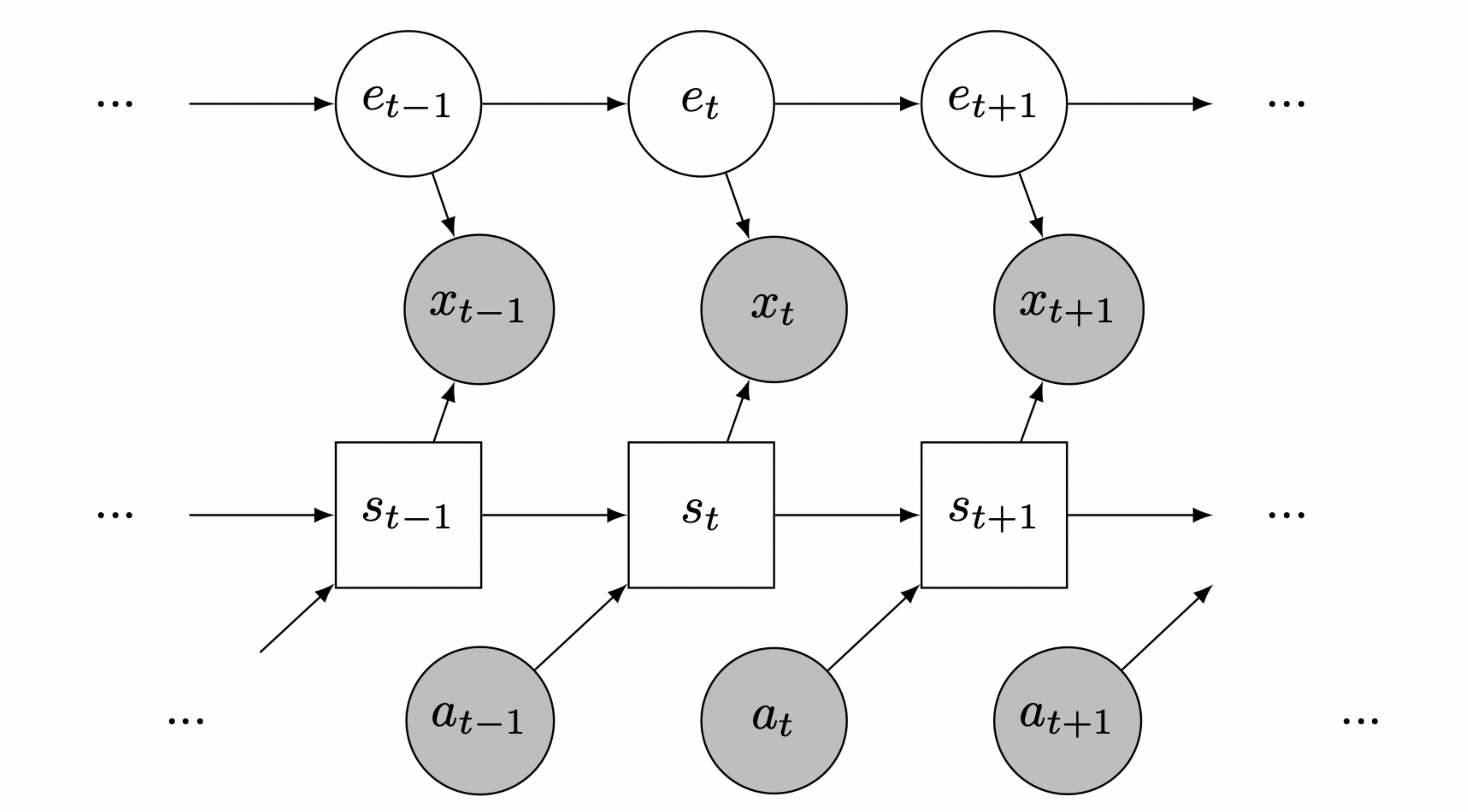
Multistep Inverse Is Not All You Need

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

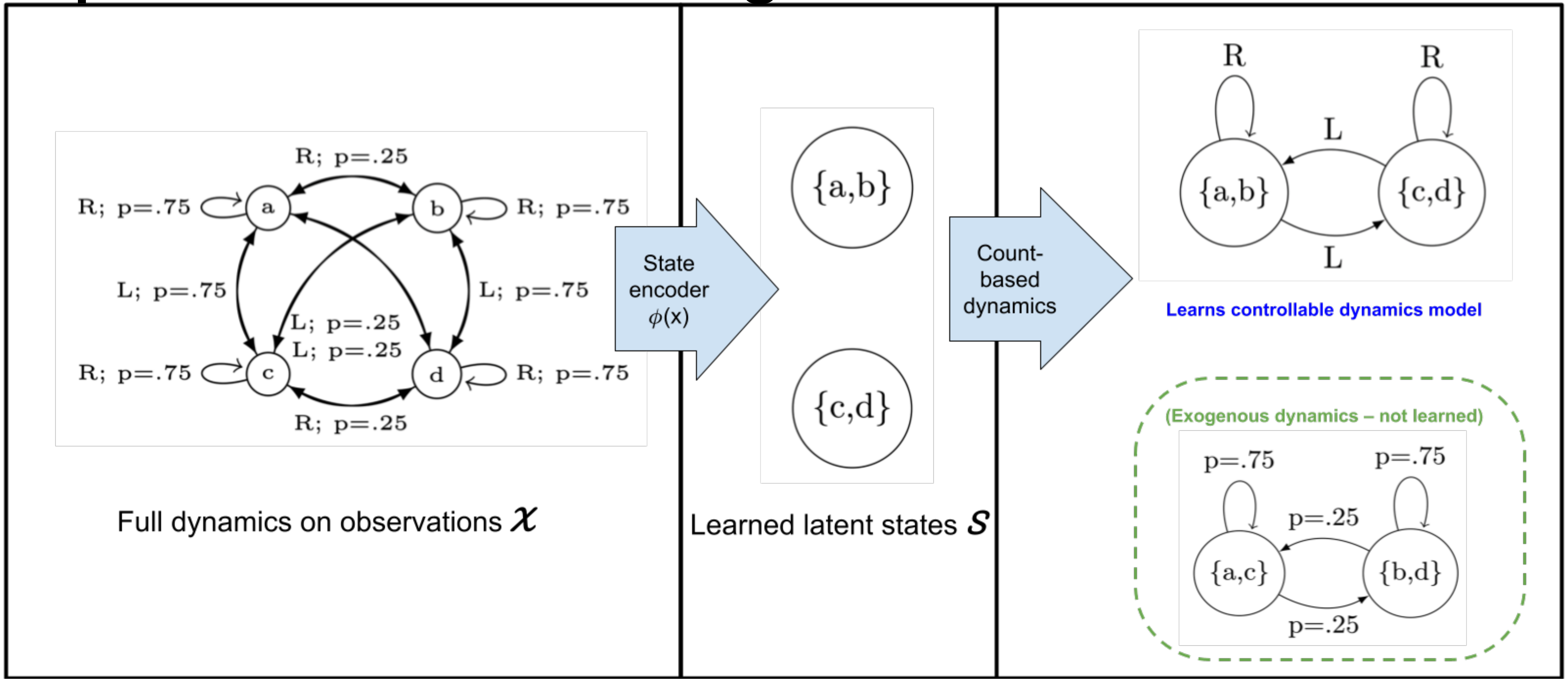
1: The University of Texas at Austin. 2: Sony AI. Correspondence to alevine0@cs.utexas.edu

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



- State x ∈ X can be factored into:
  - Endogenous state s ∈ S, discrete, evolves deterministically according to actions
  - Exogenous state e ∈ $\mathcal{E}$, stochastic, independent of actions (***noise***)
- Factorization is *not* known a priori, and s and e are *not* observed.

