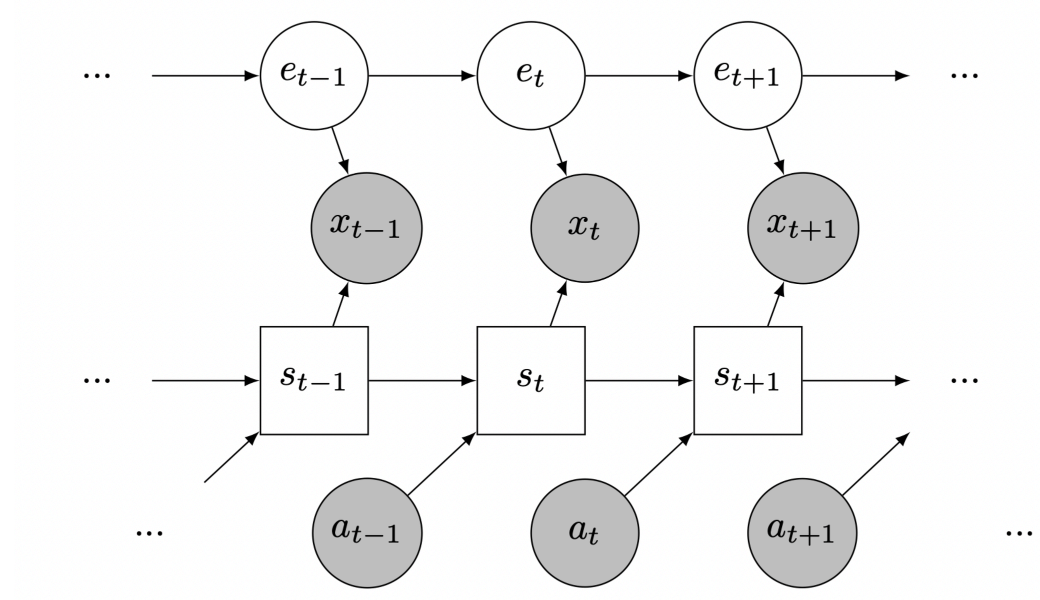
Multistep Inverse is Not All You Need

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

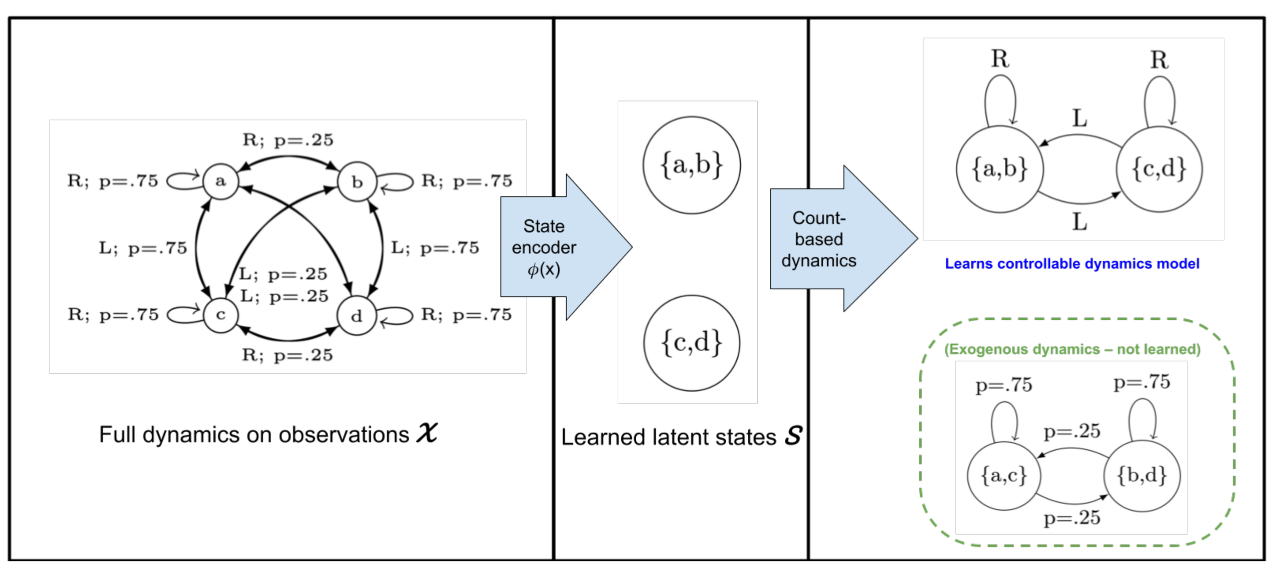
1: The University of Texas at Austin. 2: Sony AI. Correspondence to alevine0@cs.utexas.edu

