Offline RL in Regular Decision Processes: Sample Efficiency via Language Metrics

Ahana Deb[1] Roberto Cipollone[2] Anders Jonsson[1] Alessandro Ronca[3] M. Sadegh Talebi[4]

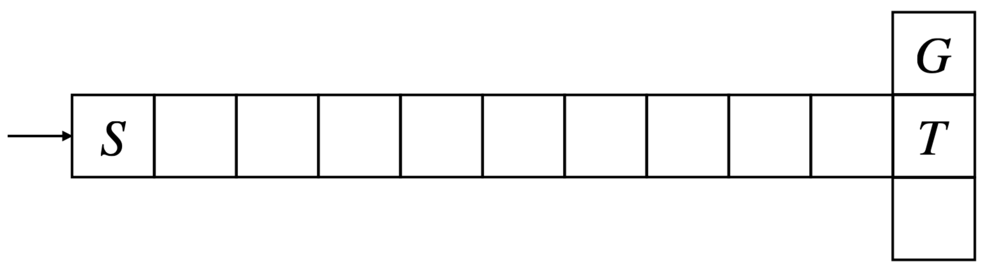
[1]Universitat Pompeu Fabra [2]Leonardo S.p.A. [3]University of Oxford [4]University of Copenhagen

Regular Decision Processes

Episodic Non-Markov Decision Process $\langle \mathcal{O}, \mathcal{A}, \mathcal{R}, \bar{T}, \bar{R}, H \rangle$, where $\bar{T} : (\mathcal{AO})^* \times \mathcal{A} \to \Delta(\mathcal{O})$ and $\bar{R} : (\mathcal{AO})^* \times \mathcal{A} \to \mathbb{R}$ are functions of the entire interaction history $(\mathcal{AO})^*$.

In a *Regular Decision Process (RDP)*, the functions $\bar{T}$ and $\bar{R}$ depend *regularly* on the interaction history, i.e., $\bar{T}$ and $\bar{R}$ can be represented by a Probabilistic-Deterministic Finite Automaton (PDFA).

Example. T-maze [1] with corridor length $N = 10$ and noisy observations.



The observation at the initial position S indicates the position of the goal G at the end of the corridor for the current episode.

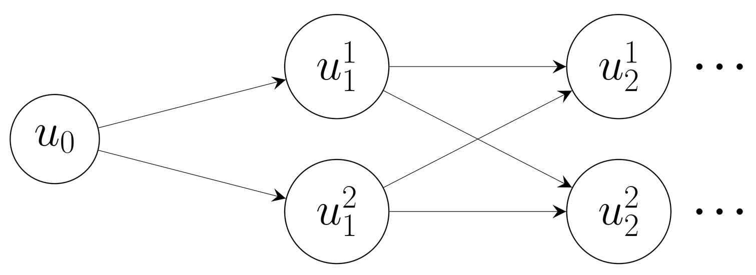
Objective

Given a dataset $\mathcal{D}$ of episodes, collected from an unknown RDP $\mathbf{R}$ and unknown behavior policy π^b, compute a near-optimal policy for $\mathbf{R}$, using the smallest $\mathcal{D}$ possible.

Fact: A near-optimal policy can be computed from the PDFA of the RDP.
Question: Can we *learn* the PDFA of an RDP from an interaction history?

AdaCT-H | State-of-the-art for PDFA Learning

AdaCT-H [3] computes the PDFA of the underlying RDP, and it enjoys a *polynomial sample complexity* in the problem parameters. It builds the *graph of transitions*, by comparing candidate states against the states already discovered.



AdaCT-H employs the *prefix distance*, defined as

$$L_\infty^{\mathsf{p}}(p_1, p_2) = \max_{u \in [0,\ell], e \in \mathcal{E}_u} |p_1(e\,*) - p_2(e\,*)|\,.$$

The *sample complexity* depends inversely on the L_∞^{p}*-distinguishability* which is the largest value μ_0 such that for each $p_1 \neq p_2$ on suffixes,

$$\mu_0 \le L_\infty^{\mathsf{p}}(p_1, p_2).$$

The **main bottleneck** of AdaCT-H is that L_∞^{p}-distinguishability tends to be exceedingly small. In our running example, the T-maze domain, the L_∞^{p}-distinguishability μ_0 *decreases exponentially* with the corridor length N.

$$\mu_0 \le L_\infty^{\mathsf{p}}(p_1, p_2) \le p_i(a_1 o_1 r_1\; a_2 o_2 r_2 \cdots a_N o_N r_N) \le 2^{-N}$$

Contribution

A practical implementation of AdaCT-H that reduces sample, memory, and time complexity, by means of the following innovations.

