

There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning

Nathan Grinsztajn, Johan Ferret,
Olivier Pietquin, Philippe Preux,
Matthieu Geist

Reading notes

Malik-Manel Hashim

Agenda

- Overview
- Approach
- Reversibility and Reversibility Estimation
- Reversibility-Aware Reinforcement Learning
- Experiments
- Summary
- Conclusion and Critique

Agenda

- Overview
- ~~Approach~~
- ~~Reversibility and Reversibility Estimation~~
- ~~Reversibility Aware Reinforcement Learning~~
- ~~Experiments~~
- ~~Summary~~
- Conclusion and Critique

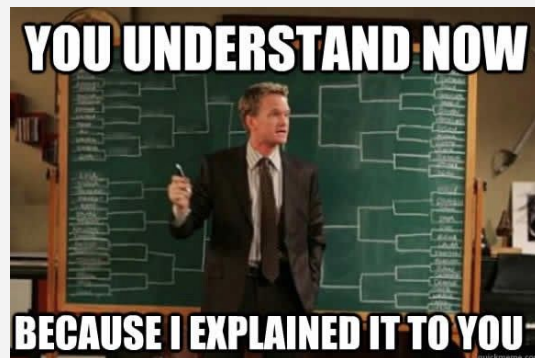
1. *Idea*
2. *Math*
3. *How to RL*
4. *Tests and results*



[2]

Overview

- Knowing reversibility of an action = Knowing its potential risk
- A always before B $\Rightarrow A \rightarrow B$ not reversible
 - Simple binary classification
- Can be used for exploration / control
- Performs great (on what they tested it on)

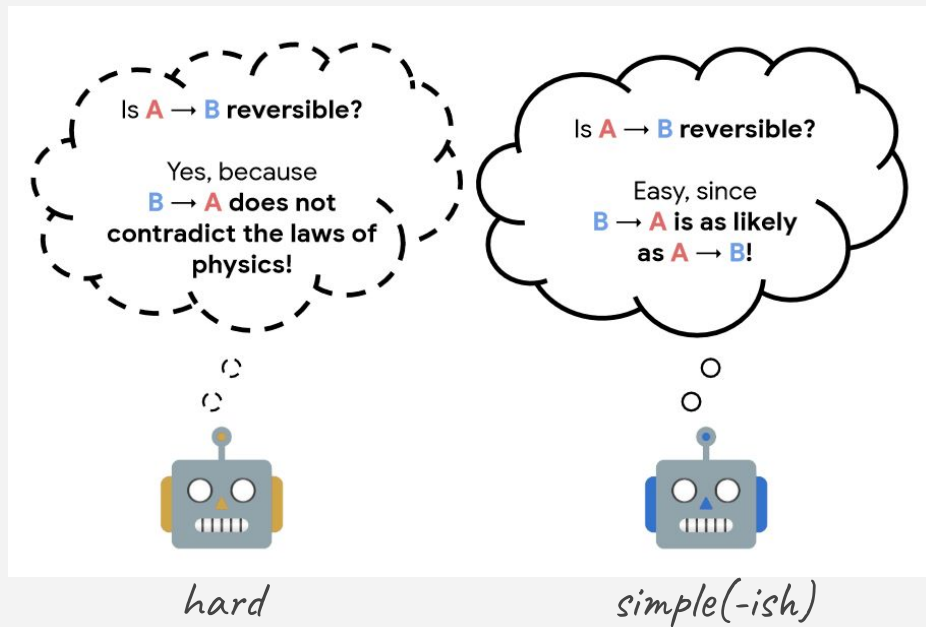


(The Idea)

Approach

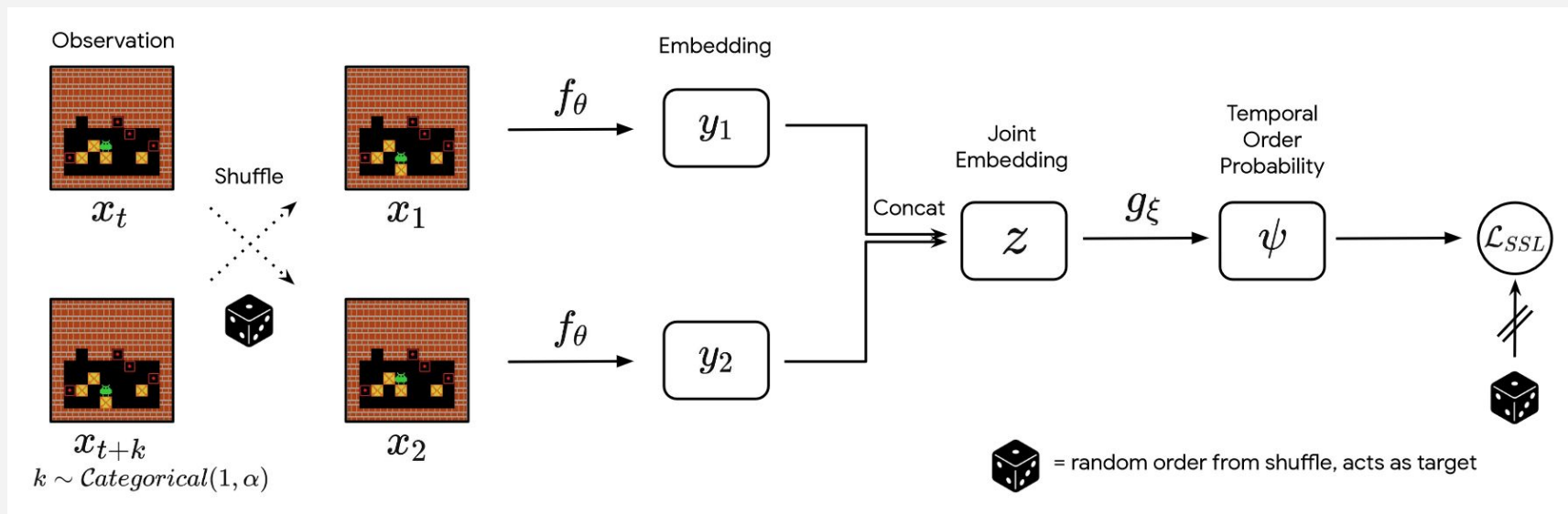
- Reversibility = Safety
- Approximation via temporal order
- Can be learned through a surrogate task

[1, p. 2]



(The Idea)

Approach



(The Math)

Reversibility and Reversibility Estimation

Degree of reversibility within K steps

$$\phi_K(s, a) := \sup_{\pi} p_{\pi}(s \in \tau_{t+1:t+K+1} \mid s_t = s, a_t = a)$$

Degree of reversibility

$$\phi(s, a) := \sup_{\pi} p_{\pi}(s \in \tau_{t+1:\infty} \mid s_t = s, a_t = a)$$

(The Math)

Reversibility and Reversibility Estimation

Degree of reversibility within K steps

$$\phi_{\pi,K}(s, a) := \sup_{\pi} p_{\pi}(s \in \tau_{t+1:t+K+1} \mid s_t = s, a_t = a)$$

Degree of reversibility

$$\phi_{\pi}(s, a) := \sup_{\pi} p_{\pi}(s \in \tau_{t+1:\infty} \mid s_t = s, a_t = a)$$

(The Math)

Reversibility and Reversibility Estimation

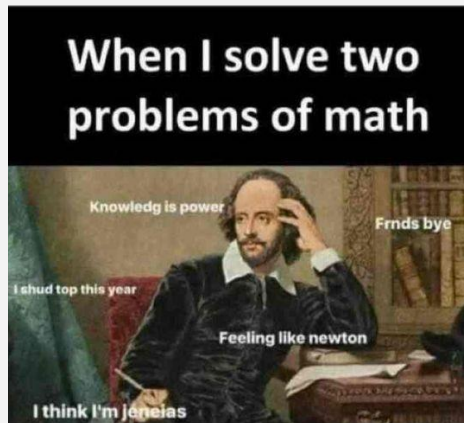
- Reversibility can be predicted via *precedence*
→ s or s' first *on average*

Finite-horizon precedence estimator

$$\psi_{\pi,T}(s, s') = \mathbb{E}_{\tau \sim \pi} \mathbb{E}_{s_t=s, s_{t'}=s' \atop t, t' < T} [\mathbb{1}_{t' > t}]$$

Empirical reversibility

$$\bar{\phi}_{\pi}(s, a) = \mathbb{E}_{s' \sim P(s, a)} [\psi_{\pi}(s', s)]$$



[4]

(The Math)

Reversibility and Reversibility Estimation

Reversibility: $\phi_{\pi}(s, a) := p_{\pi}(s \in \tau_{t+1:\infty} \mid s_t = s, a_t = a)$

Empirical reversibility: $\bar{\phi}_{\pi}(s, a) = \mathbb{E}_{s' \sim P(s, a)} [\psi_{\pi}(s', s)]$

Relation of reversibility and empirical reversibility

$$\bar{\phi}_{\pi}(s, a) \geq \frac{\phi_{\pi}(s, a)}{2}$$

(The Math)

Reversibility and Reversibility Estimation

Reversibility: $\phi_{\pi}(s, a) := p_{\pi}(s \in \tau_{t+1:\infty} \mid s_t = s, a_t = a)$ [5]

Empirical reversibility: $\bar{\phi}_{\pi}(s, a) = \mathbb{E}_{s' \sim P(s, a)} [\psi_{\pi}(s', s)]$

Relation of reversibility and empirical reversibility

$$\bar{\phi}_{\pi}(s, a) \geq \frac{\phi_{\pi}(s, a)}{2}$$



Reversibility and Reversibility Estimation

A.3 Proofs of Theorem 1 and Theorem 2

In the following, we prove simultaneously Theorem 1 and Theorem 2. We begin by two lemmas.