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



- Observation $x \in X$ can be factored into latent states:
  - Endogenous state $s \in S$, discrete, evolves deterministically
  - Exogenous state $e \in \mathcal{E}$, stochastic, indep. of actions (***noise***)

2. Representation Learning under Ex-BMDP Framework



- Task: learn encoder ϕ to map $x \in X$ to $s \in S$.
- Existing Methods:
  - Efroni et al. (2022a, 2022b), Mhammedi (2023): *finite-horizon* setting, learn separate encoders ϕ_t at each t.
  - Lamb et al. (2022): ***infinite-horizon setting*** with ***no resets***
    - ***Bounded diameter* assumption: $\forall$ s,s' $\in$ S, d(s,s') $\leq$ D**

3. Multistep Inverse (Lamb et al., 2022)

- ***AC-State:*** predict a_t given $\phi(x_t)$, $\phi(x_{t+k})$, k:

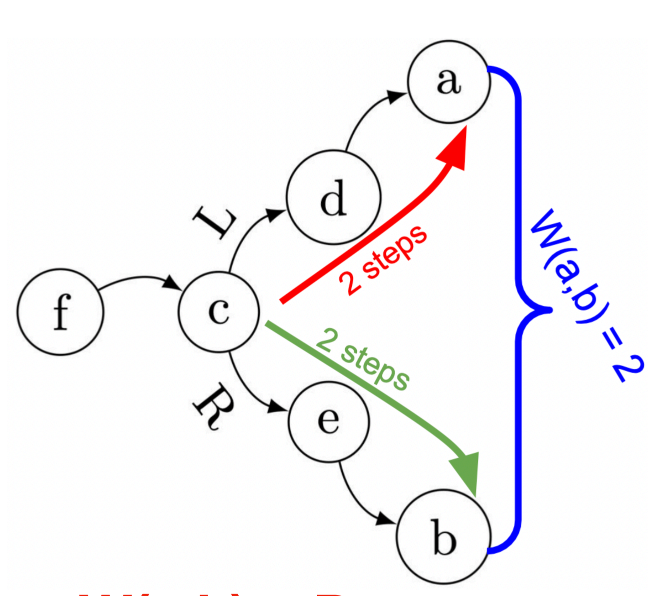
$$\mathcal{L}_{\text{AC-State}}(\phi_\theta) := \min_f \mathop{\mathbb{E}}_{k\sim\{1,\dots,D\}} \mathop{\mathbb{E}}_{(x_t,a_t,x_{t+k})} -\log(f_{a_t}(\phi_\theta(x_t),\phi_\theta(x_{t+k});k))$$

$$\{\theta\}^* := \{\theta^{**} | \theta^{**} = \arg\min_\theta \mathcal{L}_{\text{AC-State}}(\phi_\theta)\}$$

$$\theta^* := \arg\min_{\theta\in\{\theta\}^*} \|\text{Range}(\phi_\theta)\|$$

- Must show that learned ϕ won't conflate two different states s, s' $\in$ S:
  - **Proof Sketch (re-framed):**