Representation Learning In Ex-BMDP Framework



- Learn encoder ϕ that maps x to s
- Dynamics on S can be inferred by counting
- Ignore/don't learn dynamics on $\mathcal{E}$

Representation Learning In Ex-BMDP Framework

- Why learn Control-Endogenous Representation?
  - Interpretability
  - Planning

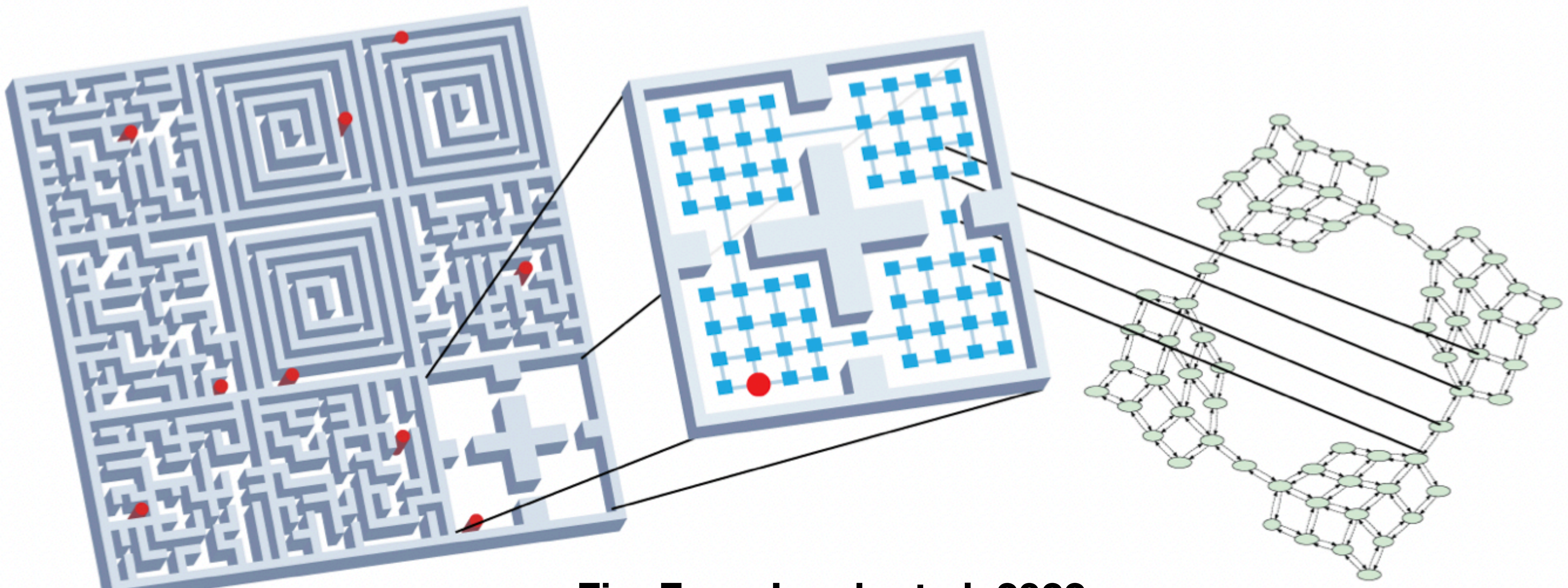


Fig. From Lamb et al. 2022

Representation Learning In Ex-BMDP Framework

- 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') ≤ D**

AC-State (Lamb et al., 2022)

- "Multistep Inverse": predict a_t given $\phi(x_t)$, $\phi(x_{t+k})$, k:

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

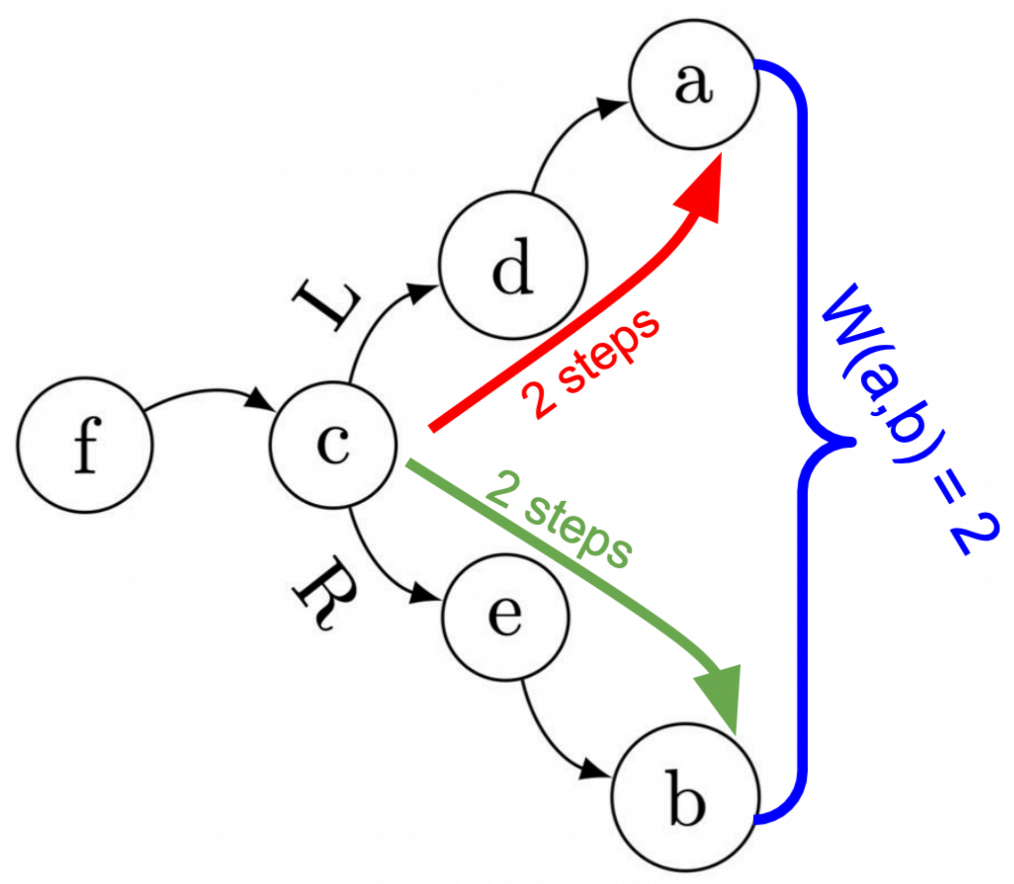
$$\{\theta\}^* := \{\theta^{**} | \theta^{**} = \arg\min_\theta \mathcal{L}_{\text{AC-State}}(\phi_\theta)\}$$

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

- Must show that learned ϕ won't conflate two different states s, s' $\in$ S.

