DYNAMIC AUTOMATON REFINEMENT AND PLANNING FOR NON-MARKOVIAN RL

Mahyar Alinejad[1] **Yue Wang**[1] **George K. Atia**[1]

[1]Dept. of Electrical & Computer Engineering, University of Central Florida, Orlando, FL, USA {mahyar.alinejad, yue.wang, george.atia}@ucf.edu

Key Contributions

- **DARP**: first framework unifying **passive DFA inference**, **online refinement**, and **efficient planning** in one loop.
- **Zero-knowledge start** — discovers valid patterns from *no* prior positive examples.
- $3\times$ **faster** convergence than Static RPNI (38 vs. 101 episodes).
- **56.7%** pattern discovery vs. 0% for static/HiPO baselines.
- Provably: almost-sure convergence + $\mathcal{O}(\log T)$ regret.

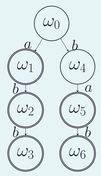
Background: Automata & Product MDPs

1) Learning a DFA from examples (RPNI):

Given *positive* traces (task completed) and *negative* traces (task failed), RPNI builds a minimal DFA that accepts all positives and rejects all negatives.

Example: Alphabet $\{a, b\}$, Pos: $\{a, ab, abb, ba, bab\}$, Neg: $\{b, aa, bb\}$

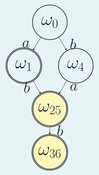
Prefix Tree



merge

Merge ω_1, ω_4 **rejected**: accepts neg. "b"

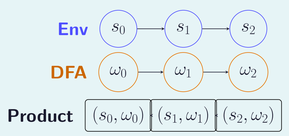
Minimal DFA



ω_{25}: merged ω_2, ω_5

ω_{36}: merged ω_3, ω_6

2) Product MDP: augment each environment state s with the current DFA state ω, turning history-dependent rewards into *Markovian* rewards on the joint state (s, ω).



Reward $= 1$ only when the DFA reaches an *accepting state* $\omega \in F$.

Problem: Non-Markovian RL

An NMRDP $\mathcal{M} = (S, s_0, A, P, \mathsf{AP}, L, R)$ has rewards $R : (2^{\mathsf{AP}})^* \to \{0, 1\}$ depending on the full trace $w = L(s_0)L(s_1)\cdots$ The reward is encoded by an **unknown** DFA $\mathcal{A}^* = (\Omega, \Sigma, \delta, \omega_0, F)$ that must be inferred online. **Joint objective:**

$$(\pi^*, \mathcal{A}^*) = \arg\max_{\pi, \mathcal{A}} \mathbb{E}_\pi\Big[R(w)\Big].$$

Initial knowledge gap. Agent begins with set of labeled trajectories $(P_{\text{init}}, N_{\text{init}})$, positives yielded reward and negatives did not. Examples may be *incomplete* or even *entirely absent* ($P_{\text{init}} = \varnothing$, zero-start), so the automaton must be **refined online** as new evidence arrives.

Gap in Prior Work

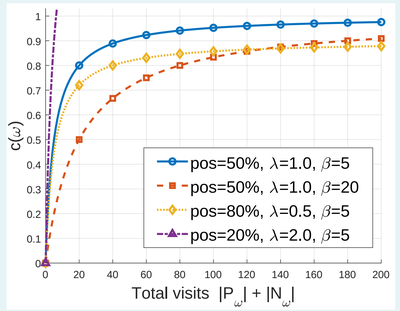
Method	Passive	Online refine	Zero-start
Static RPNI [3]	✓	×	×
AGI / L^* [4]	×	✓	✓
HiPO [5]	✓	×	×
DARP (ours)	✓	✓	✓

The DARP Method

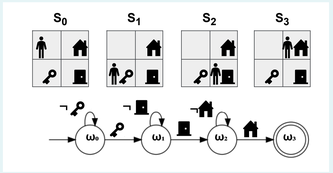
① Confidence Score per Automaton State

$$c(\omega) = \frac{|P_\omega| + \lambda\,|N_\omega|}{|P_\omega| + |N_\omega| + \beta}, \qquad \lambda = 0.5,\ \beta = 5$$

P_ω / N_ω: positive/negative trace visits through ω. Low $c(\omega) \Rightarrow$ state is unreliable and needs refinement.



$c(\omega)$ grows with evidence; β prevents early saturation.



Sequential task: key→door→home. DFA accepts only the correct ordering.