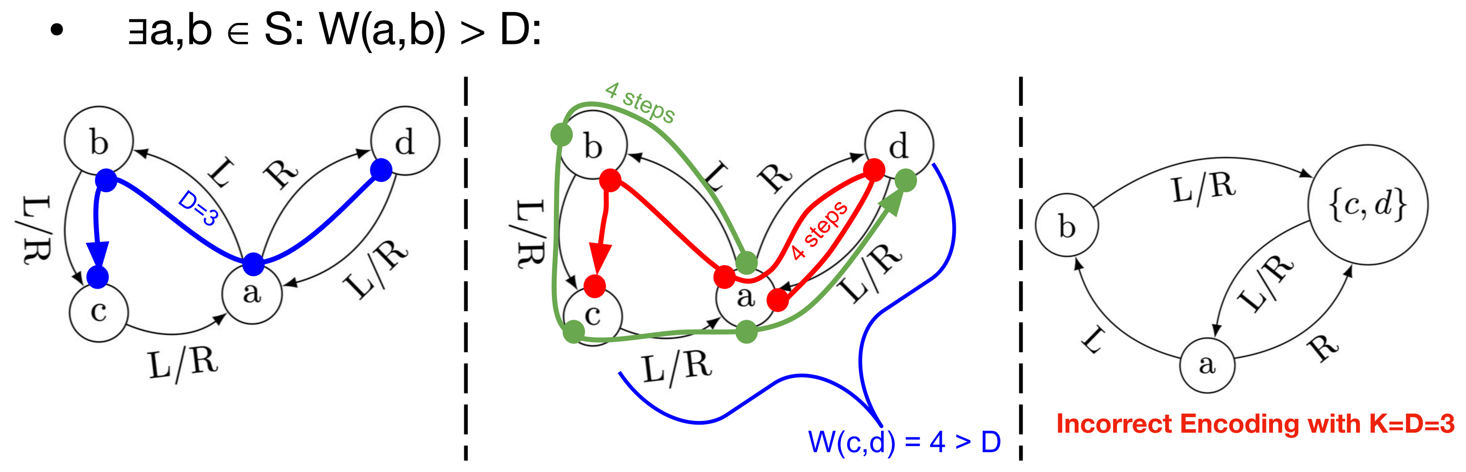
For a,b $\in$ S, let **W(a,b)** be the min. k such that $\exists$c $\in$ S, such that a and b can both be reached from c in exactly k steps. Compare $P(a_t \mid s_t = c, s_{t+k} = a)$ vs. $P(a_t \mid s_t = c, s_{t+k} = b)$. These distributions have ***disjoint support.*** Otherwise W(a,b) < k. Therefore ϕ must distinguish a, b.



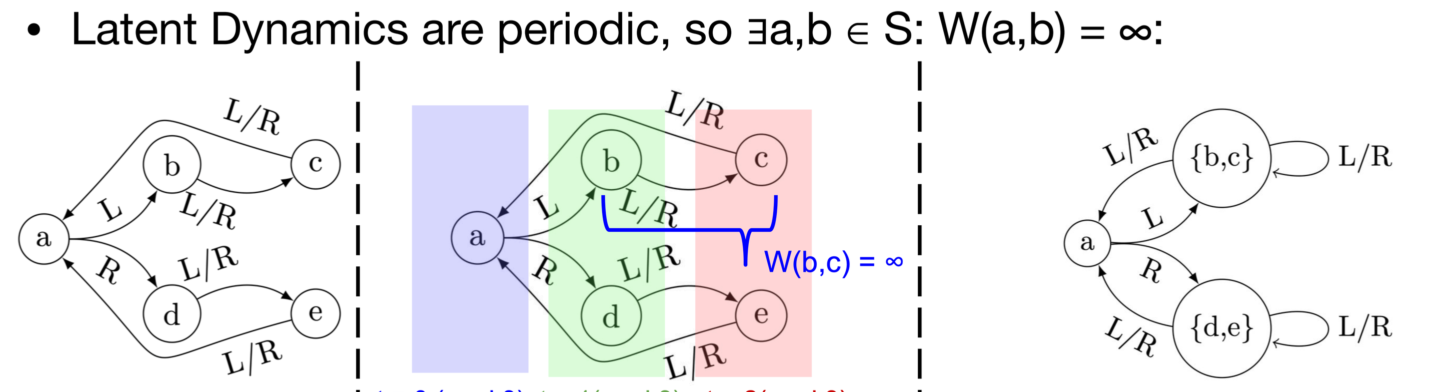
- **Flawed implicit assumption: W(a,b) ≤ D.**

