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

[Nathan et al., *NeurIPS* 2021]

Agenda

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

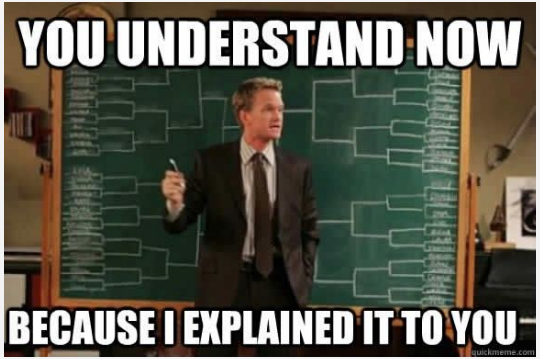
1. Idea
2. Math
3. How to RL
4. Tests and results

[2]

[Nathan et al., *NeurIPS* 2021]

Overview

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



[3]

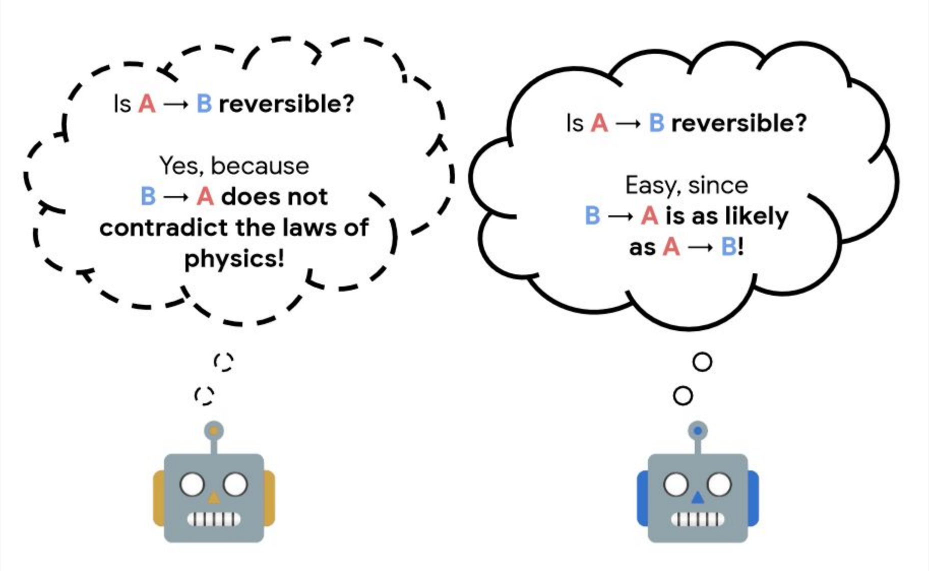
[Nathan et al., *NeurIPS* 2021]

(The Idea)

Approach

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

[1, p. 2]


Is A → B reversible?
Yes, because B → A does not contradict the laws of physics!
Is A → B reversible?
Easy, since B → A is as likely as A → B!