AC-State (Lamb et al., 2022)

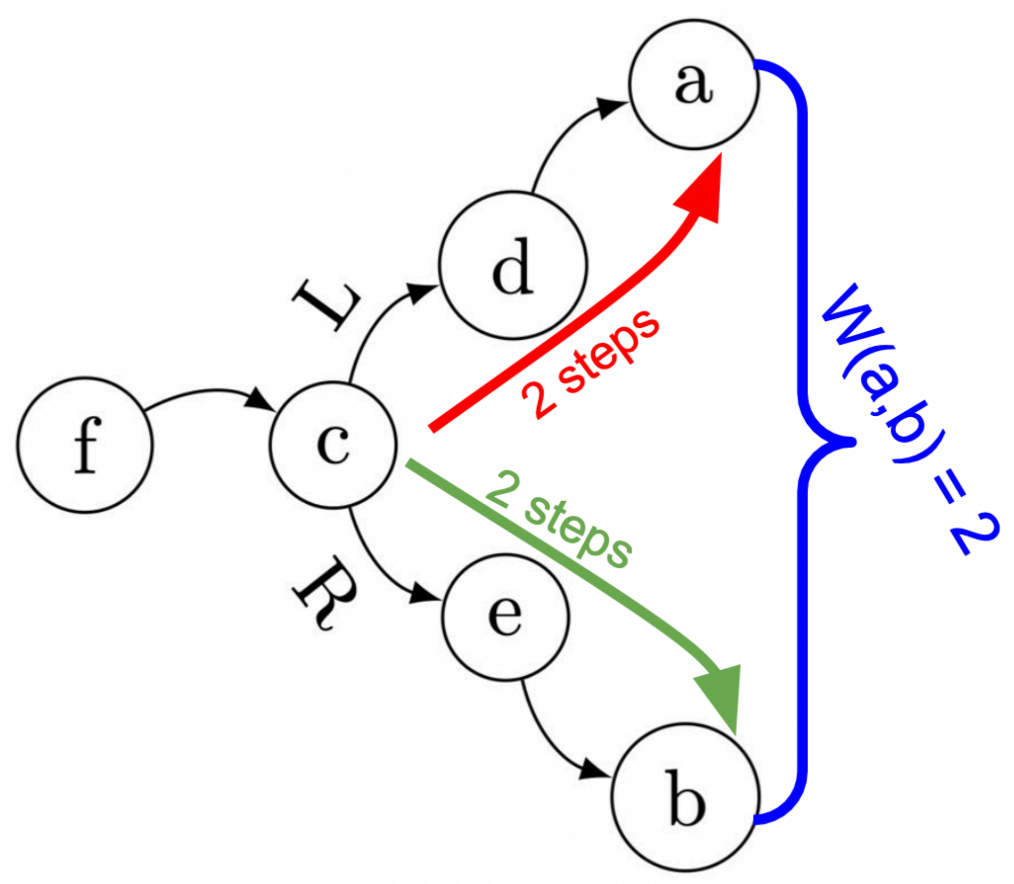
- **Proof Sketch (re-framed):**
  - For a,b ∈ S, Let "witness distance" W(a,b) be the minimum k such that ∃c∈ S, such that a and b can both be reached from c in exactly k steps.



  - Compare $P(a_t \mid s_t = c, s_{t+k} = a)$ vs. $P(a_t \mid s_t = c, s_{t+k} = b)$
  - Distributions have *disjoint support!* Otherwise $W(a,b) < k$. Therefore ϕ must distinguish a, b.
  - Bounded diameter: $\forall$ a,b ∈ S, $W(a,b) \leq D \rightarrow k \sim U(\{1,\ldots D\})$ steps is sufficient.

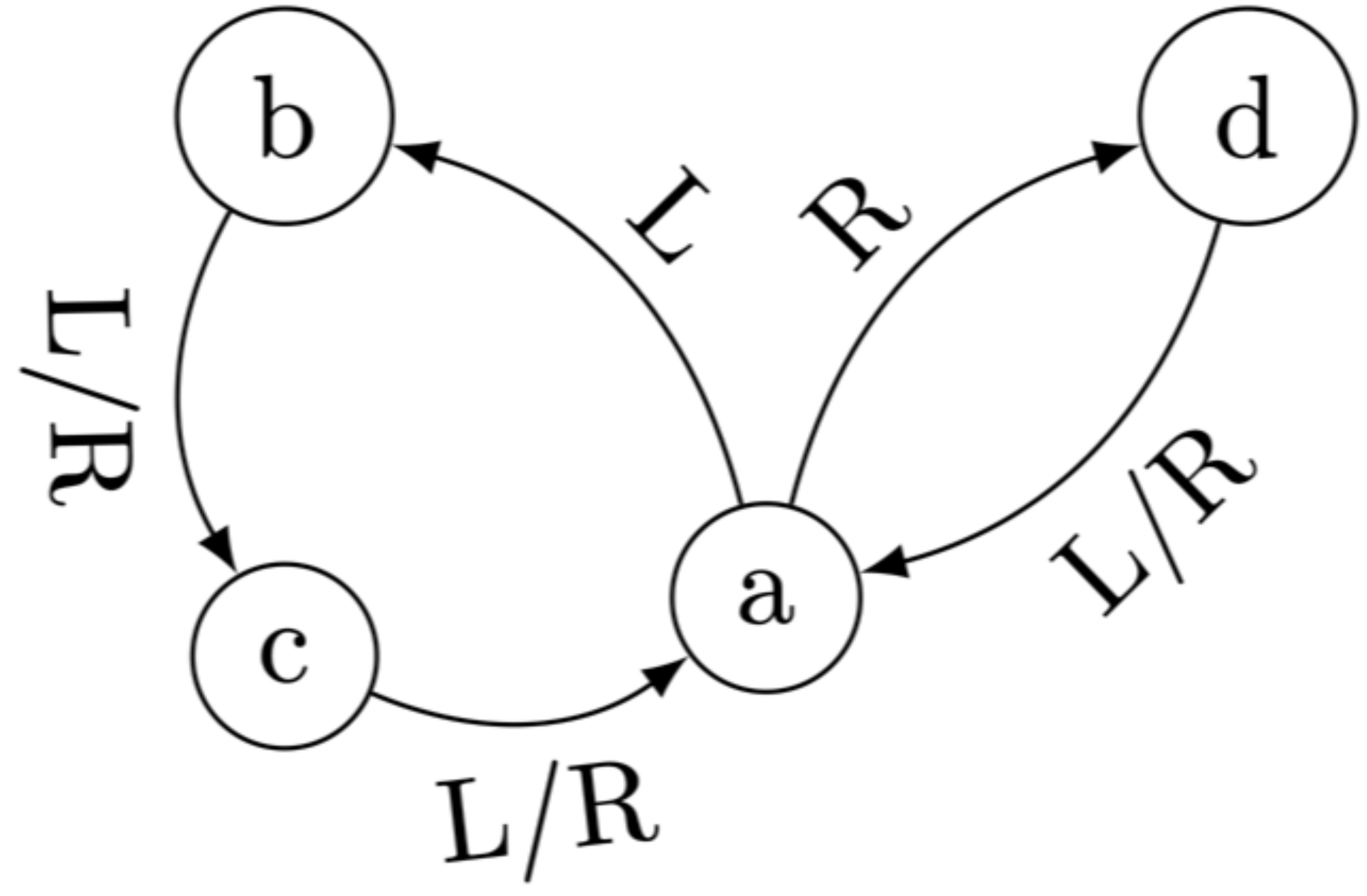
AC-State (Lamb et al., 2022)

- **Proof Sketch (re-framed):**
  - For a,b ∈ S, Let "witness distance" W(a,b) be the minimum k such that ∃c∈ S, such that a and b can both be reached from c in exactly k steps.