4. Multistep Inverse Is Not All You Need

- AC-State can ***fail*** if ***either***:
  - $\exists$a,b $\in$ S: W(a,b) > D:



- Latent Dynamics are periodic, so $\exists$a,b $\in$ S: W(a,b) = ∞:



5. ACDF: A Fix for Multistep Inverse

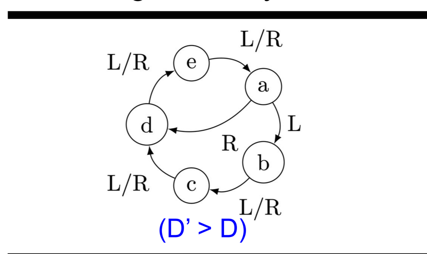
$$\mathcal{L}_{\text{ACDF}}(\phi_\theta) := \min_f \mathop{\mathbb{E}}_{k\sim\{1,\dots,D'\}} \mathop{\mathbb{E}}_{(x_t,a_t,x_{t+k})} -\log(f_{a_t}(\phi_\theta(x_t),\phi_\theta(x_{t+k});k))$$
$$+ \min_g \mathop{\mathbb{E}}_{(x_t,a_t,x_{t+1})} -\log(g_{\phi_\theta(x_{t+1})}(\phi_\theta(x_t),a_t)).$$

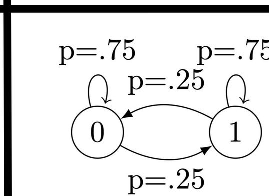
- D is replaced by D', which is any upper bound on finite W(a,b)
  - **Theorem**: If W(a,b) is finite, then W(a,b) ≤ $2D^2$ + D
    - Tight up to constant multiplicative factor
  - In practice, maximum number of steps is hyperparameter, K.
- Added ***latent forward model*** g: predict $\phi(x_{t+1})$ given $\phi(x_t)$ and a_t.
- **Theorem**: Encoders which minimize ACDF loss encode a correct endogenous latent representation.
- **AC-State + D' + Forward** model = **ACDF.**

6. Results

- **Tabular Setting:**
  - To compare AC-State and ACDF with no error from function approximation or optimization.
  - Measured success rate for learning correct encoder under tabular dynamics, for varying numbers of training samples and max. number of steps K of multistep-inverse dynamics prediction.

Endogenous Dynamics T | Exogenous Noise $\mathcal{T}_e$ | AC-State Success Rate | ACDF Success Rate


(D' > D)

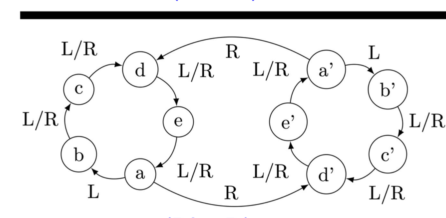


AC-State Success Rate:

Env. steps:	200	400	800	1600	3200
K=1	0%	0%	0%	0%	0%
K=2	0%	0%	0%	0%	0%
K=3	0%	0%	0%	0%	0%
K=4	0%	0%	0%	0%	0%
K=5	0%	0%	0%	0%	0%
K=6	0%	0%	0%	0%	0%
K=7	76%	100%	100%	100%	100%

ACDF Success Rate:

Env. steps:	200	400	800	1600	3200
K=1	100%	100%	100%	100%	100%
K=2	100%	100%	100%	100%	100%
K=3	100%	100%	100%	100%	100%
K=4	100%	100%	100%	100%	100%
K=5	100%	100%	100%	100%	100%
K=6	100%	100%	100%	100%	100%
K=7	100%	100%	100%	100%	100%


(D' > D)

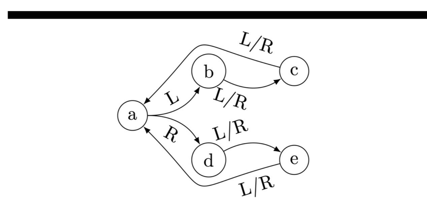
(None)

