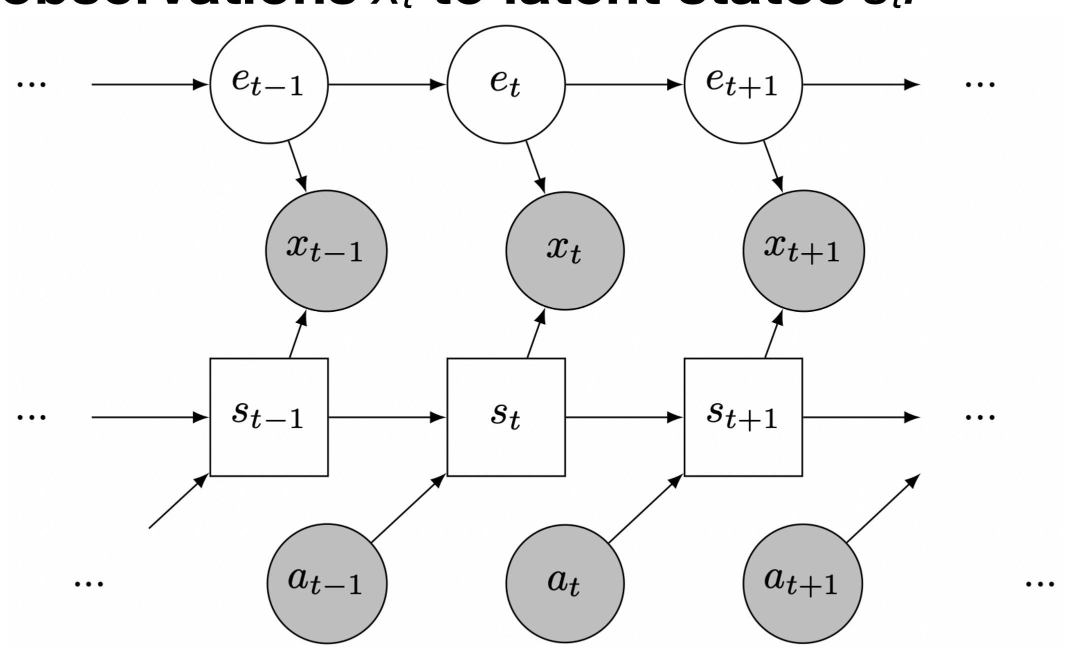
Learning a Fast Mixing Exogenous Block MDP using a Single Trajectory

Alexander Levine[1], Peter Stone[1,2], and Amy Zhang[1]

1: The University of Texas at Austin. 2: Sony AI. Correspondence to alevine0@cs.utexas.edu

Ex-BMDP Model (Efroni et al., 2022)

- Observation $x_t \in X$ can be factored into *controllable* state $s_t \in S$ and *noise* state $e_t \in \mathcal{E}$.
- Controllable state evolves deterministically, according to actions: $s_{t+1} = T(s_t, a_t)$.
- Noise (exogenous) state evolves as a Markov chain, independent of actions : $e_{t+1} \sim T_e(e_t)$.
- Observation $x_t \sim Q(s_t, e_t)$; e_t and s_t are not observed and factorization not known *a priori*.
- X and $\mathcal{E}$ can be continuous or large, S is assumed to be discrete and small.
- **Goal: learn an encoder ϕ to map observations x_t to latent states s_t.**



(Fig. From Levine et al. 2024)

Related Work

- Efroni et al. (2022): Proposed provably sample-efficient algorithm, PPE, for learning Ex-BMDP representations in the *finite horizon* setting, where the latent state s resets to a specific s_1 after (almost) every episode.
  - Also allows for near-deterministic latent dynamics T, rather than full determinism.
- Lamb et al. (2023), Levine et al. (2024): proposed algorithms for the infinite-horizon, no-reset setting, but without sample-complexity guarantees.
- **This work: we propose a provably sample-efficient algorithm for Ex-BMDP representation learning in the infinite-horizon, no reset setting.**

