Tractable Offline Learning of Regular Decision Processes

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

[1]*Universitat Pompeu Fabra* [2]*Sapienza University of Rome* [3]*University of Oxford* [4]*University of Copenhagen*

Regular Decision Process

- In an episodic Non-Markov Decision Process $\langle O, A, R, T, R, H\rangle$, the transition probabilities $T : (AO)^* \times A \to \Delta(O)$ and rewards $R : (AO)^* \times A \to \mathbb{R}$ are functions of the entire interaction history $(AO)^*$
- In a Regular Decision Process (RDP), T and R depend regularly on the interaction history, and the dynamics can be represented by a Probabilistic-Deterministic Finite Automaton (PDFA)
- Example: T-maze domain

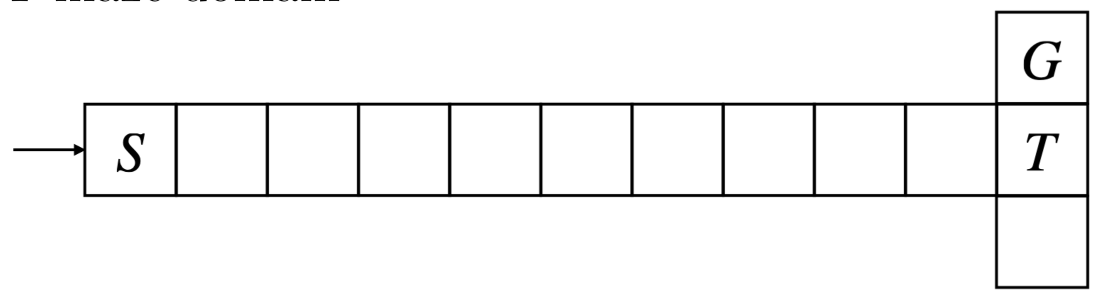

Figure: T-maze [1] with corridor length $N = 10$. The observation at the initial position S indicates the position of the goal G at the end of the corridor for the current episode.

Objective

Objective: Given a dataset D of episodes, collected from an unknown RDP R and unknown behavior policy π^b, compute a near-optimal policy for R, using the smallest D possible.
Question: A near-optimal policy can be computed from the PDFA of the RDP. Can we **learn** the PDFA of an RDP **R** from an interaction history?

AdaCT-H

AdaCT-H [3] returns the PDFA of an RDP **R** and achieves a sample complexity with polynomial dependency on the problem parameters

Input: Dataset $\mathcal{D}$ of traces in Γ^{H+1}, failure probability $0 < \delta < 1$
Output: Set $\mathcal{Q}$ of RDP states, transition function $\tau : \mathcal{Q} \times \mathcal{AO} \to \mathcal{Q}$