Lemma 1. Given a trajectory τ , we denote by $\#_T(s \rightarrow s')$ the number of pairs (s, s') in $\tau_{1:T}$ such that s appears before s' . We present a simple formula for $\psi(s', s)$ according to the structure of the state trajectory:

$$\psi_{s,T}(s, s') = \frac{\mathbb{E}_{\tau \sim \pi} [\#_T(s \rightarrow s')]}{\mathbb{E}_{\tau \sim \pi} [\#_T(s \rightarrow s') + \#_T(s' \rightarrow s)]}.$$

Proof. In order to simplify the notations, we leave implicit the fact that indices are always sampled within $[0, T]$:

$$\begin{aligned} \psi_{s,T}(s, s') &= \mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s, s_{t'} = s'} [\mathbb{1}_{t' > t}]] \\ &= \frac{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{t' > t} \mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}. \end{aligned}$$

Similarly, we have:

$$\mathbb{E}_\pi \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}] = \frac{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{t' > t} \mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}{\mathbb{E}_{t \neq t'} [\mathbb{1}_{t' > t}]}.$$

Combining it with our previous equation:

$$\begin{aligned} \psi_{s,T}(s, s') &= \frac{\mathbb{E}_\pi \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}] \mathbb{E}_{t \neq t'} [\mathbb{1}_{t' > t}]}{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]} \\ &= \frac{1}{2} \frac{\mathbb{E}_\pi \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}. \end{aligned}$$

Looking at the denominator, we can notice:

$$\begin{aligned} \mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}] &= \frac{1}{2} \mathbb{E}_\pi \mathbb{E}_{t < t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}] + \frac{1}{2} \mathbb{E}_\pi \mathbb{E}_{t' < t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}], \\ &= \frac{1}{2} \mathbb{E}_\pi \mathbb{E}_{t < t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'} + \mathbb{1}_{s_{t'} = s} \mathbb{1}_{s_t = s'}], \end{aligned}$$

which comes from the fact that t and t' play a symmetrical role. Thus,

$$\psi_{s,T}(s, s') = \frac{\mathbb{E}_{\tau \sim \pi} \mathbb{E}_t \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}{\mathbb{E}_{\tau \sim \pi} \mathbb{E}_t \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'} + \mathbb{1}_{s_{t'} = s} \mathbb{1}_{s_t = s'}]}.$$

Since

$$\begin{aligned} \mathbb{E}_{\tau \sim \pi} [\#_T(s \rightarrow s')] &= \sum_{i < j \leq T} \mathbb{1}_{s_i = s} \mathbb{1}_{s_j = s'}, \\ &= \binom{T}{2} \sum_{i < j \leq T} \left(\frac{1}{\binom{T}{2}} \mathbb{1}_{s_i = s} \mathbb{1}_{s_j = s'} \right), \\ &= \binom{T}{2} \mathbb{E}_{\tau \sim \pi} \mathbb{E}_t \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}], \end{aligned}$$

Reversibility and Reversibility Estimation

$$\psi_{s,r}(s,s') = \frac{\binom{2}{2} E_{r \rightarrow s} \{ \#r(s \rightarrow s') \}}{\binom{2}{2} E_{r \rightarrow s} \{ \#r(s \rightarrow s') + \#r(s' \rightarrow s) \}}$$

$$\psi_{s,r}(s,s') = \frac{E_{r \rightarrow s} \{ \#r(s \rightarrow s') \}}{E_{r \rightarrow s} \{ \#r(s \rightarrow s') + \#r(s' \rightarrow s) \}}.$$

where s appears

Lemma 2. Assume that we are given a fixed trajectory where s appears n_s times, s_1, s_2, \dots, s_{n_s} are the i th occurrences of s and $s_{n_s+1}, s_{n_s+2}, \dots, s_{n_s+n_s+1}$ are the $(i+1)$ th occurrence of s . In this case,

$$\pi(s \rightarrow s) = \sum_{i=1}^{n_s} \pi(s_i \rightarrow s) + \sum_{i=1}^{n_s} \pi(s_{n_s+i} \rightarrow s). \quad (5)$$

where $n_i(s)$ denotes the number of times s appears between s_i and s_{i+1} . We also have

$$\pi(s_i \rightarrow s) = \dots = \pi(s_{n_s+i} \rightarrow s) = \pi(s_{n_s+i+1} \rightarrow s).$$

Thus,

$$\pi(s \rightarrow s) = \sum_{i=1}^{n_s} \pi(s_i \rightarrow s) + \sum_{i=1}^{n_s} \pi(s_{n_s+i} \rightarrow s) = \sum_{i=1}^{n_s} \pi(s_i \rightarrow s) + \sum_{i=1}^{n_s} \pi(s_i \rightarrow s). \quad (6)$$
[illegible]

we suppose that $n_1(x^i) = n_2(x^i) = 0$. Then $\# \pi(x^i \rightarrow x^j) = \sum_{k=1}^n \sum_{l=1}^n n_1(x^k) \cdot n_2(x^l) = 0$. Thus, $\# \pi(x^i \rightarrow x^j) = \# \pi(x^j \rightarrow x^i)$. \square

Proof. Eq (4) comes directly from $\# \pi(x \rightarrow x') = \sum_{k=1}^n \sum_{l=1}^n n_1(x^k) \cdot n_2(x^l) = \sum_{k=1}^n n_1(x^k) \cdot \sum_{l=1}^n n_2(x^l) = \sum_{k=1}^n n_1(x^k) \cdot \# \pi(x' \rightarrow x)$. Thus, Exp. (5), we first notice that $\# \pi(x \rightarrow x') + \# \pi(x' \rightarrow x) = k \cdot \sum_{k=1}^n n_1(x^k) = 2 \cdot \sum_{k=1}^n n_1(x^k) \cdot \sum_{l=1}^n n_2(x^l) = 2 \cdot \# \pi(x \rightarrow x') = 2 \cdot \# \pi(x' \rightarrow x)$. \square

Thus, $\# \pi(x \rightarrow x') - \# \pi(x' \rightarrow x) = 2 \cdot \sum_{k=1}^n n_1(x^k) \cdot \sum_{l=1}^n n_2(x^l) - \left(\sum_{k=1}^n n_1(x^k) + k \cdot n_2(x') + k \cdot (k-1) \cdot n_1(x') \right) = 2 \cdot \left(k \cdot n_2(x') + n_1(x') \cdot \sum_{l=1}^n n_2(x^l) \right) - (k \cdot n_2(x') + k \cdot n_2(x') + k \cdot (k-1) \cdot n_1(x')) = k \cdot n_2(x') - k \cdot n_2(x')$.

Since $\sum_{k=1}^n n_1(x^k) \cdot \sum_{l=1}^n n_2(x^l)$ converges when T goes to infinity, $\sum_{k=1}^n n_1(x^k) \cdot \sum_{l=1}^n n_2(x^l) \rightarrow 0$, we can link reversibility and empirical

[illegible]

Proof. For a policy π , let $\hat{\psi}_t(s, s')$ denote $\hat{\mathbb{P}}_t(s' | s, \pi)$. In obtaining (17), we used the reversibility with respect to $\hat{\psi}_t(s, s')$. In obtaining (18), we used the fact that $\hat{\psi}_t(s, s') = \hat{\mathbb{P}}_t(s' | s, \pi)$ if and only if $\hat{\psi}_t(s', s) = \hat{\mathbb{P}}_t(s | s', \pi)$. For a policy π , let $\hat{\psi}_t(s, s')$ denote $\hat{\mathbb{P}}_t(s' | s, \pi)$. In obtaining (17), we used the reversibility with respect to $\hat{\psi}_t(s, s')$. In obtaining (18), we used the fact that $\hat{\psi}_t(s, s') = \hat{\mathbb{P}}_t(s' | s, \pi)$ if and only if $\hat{\psi}_t(s', s) = \hat{\mathbb{P}}_t(s | s', \pi)$. For a policy π , let $\hat{\psi}_t(s, s')$ denote $\hat{\mathbb{P}}_t(s' | s, \pi)$. In obtaining (17), we used the reversibility with respect to $\hat{\psi}_t(s, s')$. In obtaining (18), we used the fact that $\hat{\psi}_t(s, s') = \hat{\mathbb{P}}_t(s' | s, \pi)$ if and only if $\hat{\psi}_t(s', s) = \hat{\mathbb{P}}_t(s | s', \pi)$.