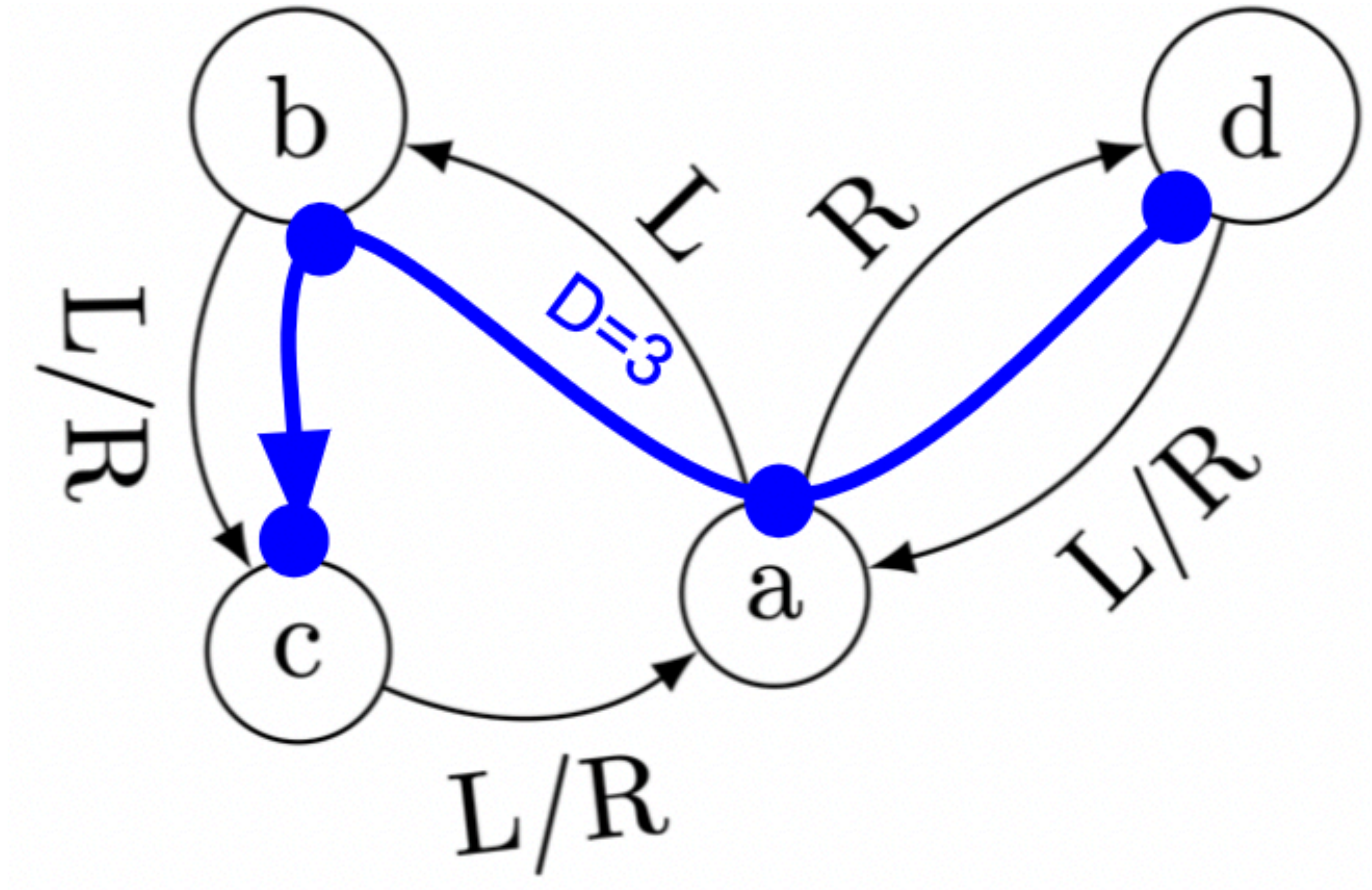


  - Compare $P(a_t \mid s_t = c, s_{t+k} = a)$ vs. $P(a_t \mid s_t = c, s_{t+k} = b)$
  - Distributions have *disjoint support!* Otherwise W(a,b) < k. Therefore ϕ must distinguish a, b.
  - Bounded diameter: ~~∀ a,b ∈ S, W(a,b) < D~~ → k ~ U({1,…D}) steps is sufficient.