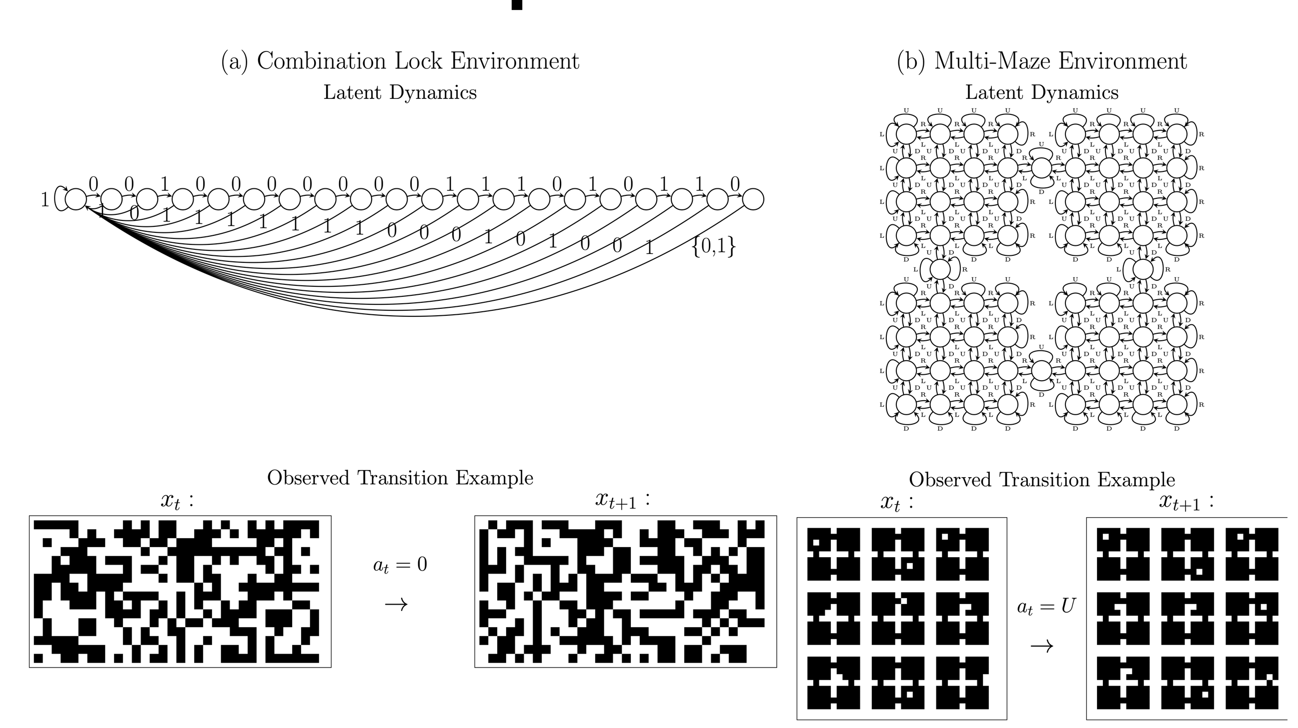
Problem Setting and Guarantees

- Agent interacts with the Ex-BMDP in a ***single trajectory***, with ***no ability to reset*** the environment.
  - Models cases, such as in robotic navigation, where manually resetting the environment repeatedly during training could be costly.
- **Core Difficulty**: In the (near) deterministic, episodic setting (Efroni et al. 2022), taking the same action sequence $a_1,\ldots,a_t$ for repeated episodes (usually) yields i.i.d. samples of a single latent state s_t. **Not possible in the no-reset, single trajectory setting.**
- We assume that the noise state e_t *mixes fast*:

$$\forall e \in \mathcal{E},\ \ \|\Pr(e_{t+\hat{t}_{\text{mix}}} = e' | e_t = e) - \pi_{\mathcal{E}}(e')\|_{\text{TV}} \leq \frac{1}{4},$$

where $\pi_{\mathcal{E}}$ is the stationary distribution of the noise state, and $\hat{t}_{\text{mix}}$ is a known upper-bound on the mixing time. (Necessary assumption)
- **Our proposed algorithm, STEEL, has sample-complexity polynomial in $|S|$ and $\hat{t}_{\text{mix}}$, and logarithmic in the size of the hypothesis class of the encoder ϕ, with *no explicit dependence* on $|X|$ and $|\mathcal{E}|$.**

