** of unknown quantity $y = f^T\Phi(s,a) + \epsilon \quad$ where: $\quad \epsilon \sim \mathcal{N}(0, \sigma^2)$
- **Fixed budget** of T trajectories of length H

Problem: Estimate a linear functional of the unknown function (potentially in an unsafe region).

$$\text{find} \quad \mathbf{C}\hat{f}_T \quad \text{where:} \quad \hat{f}_T = \arg\min_{f \in \mathcal{H}_k} \sum_{i=1}^{T} \sum_{(s,a) \in \tau_i} (f^T\Phi(s,a) - y_{i,s,a,})^2 + \lambda \|f\|_{\mathcal{H}_k}$$

[1] Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration, 2019.

[2] Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. 2023.

[3] Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps, 2022.

Method

As typical in Experimental Design, we minimize the second moment of the estimation error:

$$\mathbf{E}_T = \mathbb{E}_\epsilon[\mathbf{C}(f - \hat{f}_T)(f - \hat{f}_T)^T\mathbf{C}^T]$$

Safety:

- We define a **safe** state-action distribution polytope:

$$\mathcal{D}^{safe} := \{d | d(s,a) \geq 0, \sum_{s,a} d(s,a) = 1, \sum_{a} d(s',a) = \sum_{s,a} d(s,a)P(s'|s,a), \\ \forall (s,a) \in \mathcal{B}_{unsafe}. d(s,a) = 0\}$$

Current (Naive) Approach:

- Mutný et al. [2] derive convex upper bound for $\mathbf{E}_T$
- Their objective, defined over $\mathcal{D}^{safe}$ is the following:

$$\min_{d_\pi \in \mathcal{D}^{safe}} U(d_\pi) \quad \text{with:} \quad U(d_\pi) = s\left(\sum_{(s,a) \in \mathcal{S}\times\mathcal{A}} \frac{H d_\pi(s,a)}{\sigma^2_{s,a}} \Phi(s,a)\Phi(s,a)^T + \frac{\lambda \mathbf{I}}{T}\right)$$

Where $s(\cdot)$ is a convex scalarization function, such as $-\log(\det(\cdot))$

Note:

- All the features $\Phi(s,a)$ from the unsafe region cancel, since $d_\pi(s,a) = 0 \quad \forall (s,a) \in \mathcal{B}_{unsafe}$.
- Thus, we lose the correlation information to the unsafe region.

Reweighting-Based Approach (Ours):

- Introduce a reweighting $w : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to the objective

$$\min_{d_\pi \in \mathcal{D}^{safe}} U(d_\pi) \quad \text{with:} \quad U(d_\pi) = s\left(\sum_{(s,a) \in \mathcal{S}\times\mathcal{A}} \frac{H d_\pi(s,a) w(s,a)}{\sigma^2_{s,a}} \Phi(s,a)\Phi(s,a)^T + \frac{\lambda \mathbf{I}}{T}\right)$$

The weights are defined as follows:

$$w(s,a) = \sum_{(s',a') \in \mathcal{B}_{unsafe}} \left| \frac{\Phi(s,a)^T\Phi(s',a')}{\sqrt{\Phi(s,a)^T\Phi(s,a)}\sqrt{\Phi(s',a')^T\Phi(s',a')}} \right| + 1$$

- Weights transport information from unsafe to safe region.
- High weight $\Rightarrow$ High correlation to unsafe region $\Rightarrow$ Higher visitation

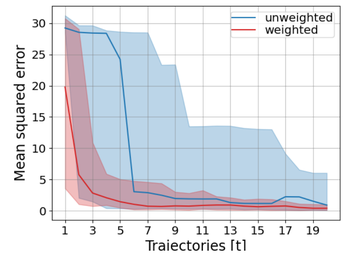
Optimization

- $\mathcal{D}^{safe}$ preserves convexity.
- $U(d_\pi)$ is also convex
- Thus, we can use the machinery of Convex RL [1, 3].
- We use the adaptive optimization scheme introduced by Mutný et al. [2].

Experiments

- Synthetic grid world with three different features (pictograms)
- Invariant correlation structure
- Unsafe state space in grey, empirical state visitation heatmap in red
- Comparison of mean squared estimation error over the whole domain

(a) **unweighted** objective (b) **weighted** objective



(c)

Figure 2

(a) Unweighted objective produces policy with more trajectories from below
(b) Weighted objective produces policy with more trajectories from top
(c) Mean squared estimation error over executed trajectories drops faster and more consistently with reweighting

Discussion

- Reweighting incentivizes visitation of the safe city state (highly correlated to unsafe region)

$\Rightarrow$ More trajectories from the top through safe city state

$\Rightarrow$ Better estimate for unsafe region and overall lower MSE

- Future experiments for more general correlation structure