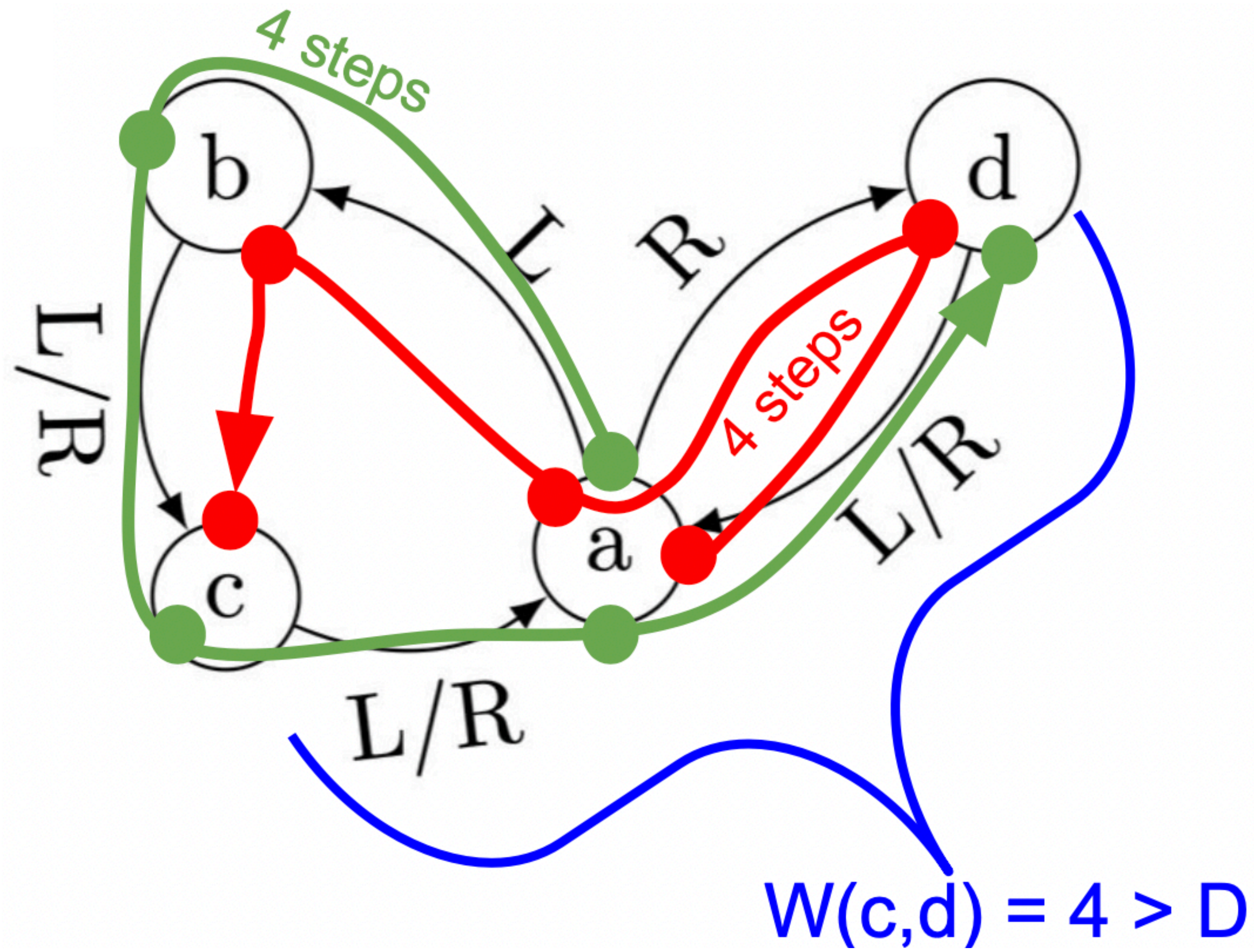
D Steps is Not All You Need



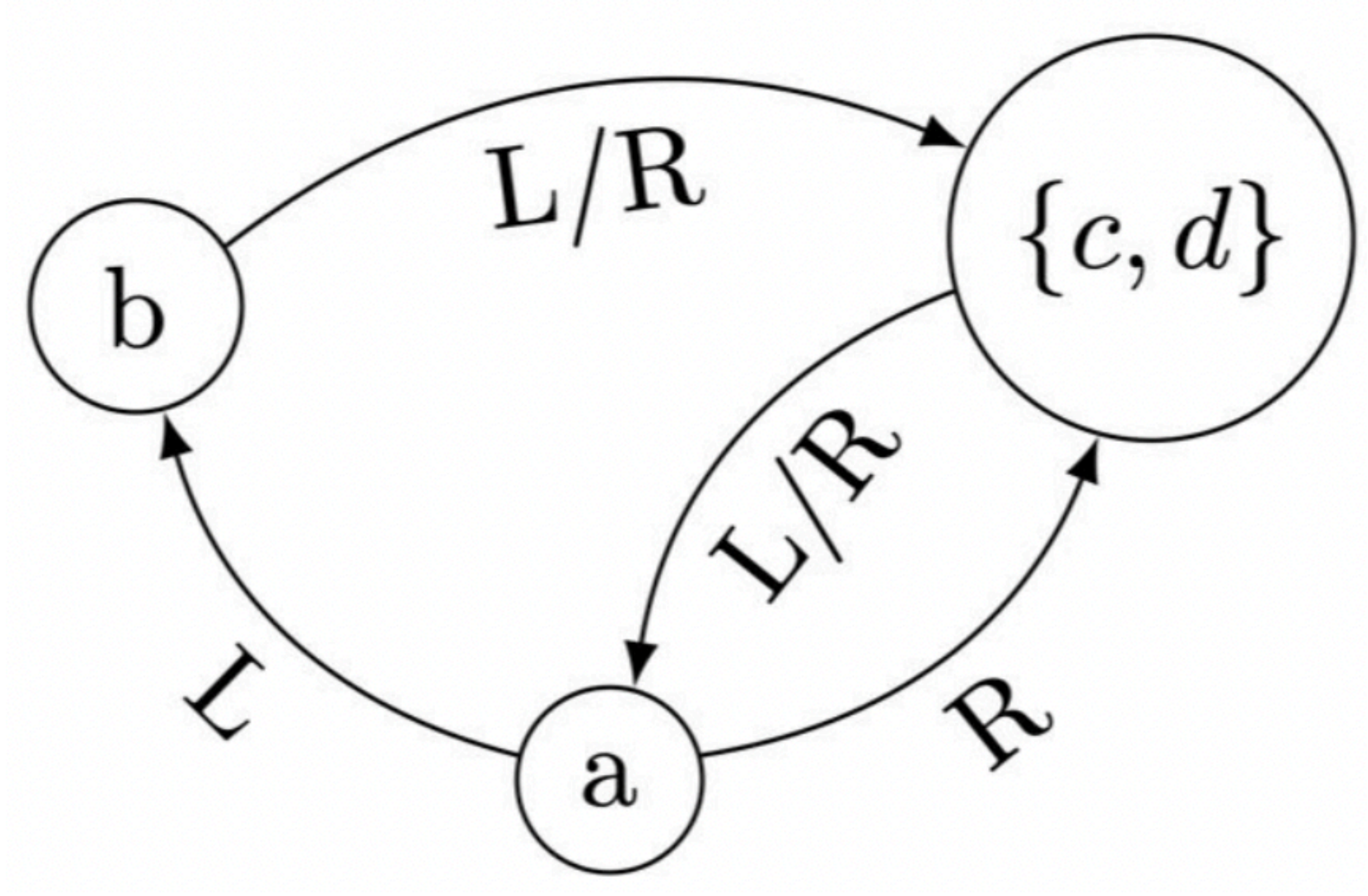
D Steps is Not All You Need



D Steps is Not All You Need



D Steps is Not All You Need



- AC-State with K=D=3 learns ***incorrect*** encoder that conflates c and d.
  - Encoder is incorrect, because we are able to control whether we're in state c or state d, but this representation doesn't show this