② Online Refinement Trigger

After each episode (trace w_t, outcome $y_t \in \{0, 1\}$), refinement fires if:

- **Inconsistency:** $\mathcal{A}_t(w_t) \neq y_t$, **or**
- **Coverage gap:** $\exists\, \omega \in \Omega$ visited with $c(\omega) < \theta_{\min}$

$$\mathcal{A}_{t+1} = \text{RPNI}\Big(\mathcal{P}_t \cup \{w_t : y_t = 1\},\ \mathcal{N}_t \cup \{w_t : y_t = 0\}\Big)$$

③ Exploration Bonus (CW-UCB)

Let ω_δ be the automaton state reached by action a from s:

$$\text{bonus}(s, a) = \alpha\sqrt{\tfrac{\log t}{N(\omega_\delta)}} \cdot \max\Big\{1 - c(\omega_\delta),\ 0\Big\}$$

Focuses exploration on *rarely visited* **and** *low-confidence* DFA regions, avoiding wasted steps in already well-understood parts of the automaton.

④ Incremental Product MDP + Q-Transfer

On refinement, only automaton states whose transitions changed are rebuilt:

$$\Delta_\Omega := \{\, \omega \in \Omega_t : \delta_t(\omega, \sigma) \neq \delta_{t+1}(\omega, \sigma) \text{ for some } \sigma \in \Sigma \,\}$$

New product states are warm-started by averaging predecessor Q-values; all unchanged states are preserved:

$$Q_{t+1}((s,\omega), a) = \begin{cases} Q_t((s,\omega), a) & \omega \notin \Delta_\Omega \\ \dfrac{1}{|\mathcal{P}(s,\omega)|} \displaystyle\sum_{(s,\omega') \in \mathcal{P}(s,\omega)} Q_t((s,\omega'), a) & \omega \in \Omega_{t+1} \setminus \Omega_t \\ \epsilon_{\text{init}} & \text{otherwise} \end{cases}$$

Theorem (Convergence & Regret)

Under ergodic exploration:

- $\mathbb{P}\Big(\lim_{t\to\infty} \mathcal{A}_t = \mathcal{A}^*\Big) = 1$ *(almost-sure convergence)*
- $\mathbb{E}[\text{Regret}(T)] = \mathcal{O}(\log T)$ *vs.* $\Omega(T)$ *for static baselines*

Experimental Results

Dungeon Quest (DQ): grid-based, 3 valid item-collection orderings.
Blind Craftsman (BC): resource-constrained, 3 optimal wood-to-tool patterns.

Pattern Discovery

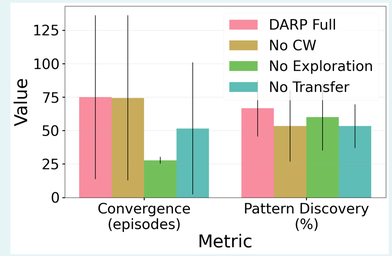
Method	Dungeon Quest		Blind Craftsman	
	Disc. (%)	Pat.	Disc. (%)	Pat.
DARP	**56.7±15.3**	**1.7/3**	**73.3±29.1**	**2.2/3**
Static RPNI	0.0	0/3	0.0	0/3
AGI	46.7±16.3	1.4/3	60.0±20.0	1.8/3
HiPO	0.0	0/3	0.0	0/3

Sample Complexity

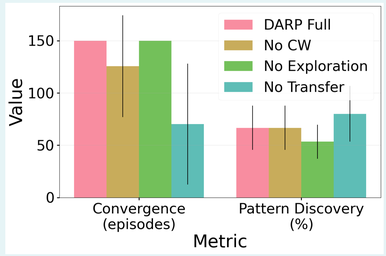
Method	Episodes to 90%		Final Success (%)	
	DQ	BC	DQ	BC
DARP	**38±37**	**116±42**	**100.0**	**62.5±41.8**
Static RPNI	101±60	>150	99.0±1.0	60.5±40.3
AGI	54±48	>150	98.0±2.0	5.5±7.6
HiPO	61±58	132±40	99.5±0.5	47.5±34.2

Mann-Whitney U: $p < 0.001$ vs. RPNI (Cohen's $d = 5.25$ DQ, 3.57 BC). All improvements significant at 99%.

Ablation Study



(a) Dungeon Quest



(b) Blind Craftsman

All three components — **confidence weighting**, **exploration bonus**, **Q-transfer** — are individually necessary; removing any one degrades performance.

Conclusions

- **DARP** jointly learns DFA structure and policy **online**, starting from zero prior knowledge of valid patterns.
- Confidence-guided refinement eliminates spurious automaton updates while targeting unreliable states.
- Outperforms all baselines in both pattern discovery and sample efficiency.
- **Future:** continuous state spaces, deep RL (DQN/PPO), multi-agent settings.

References

- R. Icarte et al. *Using reward machines.* ICML 2018.
- D. Angluin. *Learning regular sets from queries.* Inf. & Comp., 1987.
- J. Oncina, P. Garcia. *Identifying regular languages.* ALT 1992.
- N. Topper et al. *Active grammatical inference.* ICAPS 2022.
- M. Alinejad et al. *Hybrid passive inference & online planning.* ICASSP 2025.

Supported by DARPA HR0011-24-9-0427 & NSF CCF-2106339