- Exploiting the **Count-Min-Sketch (CMS)** data structure to reduce the memory complexity of storing the empirical distributions on suffixes.
- A **novel language metric** $L_\mathcal{X}$, based on the theory of formal languages, and a **hierarchy of language families** that remove the dependency on L_∞^{p}-distinguishability and yields an *exponentially more sample efficient* algorithm in domains having low complexity in language-theoretic terms.

Language Metrics

Language metrics for distributions over strings

For $\mathcal{X}$ a class of languages, the corresponding *language metric* is

$$L_\mathcal{X}(p_1, p_2) \coloneqq \max_{X \in \mathcal{X}} |p_1(X) - p_2(X)| \quad \text{where} \quad p_i(X) \coloneqq \sum_{x \in X} p_i(x).$$

Language metrics for RL

Let $\mathcal{X}_{i,j}$ be a language family parameterized on two integers i and j

$i \in [\![3]\!]$: the number of elements in $\mathcal{A}$, $\mathcal{O}$ and $\mathcal{R}$ considered jointly

$j \in [\![t]\!]$: the length of subsequences considered for comparison

Examples:

$\mathcal{X}_{1,1}$: single instance of one action, one observation or one reward

$\mathcal{X}_{3,2}$: subsequence of two instances of triplets in $\mathcal{AOR}$

The language metric $L_{\mathcal{X}_{3,t}}$ subsumes L_∞^{p}

Simple metrics suffice for simple domains. E.g., in the T-Maze,

$$\mu_0 \ge L_{\mathcal{X}_{3,1}}(p_1, p_2) \ge p_i(*\,North, Goal/1\,*) = 1/2$$

Sample Complexity Analysis

Theorem 1. AdaCT-H$(\mathcal{D}, \delta)$ returns the minimal RDP $\mathbf{R}$ with probability at least $1 - 3AOU\delta$ when CMS is used to store empirical probability estimates, the statistical test is

$$L_\infty^{\mathsf{p}}(\mathcal{Z}_1, \mathcal{Z}_2) \ge \sqrt{8 \log(4(ARO)^{H-t}/\delta) / \min(|\mathcal{Z}_1|, |\mathcal{Z}_2|)},$$

and the dataset size is $|\mathcal{D}| \ge \widetilde{\mathcal{O}}\left(\frac{HC_\mathbf{R}^* \log(1/\delta)}{d_\mathsf{m}^* \cdot \mu_0^2}\right)$, with $d_\mathsf{m}^* = \min_{t, u_t ao} d_t^*(u_t, ao)$ the minimum occupancy of the optimal policy π^*.

Theorem 2. AdaCT-H$(\mathcal{D}, \delta)$ returns the minimal RDP $\mathbf{R}$ with probability at least $1 - 2AOU\delta$ when using the language metric $L_\mathcal{X}$ to define a statistical test

$$L_\mathcal{X}(\mathcal{Z}_1, \mathcal{Z}_2) \ge \sqrt{2 \log(2|\mathcal{X}|/\delta) / \min(|\mathcal{Z}_1|, |\mathcal{Z}_2|)},$$

and the size of the dataset satisfies $|\mathcal{D}| \ge \widetilde{\mathcal{O}}\left(\frac{C_\mathbf{R}^* \log(1/\delta) \log|\mathcal{X}|}{d_\mathsf{m}^* \cdot \mu_0^2}\right)$.