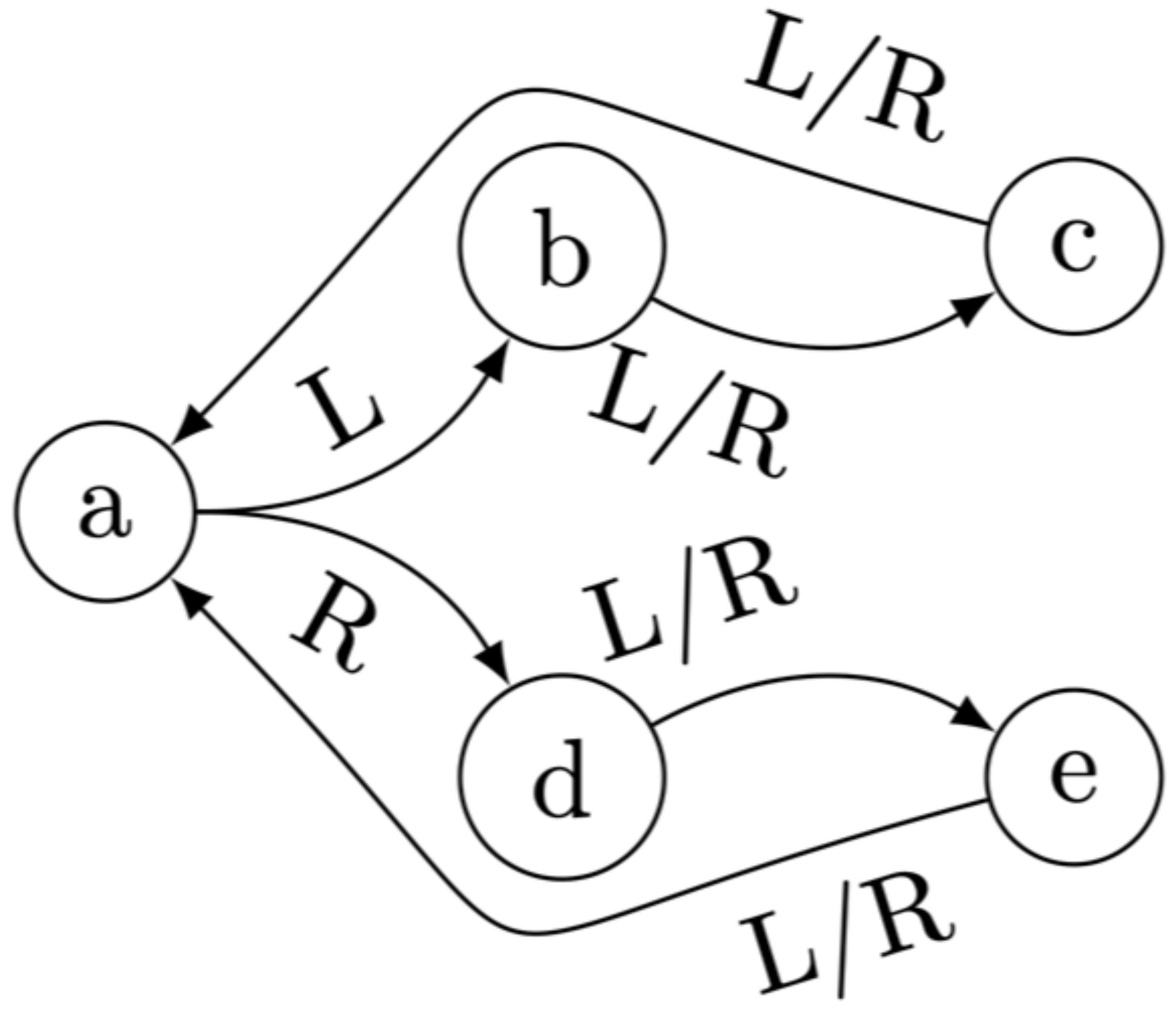
D Steps is Not All You Need

- In practice, D is not known a priori; max number of steps used is hyperparameter K.
- If not D, how many steps do we need?
- **Theorem**: If W(a,b) is finite, then $W(a,b) \leq 2D^2 + D$
  - Tight up to constant factor: we can construct dynamics where AC-State fails using $K = D^2/2 + O(D)$ steps for arbitrarily large D

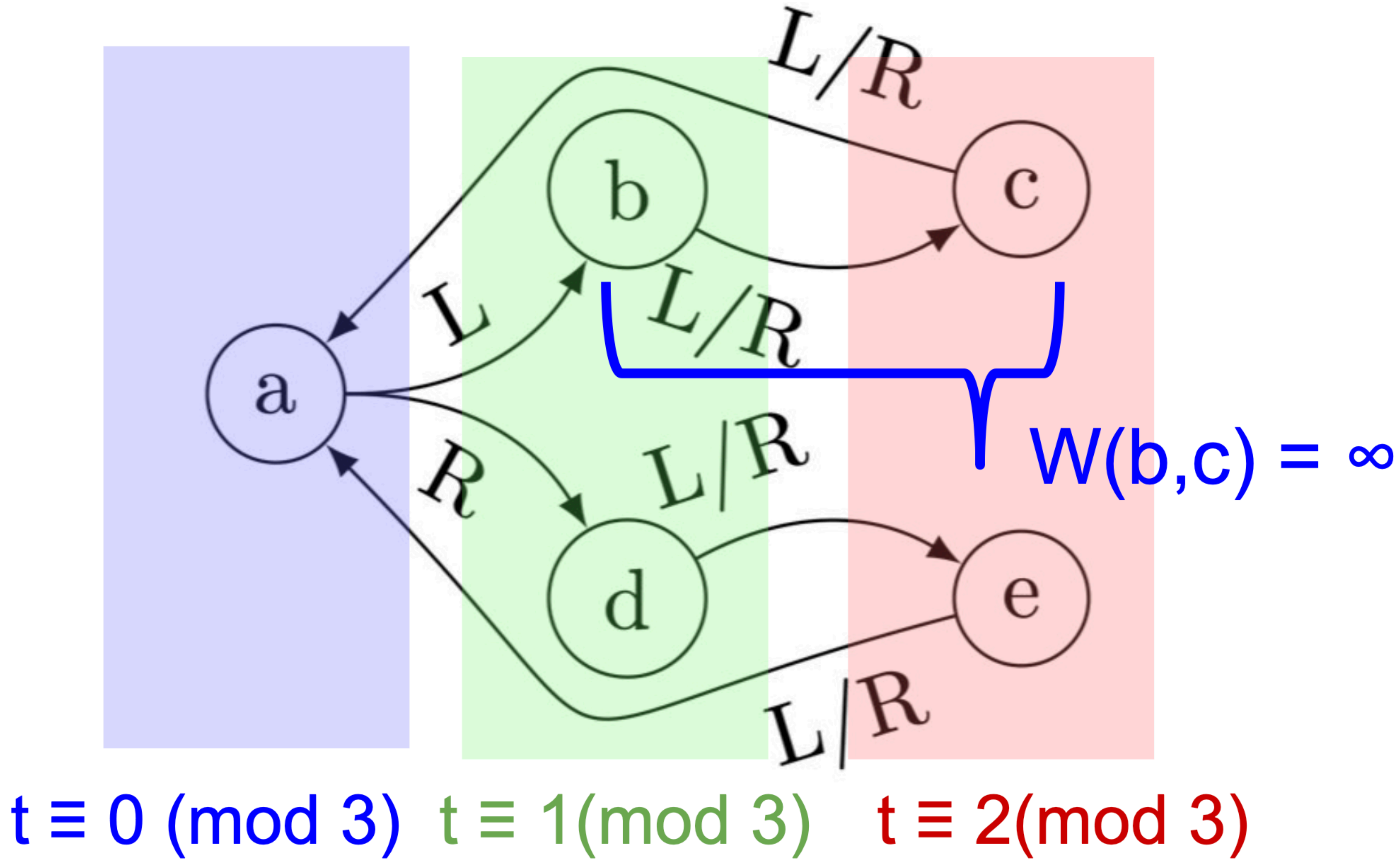
D Steps is Not All You Need

- In practice, D is not known a priori; max number of steps used is hyperparameter K.
- If not D, how many steps do we need?
- **Theorem**: ***If W(a,b) is finite***, then $W(a,b) \leq 2D^2 + D$
    - Tight up to constant factor: we can construct dynamics where AC-State fails using $K = D^2/2 + O(D)$ steps for arbitrarily large D