A.3 Proofs of Theorem 1 and Theorem 2

lemma 1. Given a trajectory τ , we denote by $\#_{\tau}(s \rightarrow s')$ the number of pairs (s, s') in $\tau_{1:T}$ such that s appears before s' . We present a simple formula for $\psi(s', s)$ according to the structure of the trajectory:

$$\psi_{\pi,T}(s,s') = \frac{\mathbb{E}_{\tau \sim \pi} [\#_T(s \rightarrow s')]}{\mathbb{E}_{\tau \sim \pi} [\#_T(s \rightarrow s') + \#_T(s' \rightarrow s)]}.$$

In order to simplify the notations, we leave implicit the fact that indices are always sampled from T .

$$\begin{aligned}\psi_{\pi,T}(s,s') &= \mathbb{E}_{\pi} \mathbb{E}_{t \neq t' | s_t = s, s_{t'} = s'} [\mathbf{1}_{t' > t}], \\ &= \frac{\mathbb{E}_{\pi} \mathbb{E}_{t \neq t'} [\mathbf{1}_{t' > t} \mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}]}{\mathbb{E}_{\pi} \mathbb{E}_{t \neq t'} [\mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}]}.\end{aligned}$$

$$E_{\pi} E_{t' > t} [1_{s_t = s} 1_{s_{t'} = s'}] = \frac{E_{\pi} E_{t' \neq t} [1_{t' > t} 1_{s_t = s} 1_{s_{t'} = s'}]}{E_{t' \neq t} [1_{t' > t}]},$$

previous equation

$$T(s, s') = \frac{\mathbb{E}_{\pi} \mathbb{E}_{t' > t} [\mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}] \mathbb{E}_{t \neq t'} [\mathbf{1}_{t' > t}]}{\mathbb{E}_{\pi} \mathbb{E}_{t \neq t'}, [\mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}]} \\ = \frac{1}{2} \frac{\mathbb{E}_{\pi} \mathbb{E}_{t' > t} [\mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}]}{\mathbb{E}_{\pi} \mathbb{E}_{t \neq t'} [\mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}]}.$$

an notice:

$$\begin{aligned} & \mathbb{E}_{\pi} \mathbb{E}_{t < t'} [1_{s_t=s} 1_{s_{t'}=s'}] + \frac{1}{2} \mathbb{E}_{\pi} \mathbb{E}_{t < t'} [1_{s_t=s} 1_{s_{t'}=s'}], \\ & \mathbb{E}_{\pi} \mathbb{E}_{t < t'} [1_{s_t=s} 1_{s_{t'}=s'} + 1_{s_t=s'} 1_{s_{t'}=s}], \end{aligned}$$

play a symmetrical role. Thus,

$$= \binom{T}{2} \mathbb{E}_{\tau \sim \mathcal{H}} \mathbb{E}_{t' \sim t} \left[\frac{\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}}{\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'} + \mathbb{1}_{s_t = s'} \mathbb{1}_{s_{t'} = s}} \right],$$

Reversibility and Reversibility Estimation

$$\psi_{s,r}(s,s') = \frac{\binom{2}{2} E_{r \rightarrow s} \{ \#r(s \rightarrow s') \}}{\binom{2}{2} E_{r \rightarrow s} \{ \#r(s \rightarrow s') + \#r(s' \rightarrow s) \}}$$

$$\psi_{s,r}(s,s') = \frac{E_{r \rightarrow s} \{ \#r(s \rightarrow s') \}}{E_{r \rightarrow s} \{ \#r(s \rightarrow s') + \#r(s' \rightarrow s) \}}.$$

where s appears

lemma 2. Assume that we are given a fixed trajectory where s appears n_s times, $s_0 \xrightarrow{n_0(x)} s \xrightarrow{n_1(x)} s \xrightarrow{n_2(x)} \dots \xrightarrow{n_{i-1}(x)} s \xrightarrow{n_i(x)}$ and the $(i+1)^{\text{th}}$ occurrence of s . In this case,

$$\#(s \rightarrow s') = \sum_{i=0}^k i \cdot n_i(s'),$$

where $n_i(s')$ denotes the number of times s' appears between the i^{th} and the $(i+1)^{\text{th}}$ occurrence of s . In this case,

$$n_0(s') = \dots = n_{i-1}(s') = 0, \text{ we also have}$$

$$n_i(s' \rightarrow s) = k \cdot n_i(s') = \sum_{j=i+1}^k n_j(s') = \sum_{j=i+1}^k n_j(s'). \text{ Thus}$$

we suppose that $n_1(s') = n_1(s) - \dots - \#(\tau \circ s' \rightarrow s) = \sum_{i=1}^k n_i(s) - \sum_{i=1}^k \sum_{j=1}^i n_j(s)$. Thus,

$$\#(\tau \circ s' \rightarrow s) = \#(\tau \circ s \rightarrow s') + \#(s' \rightarrow s) = \sum_{i=1}^k n_i(s) - k \times \sum_{i=1}^k n_i(s) + \#(\tau \circ s' \rightarrow s).$$

Proof. Eq. (4) comes directly from $\#(\tau \circ s \rightarrow s') = \sum_{i=1}^k n_i(s) - \sum_{i=1}^k \sum_{j=1}^i n_j(s)$. Thus,

Proof. Eq. (5), we first notice that $\#(\tau \circ s' \rightarrow s) = 2 \times \#(\tau \circ s' \rightarrow s) - (\#(\tau \circ s \rightarrow s') - (\#(\tau \circ s' \rightarrow s) + (k-1)n_1(s')))$. Then,

$$\begin{aligned} \#(\tau \circ s' \rightarrow s) - \#(\tau \circ s' \rightarrow s) &= 2 \times \#(\tau \circ s' \rightarrow s) - \left(\#(\tau \circ s \rightarrow s') - \left(\#(\tau \circ s' \rightarrow s) + (k-1)n_1(s') \right) \right) \\ &= 3 \left(k n_1(s') + n_1(s) - \sum_{i=1}^k \sum_{j=1}^i n_j(s) \right) - (k n_1(s') + k n_1(s) + (k-1)n_1(s')) \\ &= k n_1(s') - k n_1(s). \end{aligned}$$

Thus, $\#(\tau \circ s' \rightarrow s) - \#(\tau \circ s' \rightarrow s)$ converges when T goes to infinity.

Thus, we can link reversibility and empiric

[illegible]

We subdivide our problem into four cases, depending on τ_{s-1} .

Case 1: $P_{\theta}(s \in T_{s-1} \mid s_{s-1} = s) < 1$ and $P_{\theta}(s \in T_{s-1} \mid s_{s-1} = s) = 0$. If τ_{s-1} is recurrent for the Markov chain (induced by θ). Informally, this means that if τ_{s-1} is the state s , we tend to see s an infinite number of times, and we only see s a finite number in a given trajectory.

This implies $\phi_{\theta}(s, s) = P_{\theta}(s \in T_{s-1} \mid s_{s-1} = s) = 0$, so recurrent states can only be found after other recurrent states.

$\psi_{\theta}(s', s) = \psi(s', s)$, Eqs. (6) becomes ~ 0 .

In the following, we prove simultaneously Theorem 1 and Theorem 2. We begin by

lemma 1. *Given a trajectory τ , we denote by $\#_T(s \rightarrow s')$ the number of pairs (s, t) such that s appears before s' . We present a simple formula for $\psi(s', s)$ according to the trajectory:*

In order to simplify the notations, we leave implicit the fact that indices are a T].

$$\mathbb{E}_\pi \mathbb{E}_{t' > t} [\mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}] = \frac{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbf{1}_{t' > t} \mathbf{1}_{s_t = s} \mathbf{1}_{s_{t'} = s'}]}{\mathbb{E}_{t \neq t'} [\mathbf{1}_{t' > t}]}$$

$$T(s, s') = \frac{\mathbb{E}_\pi \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}] \mathbb{E}_{t \neq t'} [\mathbb{1}_{t' > t}]}{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]} \\ = \frac{1}{2} \frac{\mathbb{E}_\pi \mathbb{E}_{t' > t} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}{\mathbb{E}_\pi \mathbb{E}_{t \neq t'} [\mathbb{1}_{s_t = s} \mathbb{1}_{s_{t'} = s'}]}.$$
$$\frac{1}{2} \mathbb{E}_{\pi} \mathbb{E}_{t < t'} [\mathbb{1}_{s_t=s} \mathbb{1}_{s_{t'}=s'}] + \frac{1}{2} \mathbb{E}_{\pi} \mathbb{E}_{t' < t} [\mathbb{1}_{s_t=s} \mathbb{1}_{s_{t'}=s'}]$$
$$\frac{\mathbb{E}_t[\mathbb{E}_{t'}[t' \mid \mathbf{1}_{s_t=s} \mathbf{1}_{s_{t'}=s'}]]}{\mathbb{E}_t[\mathbf{1}_{s_t=s} \mathbf{1}_{s_{t'}=s'} + \mathbf{1}_{s_t=s'} \mathbf{1}_{s_{t'}=s}]} \mathbf{1}_{s_t=s} \mathbf{1}_{s_{t'}=s'},$$

$$= \binom{T}{2} \mathbb{E}_{T \sim \pi} \mathbb{E}_t \mathbb{E}_{t' > t} [\mathbf{1}_{x_t = x} \mathbf{1}_{x_{t'} = x'}]$$