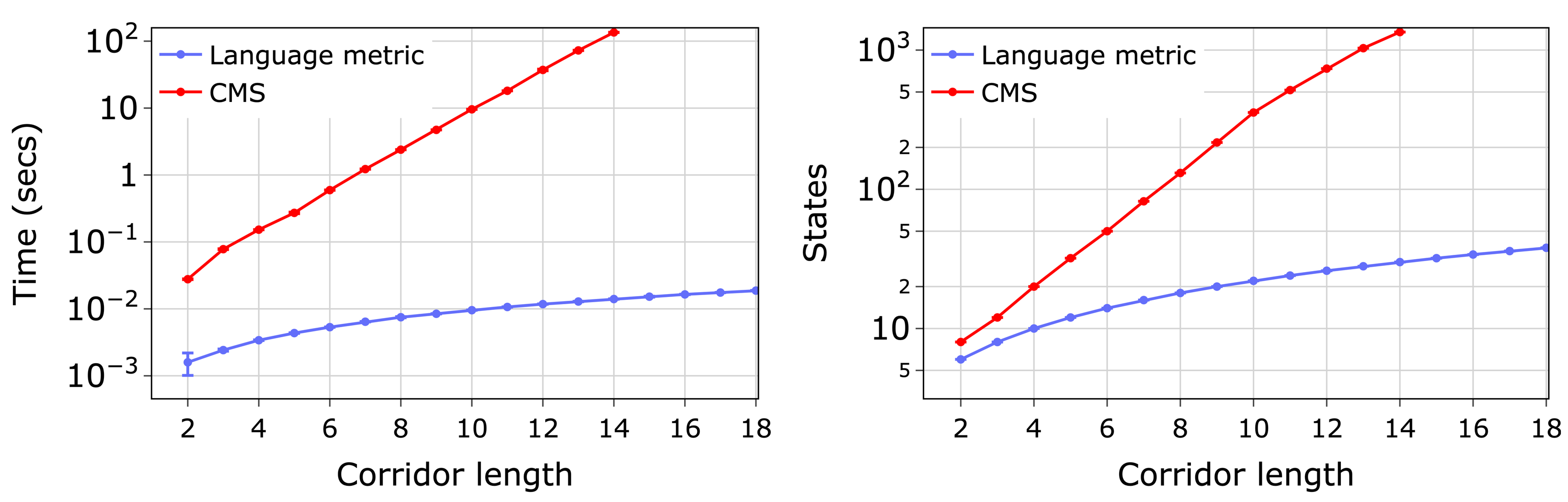
Experimental Evaluation

We provide an extensive evaluation showing that AdaCT-H equipped with the language metric $L_{\mathcal{X}_{3,1}}$ outperforms the state of the art.

		FlexFringe			CMS			Language metric		
Name	H	U	r	time	U	r	time	U	r	time
Corridor	5	11	**1.0**	0.03	11	**1.0**	0.3	11	**1.0**	**0.01**
T-maze(c)	5	29	0.0	0.11	104	**4.0**	10.1	18	**4.0**	**0.26**
Cookie	9	220	**1.0**	0.36	116	**1.0**	6.05	91	**1.0**	**0.08**
Cheese	6	669	$0.69 \pm .04$	19.28	1158	$0.4 \pm .05$	207.4	1178	$\mathbf{0.87 \pm .03}$	**12.11**
Mini-hall	15	897	$0.33 \pm .04$	25.79	-	-	-	6098	$\mathbf{0.86 \pm .03}$	**29.90**

Quantities: H is the horizon, U is the number of states of the learned automaton, r is the reward of the computed policy averaged over 100 episodes.

A clear exponential gain is observed in the figures below, reporting *run-time* and *number of states* for increasing corridor length in the T-maze.



References

[1] B. Bakker. Reinforcement learning with long short-term memory. In *NeurIPS*, 2001.

[2] R. Baumgartner and S. Verwer. Learning state machines from data streams. In *ICGI*, 2023.

[3] R. Cipollone, A. Jonsson, A. Ronca, and M. S. Talebi. Provably efficient offline reinforcement learning in regular decision processes. In *NeurIPS*, 2023.

Ahana Deb ahana.deb@upf.edu
Roberto Cipollone cipollone.rt@gmail.com
Anders Jonsson anders.jonsson@upf.edu
Alessandro Ronca alessandro@ronca.me
M. Sadegh Talebi sadegh.talebi@di.ku.dk