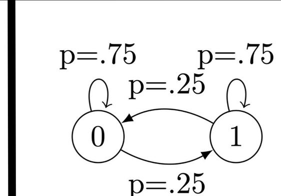
AC-State Success Rate:

Env. steps:	1000	2000	4000	8000	16000
K=10	0%	0%	0%	0%	0%
K=13	0%	0%	0%	0%	0%
K=16	0%	0%	0%	0%	0%
K=19	0%	0%	2%	0%	0%
K=22	0%	0%	2%	54%	98%
K=25	0%	0%	0%	18%	80%
K=28	0%	0%	0%	4%	38%

ACDF Success Rate:

Env. steps:	1000	2000	4000	8000	16000
K=10	0%	2%	0%	0%	0%
K=13	0%	12%	22%	64%	96%
K=16	0%	22%	96%	100%	100%
K=19	0%	12%	88%	100%	100%
K=22	0%	0%	68%	100%	100%
K=25	0%	0%	42%	98%	100%
K=28	0%	0%	32%	98%	100%


(Periodic)

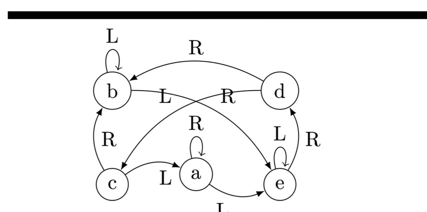


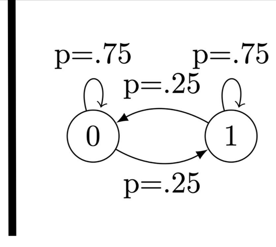
AC-State Success Rate:

Env. steps:	100	200	400	800	1600
K=1	0%	0%	0%	0%	0%
K=2	0%	0%	0%	0%	0%
K=3	0%	0%	0%	0%	0%
K=4	0%	0%	0%	0%	0%

ACDF Success Rate:

Env. steps:	100	200	400	800	1600
K=1	30%	14%	12%	8%	6%
K=2	92%	100%	100%	100%	100%
K=3	86%	98%	100%	100%	100%
K=4	84%	98%	100%	100%	100%


("Control": D' ≤ D; Aperiodic)



AC-State Success Rate:

Env. steps:	100	200	400	800	1600
K=1	0%	0%	0%	0%	0%
K=2	74%	100%	100%	100%	100%
K=3	24%	70%	100%	100%	100%
K=4	4%	19%	74%	97%	100%
K=5	0%	0%	44%	92%	100%

ACDF Success Rate:

Env. steps:	100	200	400	800	1600
K=1	98%	100%	100%	100%	100%
K=2	91%	100%	100%	100%	100%
K=3	68%	100%	100%	100%	100%
K=4	18%	88%	100%	100%	100%
K=5	4%	50%	98%	100%	100%

- **Function Approximation Setting:**
  - Gridworld-like maze navigation task and network architecture from released code of Lamb et al. (2022).
  - Compared original maze environment to a ***periodic*** variant of the environment, and original AC-State loss function to ACDF.
  - Evaluation based on success of encoder for open-loop planning.

	Baseline/AC-State	Baseline/ACDF	Periodic/AC-State	Periodic/ACDF
Success Rate	20/20 training runs	20/20 " "	1/20 " "	19/20 " "

7. Future Work

- Sample-complexity guarantees:
  - Neither AC-State nor ACDF have sample-complexity guarantees.
  - While sample-efficient algorithms have been proposed for finite-horizon Ex-BMDPs (Efroni et al. 2022a, 2022b; Mhammedi 2023), a method which such guarantees has not yet been proposed in the reset-free setting.
- State generalization/structured states:
  - Existing Ex-BMDP algorithms assume that *every possible* endogenous latent state is frequently visited during training.
  - There is a need to efficiently learn latent dynamics with combinatorial structure.

References

- Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. COLT. 2022a.
- Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. ICLR. 2022b.
- 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. 2022.
- Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. Representation learning with multi- step inverse kinematics: An efficient and optimal approach to rich-observation rl. ICML. 2023.