$\mathcal{Q}_0 \leftarrow \{q_0\}, \mathcal{Z}(q_0) \leftarrow \mathcal{D}$ // initial state
for $t = 0, \dots, H$ **do**
  $\mathcal{Q}_{c,t+1} \leftarrow \{qao \mid q \in \mathcal{Q}_t, ao \in \mathcal{AO}\}$ // get candidate states
  foreach $qao \in \mathcal{Q}_{c,t+1}$ **do** $\mathcal{Z}(qao) \leftarrow \{e_{t+1:H} \mid aroe_{t+1:H} \in \mathcal{Z}(q)\}$ // compute suffixes
  $q_m a_m o_m \leftarrow \arg\max_{qao \in \mathcal{Q}_{c,t+1}} |\mathcal{Z}(qao)|$ // most common candidate
  $\mathcal{Q}_{t+1} \leftarrow \{q_m a_m o_m\}, \tau(q_m, a_m o_m) = q_m a_m o_m$ // promote candidate
  $\mathcal{Q}_{c,t+1} \leftarrow \mathcal{Q}_{c,t+1} \setminus \{q_m a_m o_m\}$ // remove from candidate states
  for $qao \in \mathcal{Q}_{c,t+1}$ **do**
    Similar $\leftarrow \{q' \in \mathcal{Q}_{t+1} \mid \text{not } \textsc{TestDistinct}(t, \mathcal{Z}(qao), \mathcal{Z}(q'), \delta)\}$ // confidence test
    if *Similar* $= \emptyset$ **then** $\mathcal{Q}_{t+1} \leftarrow \mathcal{Q}_{t+1} \cup \{qao\}, \tau(q, ao) = qao$ // promote candidate
    else $q' \leftarrow$ element in *Similar*, $\tau(q, ao) = q', \mathcal{Z}(q') \leftarrow \mathcal{Z}(q') \cup \mathcal{Z}(qao)$ // merge states
  end
end
return $\mathcal{Q}_0 \cup \dots \cup \mathcal{Q}_{H+1}, \tau$
Function $\textsc{TestDistinct}(t, \mathcal{Z}_1, \mathcal{Z}_2, \delta)$
  return $L^{\mathsf{p}}_\infty(\mathcal{Z}_1, \mathcal{Z}_2) \geq \sqrt{2\log(8(ARO)^{H-t}/\delta)/\min(|\mathcal{Z}_1|, |\mathcal{Z}_2|)}$

The bottleneck is the statistical test on the prefix distance defined as

$$L^{\mathsf{p}}_\infty(p_1, p_2) = \max_{u \in [0,\ell], e \in E_u} |p_1(e*) - p_2(e*)|$$

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

$$L^{\mathsf{p}}_\infty(p_1, p_2) \geq \mu_0 > 0$$

For example, in T-maze, L^{p}_∞-distinguishability **decreases exponentially** with the corridor length N.

Contributions

A practical implementation of AdaCT-H that reduces the memory and time complexity

- Exploit the Count-Min-Sketch (CMS) data structure to reduce the memory complexity of storing the empirical distributions on suffixes
- Develop a novel language metric L_X, based on the theory of formal languages, and define a hierarchy of language families that removes the dependency on L^{p}_∞-distinguishability and is exponentially more sample efficient in domains having low complexity in language-theoretic terms

Language Hierarchies

- We define the following sets of basic patterns $\mathcal{G}_i$ of increasing complexity

$$\mathcal{G}_1 = \{aO/R \mid a \in A\} \cup \{AO/r \mid r \in R\} \cup \{Ao/R \mid o \in O\},$$
$$\mathcal{G}_2 = \mathcal{G}_1 \cup \{ao/R \mid a \in A, o \in O\} \cup \{aO/r \mid a \in A, r \in R\} \cup \{Ao/r \mid a \in A, r \in R\}.$$

- The operator C^ℓ_k which maps any set of languages $\mathcal{G}$ to a new set of languages:

$$\mathrm{C}^\ell_k(\mathcal{G}) = \{\{x_0 G_1 \cdots x_{k-1} G_k x_k \mid x_0, \dots, x_k \in \Gamma^*, |x_0 \cdots x_k| = (\ell - k)\} \mid G_1, \dots, G_k \in \mathcal{G}\}.$$

- *Two-dimensional hierarchy* of sets $X_{i,j}$ of languages: $X_{i,j} = \cup_{k \in j} \mathrm{C}^\ell_k(\mathcal{G}_i)$, $\forall i \in 3$, $\forall j \in \ell$.
- **Induced language metric**: $L_X(p, p') \coloneqq \max_{X \in X} |p(X) - p'(X)|$, where $p(X) \coloneqq \sum_{x \in X} p(x)$.

Analysis

Theorem 1: AdaCT-H(D, δ) returns a minimal RDP R with probability at least $1 - 3AOQ\delta$ when CMS is used to store the empirical probability distributions of episode suffixes, the statistical test is

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

and the size of the dataset is at least $|D| \geq \widetilde{\mathcal{O}}(HC^*_{\mathbf{R}}\log(1/\delta)/d^*_{\min}\mu_0^2)$, where $d^*_{\min} = \min_{t,q,ao} d^*_t(q, ao)$ and $C^*_{\mathbf{R}}$ is the single-policy concentrability of **R**.