Case 3: $p_{\pi}(s \in T_{k+1} \cap \infty) | s_k = 1$ and $p_{\pi}(s' \in T_{k+1} \cap \infty) | s'_k = 1$ (s is recurrent and s' is recurrent for the Markov chain induced by π). We denote by T_k the random variable corresponding to the time of the k^{th} visit to s . A trajectory can be represented as follows:

$$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow \dots \rightarrow n_k(s) \rightarrow s_{T_k} \rightarrow n_{k+1}(s'), \dots$$

where, writing \sim the equality in distribution, $n_2(s') \sim n_3(s') \sim \dots \sim n_k(s')$ and $E_\tau n_2(s') = E_\tau n_3(s') = \dots = E_\tau n_k(s')$ using the strong Markov property. From Lemma 1 we get

We can see from Lemma 2 :

$$\frac{\mathbb{E}_T[\#T_k(s \rightarrow s') - \#T_k(s' \rightarrow s)]}{\mathbb{E}_{T \sim \pi}[\#T_k(s \rightarrow s') + \#T_k(s' \rightarrow s)]} \xrightarrow{k \rightarrow \infty} 0. \quad (7)$$

$$S_0 \xrightarrow{\quad} S \xrightarrow{\quad} S \xrightarrow{\quad} S \xrightarrow{\quad} S \xrightarrow{\quad} \dots \xrightarrow{\quad} S = ST_{\Phi_Y(s)} \xrightarrow{\quad} S_1 \xrightarrow{\quad} S = ST_{\Phi_Y(s+1)}$$

And,

$$\frac{\mathbb{E}_\tau[\#\tau(s \rightarrow s') - \#\tau(s' \rightarrow s)]}{\mathbb{E}_\tau[\#\tau(s \rightarrow s') + \#\tau(s' \rightarrow s)]} \geq \frac{\mathbb{E}_\tau[\#\tau(s)(s \rightarrow s') - \#\tau(s)(s' \rightarrow s)] - \mathbb{E}_\tau[\sum_{i=1}^{n(s)} n_i(s')]}{\mathbb{E}_\tau[\#\tau(s)(s \rightarrow s') + \#\tau(s)(s' \rightarrow s)]},$$

$$\xrightarrow{T \rightarrow \infty} 0$$

Reversibility and Reve

$$\psi_{s,r}(s,s') = \frac{\binom{2}{2} E_{r \rightarrow s} \{ \#r(s \rightarrow s') \}}{\binom{2}{2} E_{r \rightarrow s} \{ \#r(s \rightarrow s') + \#r(s' \rightarrow s) \}}$$

$$\psi_{s,r}(s,s') = \frac{E_{r \rightarrow s} \{ \#r(s \rightarrow s') \}}{E_{r \rightarrow s} \{ \#r(s \rightarrow s') + \#r(s' \rightarrow s) \}}.$$

where s appears

Lemma 2. Assume that

$$n_0(s) = \sum_{i=1}^k n_i(s) \quad (5)$$

where $n_i(s)$ denotes the number of times i appears in s . In this case,

$$\#x(s \rightarrow s') = \sum_{i=1}^k x_i(s') - \sum_{i=1}^k x_i(s) = \sum_{i=1}^k (n_i(s') - n_i(s)).$$

If we suppose that $n_i(s) = n_i(s')$, we also have

$$\#x(s \rightarrow s') = \sum_{i=1}^k (n_i(s') - n_i(s)) = \sum_{i=1}^k (n_i(s') - n_i(s)).$$

Thus

$$\#x(s \rightarrow s') = \sum_{i=1}^k (n_i(s') - n_i(s)) = \sum_{i=1}^k (n_i(s') - n_i(s)).$$

If we suppose that $n_1(s') = n_2(s') = \dots = \# \{ \tau(s' \rightarrow s) \} = \sum_{k=1}^{n(s)} \sum_{i=1}^{n_k} n_{k,i}(s') = 0$, then

$$\# \{ \tau(s' \rightarrow s) \} - \# \{ \tau(s' \rightarrow s) \} = \sum_{k=1}^{n(s)} \sum_{i=1}^{n_k} n_{k,i}(s') - \# \{ \tau(s' \rightarrow s) \}.$$

Proof. Eq (4) comes directly from $\# \{ \tau(s' \rightarrow s) \} = \sum_{k=1}^{n(s)} \sum_{i=1}^{n_k} n_{k,i}(s') = 0$. Thus

Eq (5), we first notice that $\# \{ \tau(s' \rightarrow s) \} = 2 \times \# \{ \tau(s' \rightarrow s) \} - \# \{ \tau(s' \rightarrow s) \} =$

$$= 2 \left(k n_0(s') + n_1(s') \sum_{i=1}^{n(s)} \binom{k-1}{i} \right) - (k n_0(s') + k n_0(s') + k(k-1) n_1(s'))$$

$$= k n_0(s') - k n_0(s').$$

As $\# \{ \tau(s' \rightarrow s) \} \rightarrow 0$, $\# \{ \tau(s' \rightarrow s) \}$ converges when T goes to infinity.

As $\# \{ \tau(s' \rightarrow s) \} \rightarrow 0$, we can link reversibility and empirics

$$|s' = s|$$
$$\begin{aligned}
 &= 2 \left(k n_k(s') + n(s') \frac{1}{\binom{n-1}{k-1}} \right) \\
 &= k n_k(s') - k n_0(s').
 \end{aligned}$$

Proof. For a policy π and $s, s' \in S$, let $\phi_\pi(s, a) = E_{\tau \sim P(\cdot|s,a)}[\phi_\pi(s', s)]$. In case (i), we have $\phi_\pi(s, a)$ converges to a quantity denoted by $\psi(s', s)$, and thus

$$\forall s, s' \in S, \quad \frac{\bar{\phi}^\pi(s, s')}{2} \leq \psi(s', s).$$

In cases (ii)-(iv), depending on whether s and s' are recurrent or transient, we can only be linked after s' , thus

We subdivide our problem into four cases, depending on

$$Y_{s'} \in \mathcal{C}_1 \text{ and } \rho_1(P \in \mathcal{T}_{1+1} | s) = 0 \text{ or } 1.$$

Case 1: $P_{s'}(\tau \in \mathcal{T}_{1+1} | s = s) < 1$ and $\rho_1(P \in \mathcal{T}_{1+1} | s) = 0$, as recurrent states can only be linked to a given trajectory. This implies that $\phi_{s'}(s, s') = P_{s'}(\tau \in \mathcal{T}_{1+1} | s) = 0$, as there is no trajectory with s' as recurrent state. Case 2: $P_{s'}(\tau \in \mathcal{T}_{1+1} | s = s) = 1$ and $\rho_1(P \in \mathcal{T}_{1+1} | s) = 0$, as recurrent states can only be linked to a given trajectory. This implies $\phi_{s'}(s, s') = P_{s'}(\tau \in \mathcal{T}_{1+1} | s) = 0$, as there is no trajectory with s' as recurrent state. Case 3: $P_{s'}(\tau \in \mathcal{T}_{1+1} | s = s) = 1$ and $\rho_1(P \in \mathcal{T}_{1+1} | s) = 1$, as recurrent states can only be linked to a given trajectory. This implies $\phi_{s'}(s, s') = P_{s'}(\tau \in \mathcal{T}_{1+1} | s) = 1$, as there is a trajectory with s' as recurrent state. Case 4: $P_{s'}(\tau \in \mathcal{T}_{1+1} | s = s) < 1$ and $\rho_1(P \in \mathcal{T}_{1+1} | s) = 1$, as recurrent states can only be linked to a given trajectory. This implies $\phi_{s'}(s, s') = P_{s'}(\tau \in \mathcal{T}_{1+1} | s) = 1$, as there is a trajectory with s' as recurrent state.

Lemma 1. *Given a trajectory τ , if s appears before s' . We pre-*

in order to simplify $T]$.

$$\frac{\mathbb{E}_x [\#_T(s \rightarrow s') - \#_T(s' \rightarrow s)]}{\mathbb{E}_x [\#_T(s \rightarrow s') + \#_T(s' \rightarrow s)]} \xrightarrow{T \rightarrow \infty} 0, \text{ and finally,}$$

As $\phi_s^*(s, s') = 1$ here, we immediately have $\phi_s^*(s, s') = 1$ for all $s, s' \in S$. This is tight in this case.