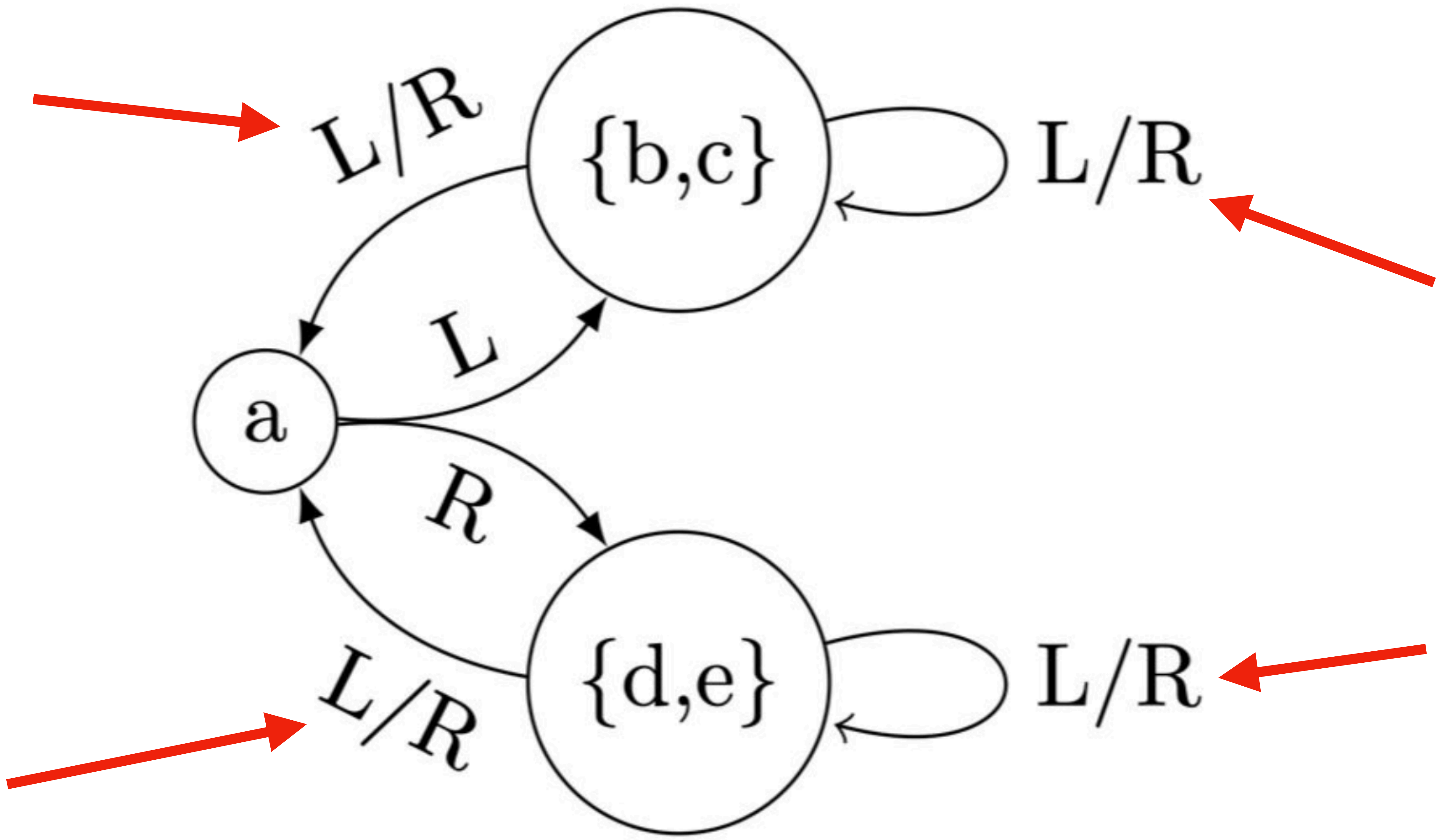
Multistep Inverse is Not All You Need



Multistep Inverse is Not All You Need



Multistep Inverse is Not All You Need



- Dynamics learned with AC-State (***For any K***) not deterministic: ***not a valid*** endogenous latent representation.

ACDF

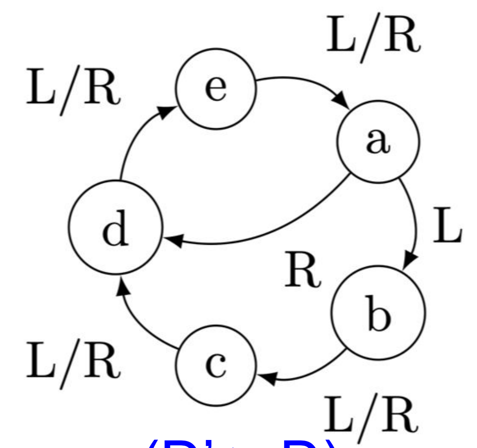
- New algorithm to fix AC-State:

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

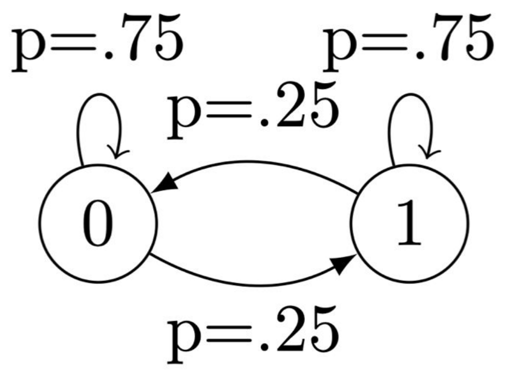
- Where:
    - D is replaced by D', any upper bound on finite witness distances (can use $D' := 2D^2+D$; in practice, a hyperparameter.)
    - Added latent forward model g: predict $\phi(x_{t+1})$ given $\phi(x_t)$ and a_t
- **AC**-State + **D'** + **Forward** model = **ACDF**
- **Theorem (informal)**: Encoders which minimize ACDF loss encode a correct endogenous latent representation.

Results: Tabular

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



(D' > D)

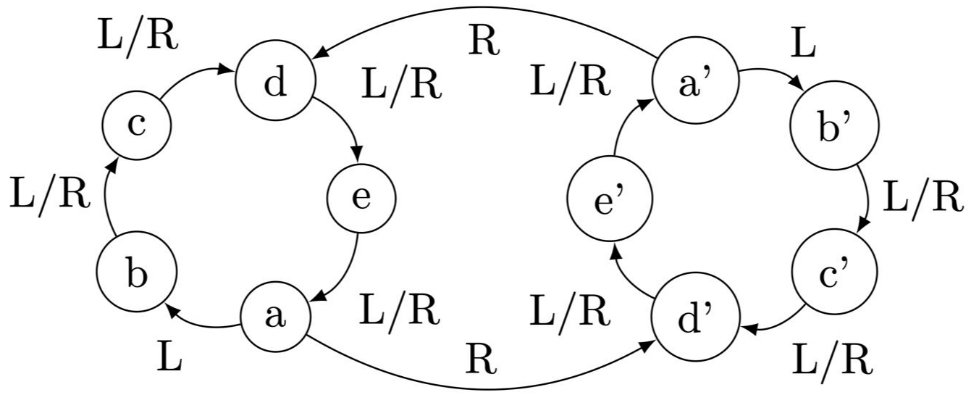


AC-State Success Rate

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)

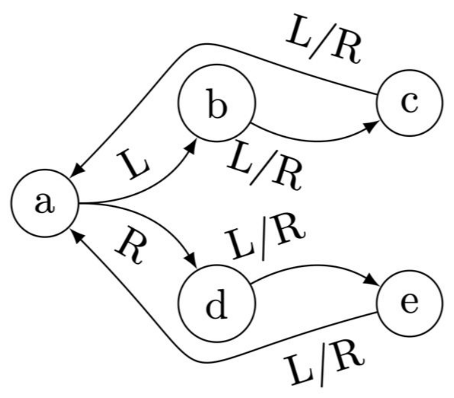
(None)