Theorem 2: AdaCT-H(D, δ) returns a minimal RDP R with probability at least $1 - 2AOQ\delta$ when the statistical test is implemented using the language metric L_X and equals

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

and the size of the dataset is at least $|D| \geq \widetilde{\mathcal{O}}(C^*_{\mathbf{R}}\log|\mathcal{X}|\log(1/\delta)/d^*_{\min}\mu_0^2)$.

Results

		FlexFringe			CMS			Language metric		
Name	H	Q	r	time	Q	r	time	Q	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	326	$\mathbf{0.81 \pm .04}$	**2.23**
Mini-hall	15	897	$0.33 \pm .04$	25.79	-	-	-	5134	$\mathbf{0.91 \pm .03}$	**23.9**

Figure: Summary of the experiments. We compare our two approaches against FlexFringe[2] for the average reward over 100 episodes, time taken(seconds) and size of the state space learned (Q).

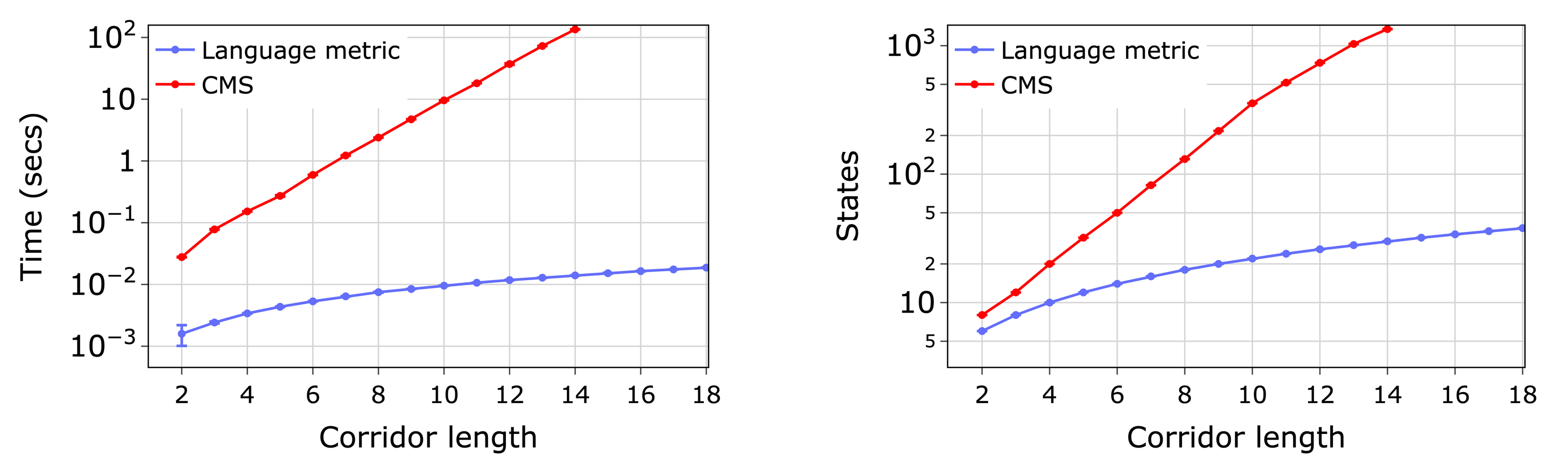

Figure: Impact of increasing the length of the corridor for the T-maze domain.

References

[1] B. Bakker. Reinforcement learning with long short-term memory. In *Neural Information Processing Systems (NeurIPS)*, pages 1475–1482, 2001.

[2] R. Baumgartner and S. Verwer. Learning state machines from data streams: A generic strategy and an improved heuristic. In *International Conference on Grammatical Inference (ICGI)*, pages 117–141, 2023.

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