[illegible]

This time, we consider a trajectory τ where s appears k times after s' , such that $\tau = \langle s', \dots, s, \dots, s, \dots \rangle$. Then, we have $\psi_{\tau}(s', s) = \frac{E_{\tau}[\psi_{\tau}(s' \rightarrow s)]}{E_{\tau}[\psi_{\tau}(s' \rightarrow s') + \psi_{\tau}(s' \rightarrow s)]}$ and we have $\psi_{\tau}(s' \rightarrow s) \leq \psi_{\tau}(s' \rightarrow s')$. Therefore, $\psi_{\tau}(s', s) \leq \psi_{\tau}(s' \rightarrow s')$, and thus:

[illegible]

Likewise

$$\begin{aligned} E_s[\theta(s' \rightarrow s) + \theta(s \rightarrow s') | k] &= E_s[k n_1(s') u_1(s') + k n_0(s') n_1(s) + k(k-1) n_1(s') n_1(s') | k], \\ &\geq -k E_s[n_1(s') E_s[n_1(s') | k, u_1(s') > 0]] \\ &\quad - k E_s[n_1(s') E_s[n_1(s') | k, u_1(s') > 0]] \text{ as in Lemma 2,} \\ &= k [E_s[n_1(s') E_s[n_1(s') + E_s[n_1(s') E_s[n_1(s')]]] + k(k-1) E_s[n_1(s') E_s[n_1(s')]]], \end{aligned}$$
$$\begin{aligned}
 & \mathbf{1}_{s_i=s} \mathbf{1}_{s_j=s'}, \\
 (2) \quad & \sum_{i < j \leq T} \frac{1}{\binom{T}{2}} \mathbf{1}_{s_i=s} \mathbf{1}_{s_j=s'}, \\
 & = \binom{T}{2} \mathbb{E}_{T \sim \pi} \mathbb{E}_t \mathbb{E}_{t' > t} [\mathbf{1}_{s_i=s} \mathbf{1}_{s_{t'}=s'}]
 \end{aligned}$$

[1, p. 16ff]

Reversi

$$\begin{aligned} & \frac{\mathbb{E}_r[\#\{s' \rightarrow s\} - \#\{s \rightarrow s'\}]}{\mathbb{E}_r[\#\{s' \rightarrow s\} + \#\{s \rightarrow s'\}]} = \frac{\sum_{i=0}^{\infty} \alpha^i p(k=i) \mathbb{E}_r[\#\{s' \rightarrow s\} - \#\{s \rightarrow s'\} | k=i]}{\sum_{i=0}^{\infty} \alpha^i p(k=i) \mathbb{E}_r[\#\{s' \rightarrow s\} + \#\{s \rightarrow s'\} | k=i]}, \\ & \geq \frac{\sum_{i=0}^{\infty} \alpha^{i+1} (1-\alpha) i \mathbb{E}_r[n_1(s)] \mathbb{E}_r[n_1(s')]}{\sum_{i=0}^{\infty} \alpha^{i+1} (1-\alpha) [i \mathbb{E}_r[n_1(s)] \mathbb{E}_r[n_1(s')] + \mathbb{E}_r[n_1(s)] \mathbb{E}_r[n_0(s')] + (i+1) \mathbb{E}_r[n_1(s)] \mathbb{E}_r[n_1(s')]]}, \end{aligned}$$

$$\begin{aligned} & \approx \frac{\sum_{i=1}^n \alpha^{i-1} (1-\alpha) i \mathbb{E}_\tau [\mathbf{u}(s^i)] \mathbb{E}_\tau [\mathbf{u}(s^i)]}{\sum_{i=1}^n \alpha^{i-1} (1-\alpha) i} + i(i-1) \mathbb{E}_\tau [\mathbf{u}(s^i)] \mathbb{E}_\tau [\mathbf{u}(s^i)] \\ & \approx \frac{\sum_{i=1}^n \alpha^{i-1} (1-\alpha) i}{\sum_{i=1}^n \alpha^{i-1} (1-\alpha) i + i(i-1)}, \\ & \approx \frac{\sum_{i=1}^n \alpha^{i-1} (1-\alpha) i}{\sum_{i=1}^n \alpha^{i-1} (1-\alpha) i^2}, \\ & \approx \frac{1}{\frac{1-\alpha}{(1-\alpha)^2}}, \\ & \approx \frac{1-\alpha}{1-\alpha} \end{aligned}$$

From Lemma 1

$$\begin{aligned} \psi^n(s', s) &= \frac{1}{2} \left(1 + \frac{E_{\tau, \sigma} [\#(s' \rightarrow s) - \#(s \rightarrow s')]}{\#(s \rightarrow s') + \#(s' \rightarrow s)} \right) \\ &\geq \frac{1}{2} \left(1 - \frac{1 - \alpha}{1 + \alpha} \right) \\ &\geq \frac{\alpha}{1 + \alpha}, \\ &\geq \frac{\alpha}{2} = \frac{\hat{\phi}_\pi(s, s')}{2}. \end{aligned}$$

As a quick summary, we divided our problem in 4 cases, and proved that in each case, for every pair of states s, s' , we have $\psi^{\pi}(s', s) \geq \frac{\psi(s, s')}{2}$.
To end the proof, we simply take the expectation over the distribution of the next states:

$$\begin{aligned} \mathbb{E}_{s' \sim P(s,a)} \psi_\pi(s', s) &\geq \frac{1}{2} \mathbb{E}_{s' \sim P(s,a)} \hat{\phi}_\pi(s, s'), \\ \hat{\phi}_\pi(s, a) &\geq \frac{\phi_\pi(s, a)}{2}. \end{aligned}$$

We get

$$\psi_{s,T}(s,s') = \overline{\binom{1}{2} E_r}$$

Lemma 2. Assume that we are given

$$s_0 \xrightarrow{\quad} s$$

where $n_1(s')$ denotes the number of s . In this case,

If we suppose that $n_1(s') =$

first notice that $\#_T(s \rightarrow s') = 2 \times \#_T(s \rightarrow s')$

$$= 2 \left(k n_A(s') + n_1(s') \sum_{i=0}^{\infty} \right)$$

$$= k_n(s^t) - k_n(s^t) = 0$$

Proof. For a policy π and $s, s' \in S$, in our case that $\phi_\pi(s, a) = \mathbb{E}_{\pi, P}(a | s, s')$. In our case, $\phi_\pi(s, a)$ converges to a quantity denoted by $\psi(s', s)$, and that:

$$V_{\pi, s'} \in S, \quad \frac{\tilde{v}(s', s)}{2} \leq \psi(s', s).$$

into four cases, depending on whether s and s' are recurrent or transient. (Formally, this means that if a trajectory τ contains s and s' , we only see s a finite

Case 1: $p_s(\tau \in T_{1:n-1} \mid a_{1:n-1} = 1) < 1$ and $p_s(\tau \in T_{1:n-1} \mid a_{1:n-1} = 0) = 0$, as recurrent states can only be linked to recurrent for the Markov chain induced by π_1 . Informally, this means that s is not a recurrent state of π_1 . In this case, we need to consider the state s we need to consider an infinite number of times, and we only need to see a finite number of times.

This implies $p_s(a_{1:n-1}^* \mid \tau \in T_{1:n-1} \mid a_{1:n-1} = 0) = 0$, as recurrent states can only be linked to other recurrent states [34]. It is not possible to find trajectories where s appears after s^* , but $\psi_1(s^*, s) = 0 = \psi(s, s^*)$. Exp. (6) becomes $0 \leq 0$.