AC-State Success Rate

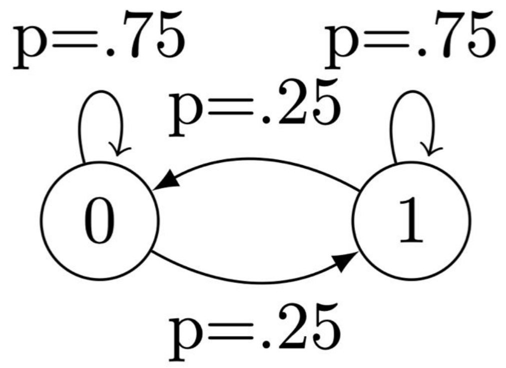
Env. steps:	1000	2000	4000	8000	16000
K=10	0%	0%	0%	0%	0%
K=13	0%	0%	0%	0%	0%
K=16	0%	0%	0%	0%	0%
K=19	0%	0%	2%	0%	0%
K=22	0%	0%	2%	54%	98%
K=25	0%	0%	0%	18%	80%
K=28	0%	0%	0%	4%	38%

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)

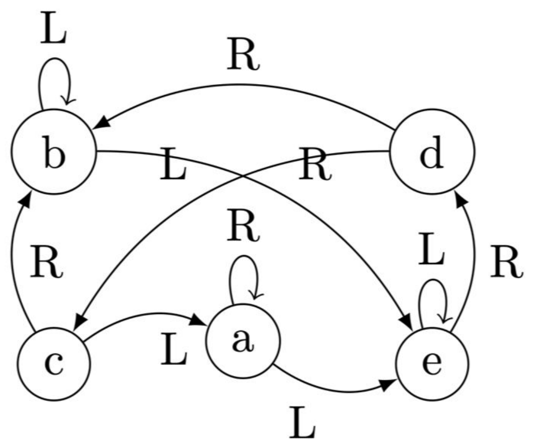


AC-State Success Rate

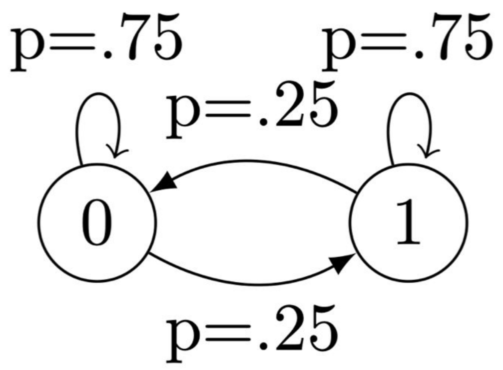
Env. steps:	100	200	400	800	1600
K=1	0%	0%	0%	0%	0%
K=2	0%	0%	0%	0%	0%
K=3	0%	0%	0%	0%	0%
K=4	0%	0%	0%	0%	0%

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

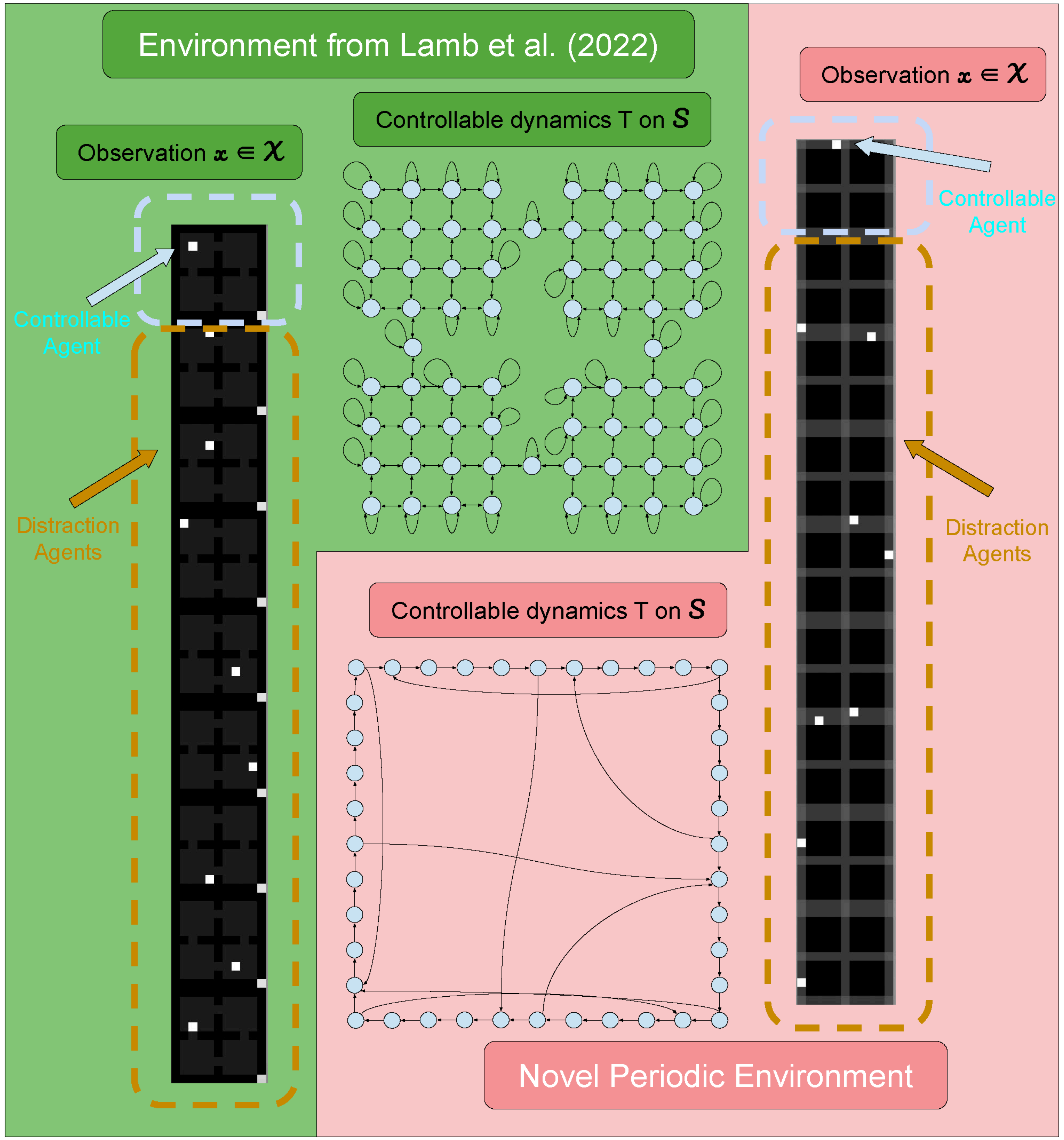
Env. steps:	100	200	400	800	1600
K=1	98%	100%	100%	100%	100%
K=2	91%	100%	100%	100%	100%
K=3	68%	100%	100%	100%	100%
K=4	18%	88%	100%	100%	100%
K=5	4%	50%	98%	100%	100%

Results: Deep Learning

- 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 " "

Results: Deep Learning



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.