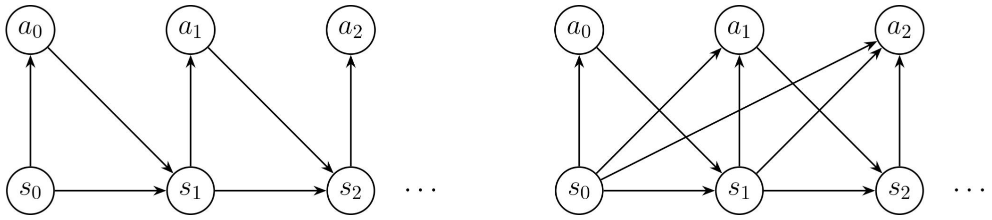
Learning non-Markovian Decision-Making from State-only Sequences

Aoyang Qin, Feng Gao, Qing Li, Song-Chun Zhu, Sirui Xie

Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS-2023)

Introduction

- ***The expressivity of Markov reward has been proved to be limited.***
- We develop a Maximum Likelihood Estimation for generative modeling of *non-Markovian Decision Process (nMDP)*, where TD-learning-based imitation is unreliable.
- The novel EM-like algorithm recover the unobserved decisions and underlying value functions from pure observations *without action labels*.



Graphical model of policy and transition in standard **MDP** and **nMDP**

Abel, David, et al. "On the expressivity of markov reward." Advances in Neural Information Processing Systems 34 (2021): 7799-7812.

Modeling and Learning

- Trajectory joint distribution:

$$p_\theta(\zeta) = p(s_0) \prod_{t=0}^{T-1} p_\alpha(a_t|s_{0:t}) p_\beta(s_{t+1}|s_t, a_t)$$

- Transition as single-mode Gaussian

$$\mathcal{N}(g_\beta(s_t, a_t), \sigma^2)$$

- Policy as multi-mode Energy-Based Model (EBM)

$$p_\alpha(a_t|s_{0:t}) = \frac{1}{Z(\alpha, s_{0:t})} \exp\left(f_\alpha(a_t; s_{0:t})\right)$$

- MLE learning, the gradient is:

$$\nabla_\theta \log p_\theta(\xi) = \mathbb{E}_{p_\theta(A|S)}[\sum_{t=0}^{T-1} (\underbrace{\nabla_\alpha \log p_\alpha(a_t|s_{0:t})}_{\text{policy/prior}}, \underbrace{\nabla_\beta \log p_\beta(s_{t+1}|s_t, a_t)}_{\text{transition}})]$$

Sampling

- Policy term involves both posterior and prior samples:

$$\begin{aligned}\delta_{\alpha,t}(S) &= \mathbb{E}_{p_\theta(A|S)}\left[\nabla_\alpha \log p_\alpha(a_t|s_{0:t})\right] \\ &= \mathbb{E}_{p_\theta(A|S)}\left[\nabla_\alpha f_\alpha(a_t; s_{0:t})\right] - \mathbb{E}_{p_\alpha(a_t|s_{0:t})}\left[\nabla_\alpha f_\alpha(a_t; s_{0:t})\right]\end{aligned}$$

- Short run Langevin MCMC for prior samples:

$$a_{t,k+1} = a_{t,k} + s\nabla_{a_{t,k}} f_\alpha(a_{t,k}; s_{0:t}) + \sqrt{2s}\epsilon_k$$

- Importance sampling for posterior samples:

$$p_\theta(a_t|s_{0:t+1}) = \frac{p_\beta(s_{t+1}|s_t, a_t)}{\mathbb{E}_{p_\alpha(a_t|s_{0:t})}\left[p_\beta(s_{t+1}|s_t, a_t)\right]} p_\alpha(a_t|s_{0:t})$$

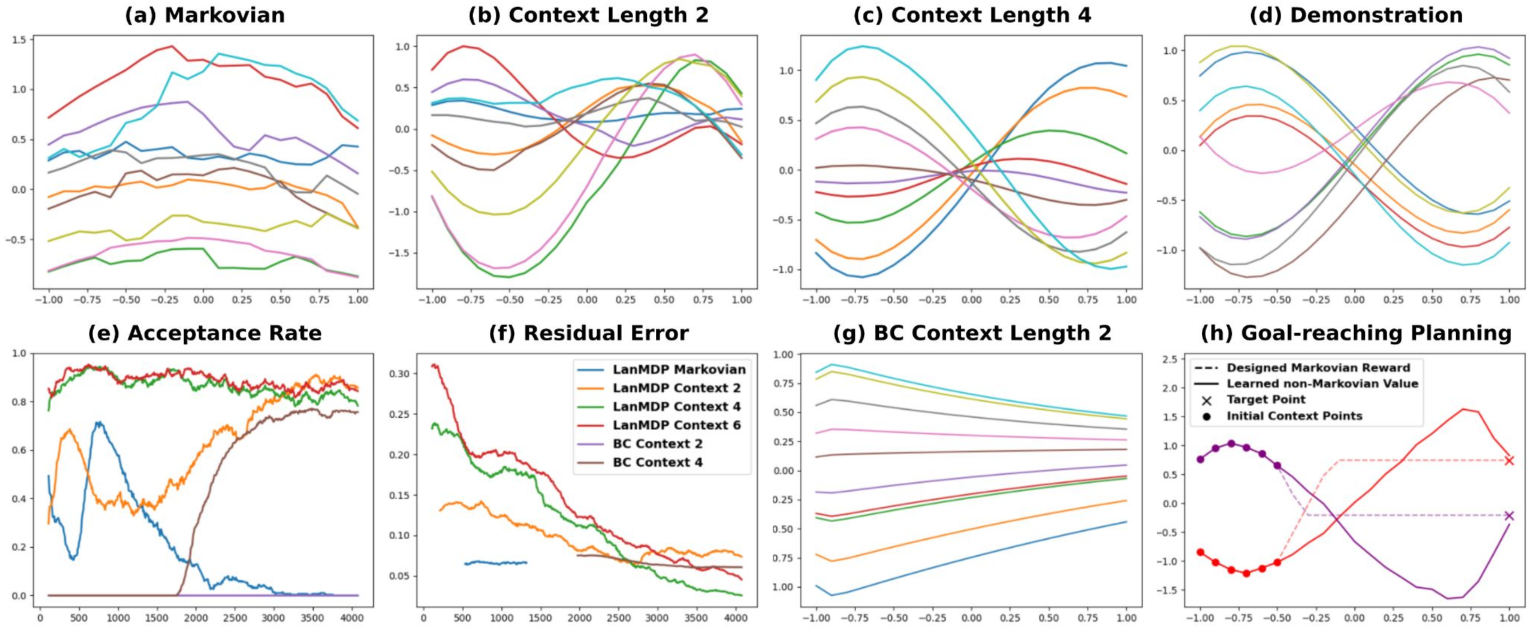
Theoretical Analysis

- We construct a sequential decision-making problem, whose objective yields the same optimal policy as MLE.
- We witness the automatic emergence of the entire family of maximum (inverse) RL.
- We derive the posterior probability of action sequences given any goal state, involving the intermediate transitions.

Decision-making as inference:
policy as prior, planning as inference.

Experiments: Curve Planning

- Policy of cubic curve planning is necessarily non-Markovian, since the historical states are needed to estimate the higher-order derivatives.



Experiments: MuJoCo

- Our model demonstrates steeper learning curves than state-only baselines.
- Our model matches or surpasses the performance of action-label baselines.