17

Therefore,

$$\frac{\mathbb{E}_T[\#T(s \rightarrow s') - \#T(s' \rightarrow s)]}{\mathbb{E}_T[\#T(s \rightarrow s') + \#T(s' \rightarrow s)]} \xrightarrow{T \rightarrow \infty} 0, \text{ and finally,}$$

As $\hat{\phi}_n(s, s') = 1$ here, we immediately get $r(s, s') = \frac{1}{2} + \frac{E_T[\#T(s \rightarrow s') - \#T(s' \rightarrow s)]}{2E_{T \sim \pi}[\#T(s \rightarrow s') - \#T(s' \rightarrow s)]} \xrightarrow{T \rightarrow \infty} 0$, and finally, $r(s, s') = 1$ here, we immediately get $r(s, s') = \frac{1}{2} + \frac{E_T[\#T(s \rightarrow s') - \#T(s' \rightarrow s)]}{2E_{T \sim \pi}[\#T(s \rightarrow s') - \#T(s' \rightarrow s)]} \xrightarrow{T \rightarrow \infty} 0$, and finally,

[illegible]

This time, we consider a trajectory τ where s appears k times after s' , such that $k \geq 1$. Then, we have $\psi_\tau(s', s) = \psi(s' \rightarrow s)$. We can then write τ as $\tau = \tau_1 \dots \tau_k$, where τ_i is a trajectory starting from s' and ending at s . We can then write τ_i as $\tau_i = \tau_i^1 \dots \tau_i^{n_i}$, where τ_i^j is a trajectory starting from s' and ending at s . We can then write τ_i^j as $\tau_i^j = \tau_i^{j,1} \dots \tau_i^{j,n_{i,j}}$, where $\tau_i^{j,k}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k}$ as $\tau_i^{j,k} = \tau_i^{j,k,1} \dots \tau_i^{j,k,n_{i,j,k}}$, where $\tau_i^{j,k,l}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l}$ as $\tau_i^{j,k,l} = \tau_i^{j,k,l,1} \dots \tau_i^{j,k,l,n_{i,j,k,l}}$, where $\tau_i^{j,k,l,m}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m}$ as $\tau_i^{j,k,l,m} = \tau_i^{j,k,l,m,1} \dots \tau_i^{j,k,l,m,n_{i,j,k,l,m}}$, where $\tau_i^{j,k,l,m,n}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n}$ as $\tau_i^{j,k,l,m,n} = \tau_i^{j,k,l,m,n,1} \dots \tau_i^{j,k,l,m,n,n_{i,j,k,l,m,n}}$, where $\tau_i^{j,k,l,m,n,p}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p}$ as $\tau_i^{j,k,l,m,n,p} = \tau_i^{j,k,l,m,n,p,1} \dots \tau_i^{j,k,l,m,n,p,n_{i,j,k,l,m,n,p}}$, where $\tau_i^{j,k,l,m,n,p,q}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q}$ as $\tau_i^{j,k,l,m,n,p,q} = \tau_i^{j,k,l,m,n,p,q,1} \dots \tau_i^{j,k,l,m,n,p,q,n_{i,j,k,l,m,n,p,q}}$, where $\tau_i^{j,k,l,m,n,p,q,r}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r}$ as $\tau_i^{j,k,l,m,n,p,q,r} = \tau_i^{j,k,l,m,n,p,q,r,1} \dots \tau_i^{j,k,l,m,n,p,q,r,n_{i,j,k,l,m,n,p,q,r}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s}$ as $\tau_i^{j,k,l,m,n,p,q,r,s} = \tau_i^{j,k,l,m,n,p,q,r,s,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,n_{i,j,k,l,m,n,p,q,r,s}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t} = \tau_i^{j,k,l,m,n,p,q,r,s,t,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,n_{i,j,k,l,m,n,p,q,r,s,t}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,n_{i,j,k,l,m,n,p,q,r,s,t,u}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,n_{i,j,k,l,m,n,p,q,r,s,t,u,v}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,n_{i,j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i}}$, where $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,j}$ is a trajectory starting from s' and ending at s . We can then write $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,j}$ as $\tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,j} = \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,j,1} \dots \tau_i^{j,k,l,m,n,p,q,r,s,t,u,v,w,x,y,z,a,b,c,d,e,f,g,h,i,j,n_{i,j,k,l,m,n,p$

This time, we consider a trajectory τ where s appears

we consider a trajectory τ where a appears k times at s' , such that it is of the form:

$$s_0 \xrightarrow{\frac{s_0' - s_0^2}{s_0(s_0' - s_0)}} s_1 \xrightarrow{\frac{s_1' - s_1^2}{s_1(s_1' - s_1)}} s_2 \xrightarrow{\frac{s_2' - s_2^2}{s_2(s_2' - s_2)}} \dots \xrightarrow{\frac{s_{n-1}' - s_{n-1}^2}{s_{n-1}(s_{n-1}' - s_{n-1})}} s_n \xrightarrow{\frac{s_n' - s_n^2}{s_n(s_n' - s_n)}} s_{n+1}$$

$$\begin{aligned} & \geq \mathbb{E}_x[\theta(s' \rightarrow s) - \theta(s \rightarrow s')]k, n_k(s') > 0], \\ & \geq \mathbb{E}_x[n_0(s')n_1(s) + (k-1)n_1(s) + n_k(s)] - n_1(s)[kn_1(s') + n_k(s')] \\ & \geq -k\mathbb{E}_x[n_1(s)k] \mathbb{E}_x[n_k(s')k, n_k(s') > 0], \\ & -k\mathbb{E}_x(n_1(s))\mathbb{E}_x(n_k(s')), \end{aligned}$$

as in Lemma 2,

Likewise,

$$\begin{aligned} E_s[\mathcal{E}(s' \rightarrow s) + \mathcal{E}(s \rightarrow s') | k] &= E_s[k n_1(s) n_0(s') + k n_0(s') n_1(s) + k(k-1) n_1(s) n_1(s') | k], \\ &= k[E_s[n_1(s) | k] E_s[n_1(s') + E_s[n_1(s) | k] E_s[n_1(s') | k], \\ &\geq -k E_s[n_1(s) | k] E_s[n_1(s')], \end{aligned}$$

$$\begin{aligned}
 & \mathbb{1}_{x_1 = s} \mathbb{1}_{x_j = s'} \\
 & \left(\frac{2}{T} \right) \sum_{1 \leq j \leq T} \frac{1}{\binom{T}{2}} \mathbb{1}_{x_1 = s} \mathbb{1}_{x_j = s'} \\
 & = \left(\frac{2}{T} \right) \mathbb{E}_{r, r'} \mathbb{E}_{\mathbf{t} \in \mathcal{T}^2} [\mathbb{1}_{x_1 = s} \mathbb{1}_{x_{r'} = s'}],
 \end{aligned}$$

is not possible to see
erified.

From Lemma 1 we get

$$\frac{+ \#T(s \rightarrow s') - \#T(s' \rightarrow s)}{+ \#T(s' \rightarrow s)}$$

$$\frac{(\rightarrow s)]}{(s' \rightarrow s)]} = \frac{-k E_r n_1(s')}{k E_r n_1(s') + k^2 E_r n_2(s')} \rightarrow 0. \quad (7)$$

note $\#_T(s)$ the random variable corresponding to the number of occurrences of s in T . The following structure:

$$\begin{aligned} & \xrightarrow{\tau_1} \dots \xrightarrow{\tau_k} s = ST_{\varphi_T(s)} \xrightarrow{\tau_{k+1}} s_1 \xrightarrow{\tau_{k+2}} s = ST_{\varphi_T(s)} \dots \\ & \leq \frac{E_T[\#\varphi_T(s)(s \rightarrow s') - \#\varphi_T(s)(s' \rightarrow s)] + E_T[\varphi_T(s)_{k+1}(s')]}{E_T[\#\varphi_T(s)(s \rightarrow s') + E_T[\#\varphi_T(s')(s' \rightarrow s)]} \\ & \xrightarrow{\tau_{k+2}} 0 \text{ as in Equ. (7)}. \end{aligned}$$

$$\frac{\#_{T(s' \rightarrow s)}}{\#_{T(s' \rightarrow s)}} \geq \frac{\mathbb{E}_T[\#_{T(s)}(s \rightarrow s') - \#_{T(s)}(s' \rightarrow s)] - \mathbb{E}_T \sum_{i=1}^{\#_{T(s)}+1} n_i(s')}{\mathbb{E}_T[\#_{T(s)}(s \rightarrow s') + \#_{T(s)}(s' \rightarrow s)]},$$

(The Math)

Reversibility and Reversibility Estimation

Reversibility: $\phi_{\pi}(s, a) := p_{\pi}(s \in \tau_{t+1:\infty} \mid s_t = s, a_t = a)$

Empirical reversibility: $\bar{\phi}_{\pi}(s, a) = \mathbb{E}_{s' \sim P(s, a)} [\psi_{\pi}(s', s)]$