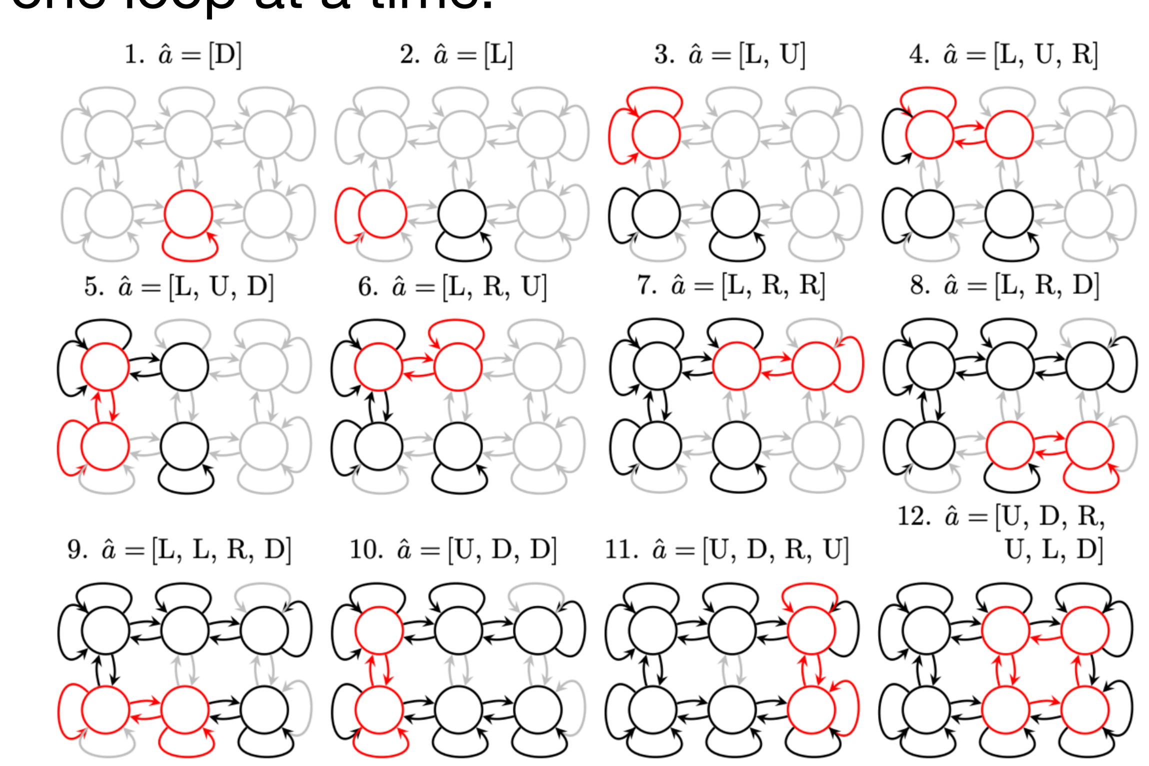
Experiments



	Combo. Lock $(K=20)$	Combo. Lock $(K=30)$	Combo. Lock $(K=40)$	Multi-Maze
Accuracy	20/20	20/20	20/20	20/20
Env. Steps	$2.00{\cdot}10^6$ $\pm1.28{\cdot}10^5$	$4.78{\cdot}10^6$ $\pm4.36{\cdot}10^5$	$9.59{\cdot}10^6$ $\pm1.13{\cdot}10^6$	$4.13{\cdot}10^7$ $\pm1.11{\cdot}10^6$

Algorithm (STEEL)

- Core Idea: Repeating any action sequence $\hat{a} = [a_1,..,a_n]$ is guaranteed to *eventually* enter a loop of latent states (of length at most $n{*}|S|$)
  - Once in a loop, we can "wait out" the mixing time $\hat{t}_{\text{mix}}$ to get near-i.i.d. samples.
  - Once we find the period of the cycle, we can collect near-i.i.d. datasets from all visited latent states.
- We can then construct the latent dynamics one loop at a time:



- Challenges :
  - How do we determine the period of a cycle?
  - How do we ensure that all latent states in S are covered by some cycle?
  - See paper to find out!

References

- Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. ICLR. 2022.
- Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J Foster, Lekan P Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. TMLR. 2023.
- Alexander Levine, Peter Stone, and Amy Zhang, Multistep inverse is not all you need. RLC 2024.