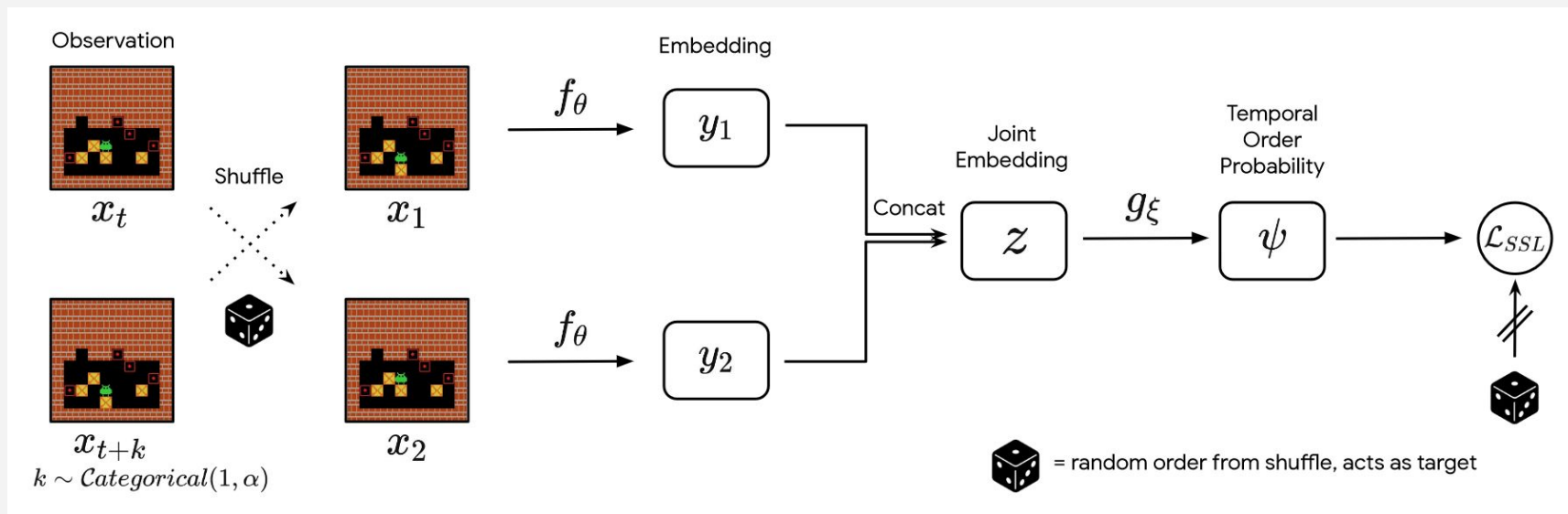
Relation of reversibility and empirical reversibility

$$\bar{\phi}_{\pi}(s, a) \geq \frac{\phi_{\pi}(s, a)}{2}$$



(How to RL)

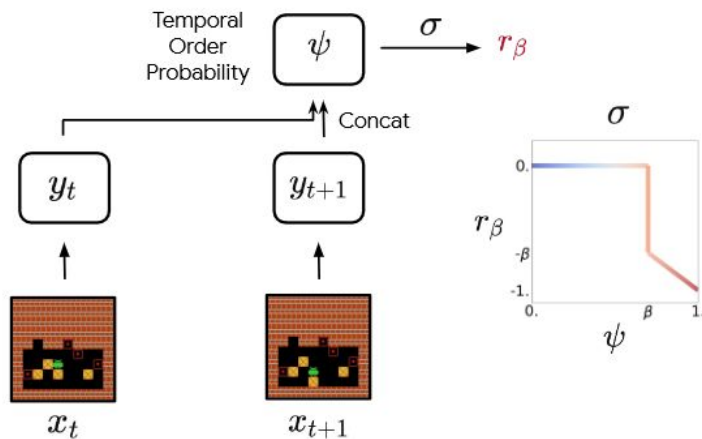
Reversibility-Aware Exploration / Control



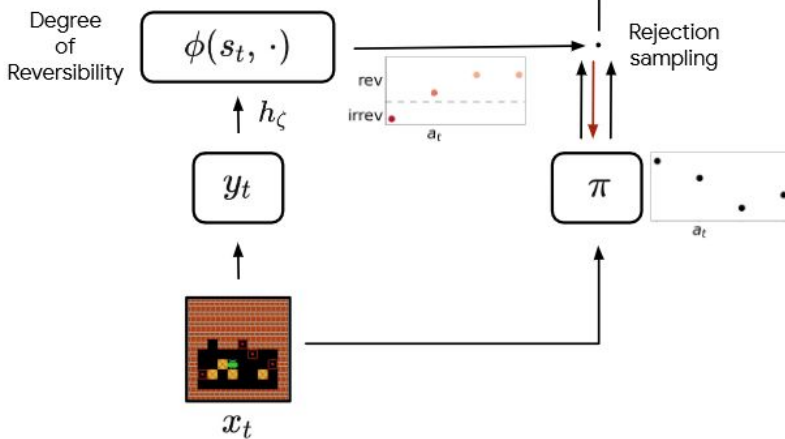
(How to RL)

Reversibility-Aware Exploration / Control

(a) RAE penalizes irreversible transitions

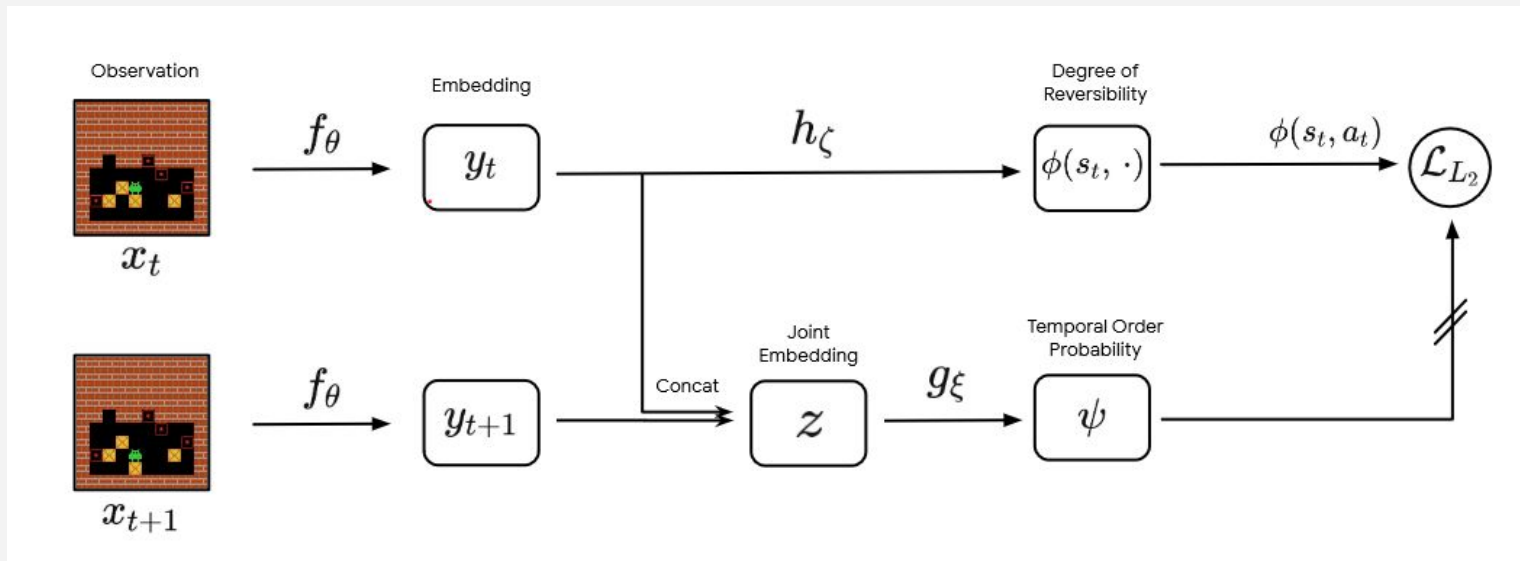


(b) RAC hijacks irreversible actions

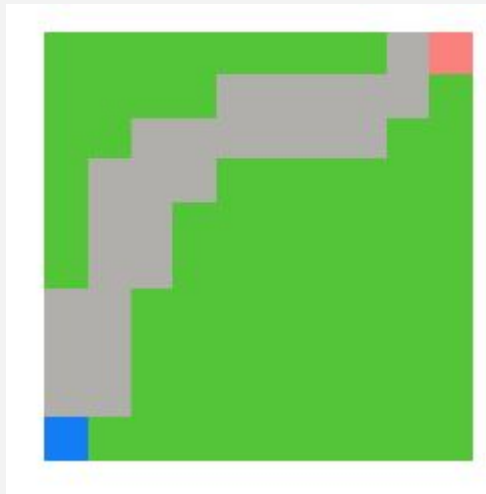


(How to RL)

Reversibility-Aware Exploration / Control

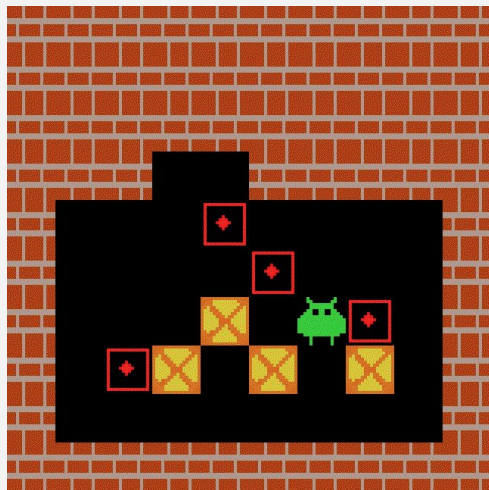


Tests and results



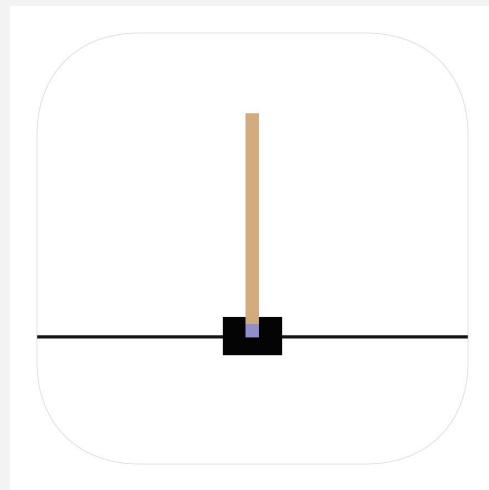
[1, p. 8]

Turf-Environment



[7]

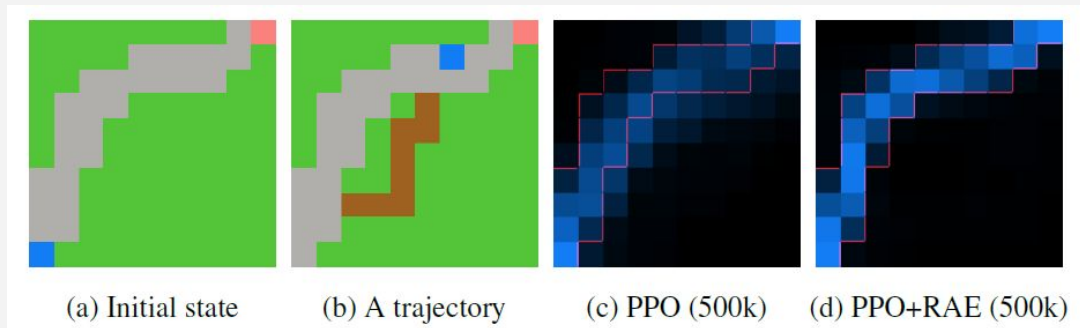
Sokoban



[8]

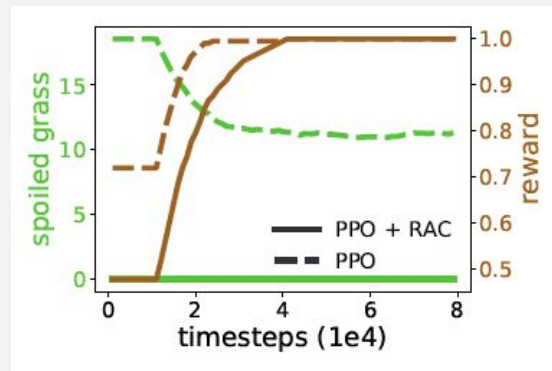
Cartpole

Tests and results – Turf



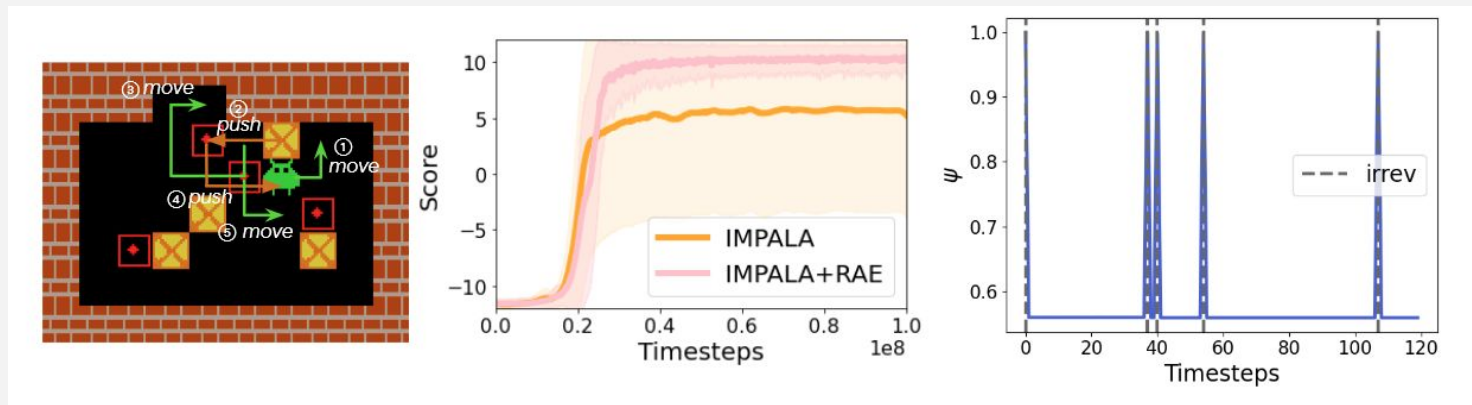
[1, p. 8]

- No irreversible actions
- Slower learning



[1, p. 9]

Tests and results - Sokoban



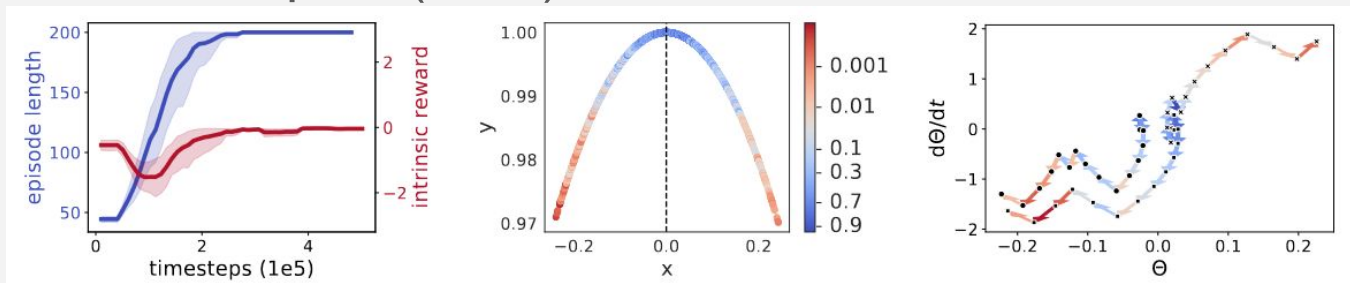
[1, p. 8]

- Very challenging environment
- Sparse irreversible actions
- Better and more consistent performance with RAE

Tests and results - Cartpole

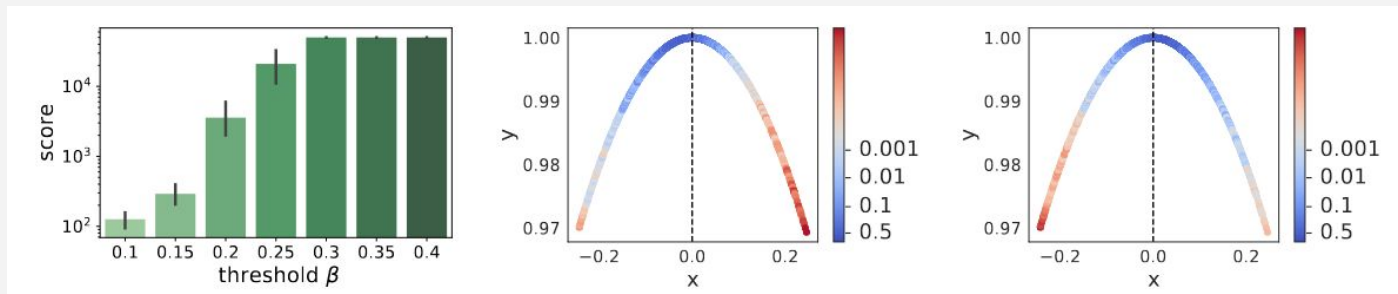
Reward-free cartpole (RAE)

[1, p. 7]



Cartpole+ (RAC)

Note: Color indicates estimated reversibility values



Action 0 (move left)

Action 1 (move right) [1, p. 9]

Summary

- Safety = Reversibility
- Definition of reversibility via precedence
- Precedence classification as surrogate task
- Reversibility-Aware Exploration / Control
- Turf / Sakoban / Cartpole



Conclusion and critique

- Simple representation of a complex task
- Extremely modular, can be used on any policy
- Requires (a lot) more testing



Sources

- Paper:

- [1] – Nathan Grinsztajn and Johan Ferret and Olivier Pietquin and Philippe Preux and Matthieu Geist, “There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning”, *NeurIPS* 2021.

- Images:

- [2]–<https://blog.ml6.eu/catching-the-ai-train-c0c496959999>
- [3]–<https://medium.com/decktopus/15-memes-everyone-who-has-given-a-presentation-will-relate-to-e4946babfc6f>
- [4]–<https://www.pinterest.de/pin/625718941963149650/>
- [5]–<https://www.pinterest.es/pin/691935930226547050/>
- [6]–<http://www.quickmeme.com/meme/3u2bs0>
- [7]–<https://mobile.twitter.com/GoogleAI/status/1455973174319910915?cxt=HHwWWhsCwner407QoAAAA>
- [8]–<https://github.com/ganeshjha/Cartpole>
- [9]–<https://imgflip.com/i/71ua6o>
- [10]–<https://towardsdatascience.com/deep-learning-a-monty-hall-strategy-or-a-gentle-introduction-to-deep-q-learning-and-openai-gym-d66918ac5b26>

Questions?