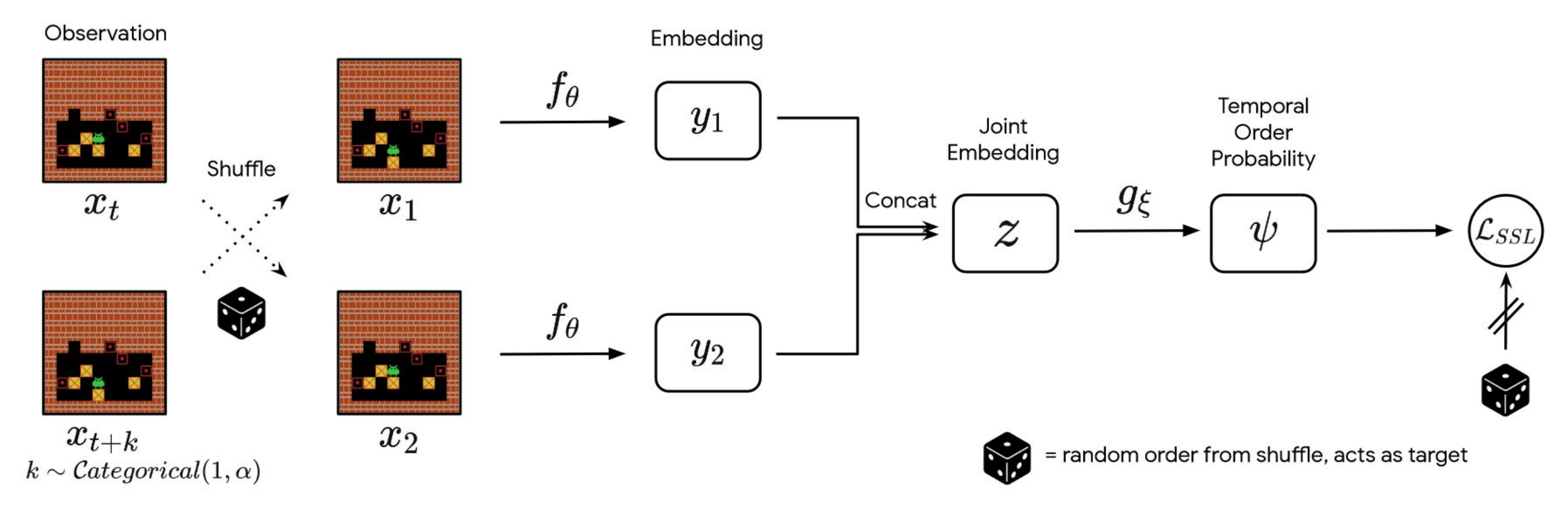
hard

simple(-ish)

[Nathan et al., *NeurIPS* 2021]

(The Idea)

Approach



[1, p. 5]

[Nathan et al., *NeurIPS* 2021]

(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

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

[Nathan et al., *NeurIPS* 2021]

(The Math)

Reversibility and Reversibility Estimation

Degree of reversibility within K steps

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

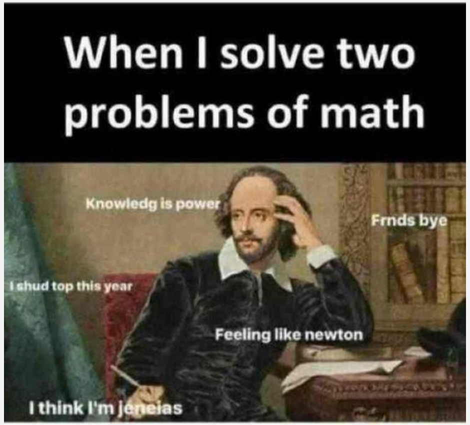
Degree of reversibility

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

[Nathan et al., *NeurIPS* 2021]

(The Math)

Reversibility and Reversibility Estimation



[4]

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

Finite-horizon precedence estimator

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

Emperical reversibility

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

[Nathan et al., *NeurIPS* 2021]

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

Relation of reversibility and empirical reversibility

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

[Nathan et al., *NeurIPS* 2021]

(The Math)

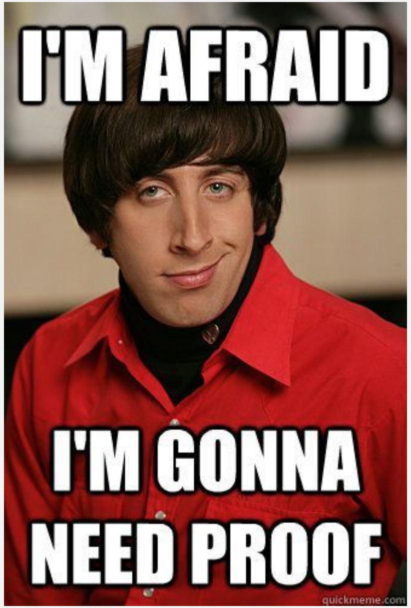
Reversibility and Reversibility Estimation

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

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

Relation of reversibility and empirical reversibility

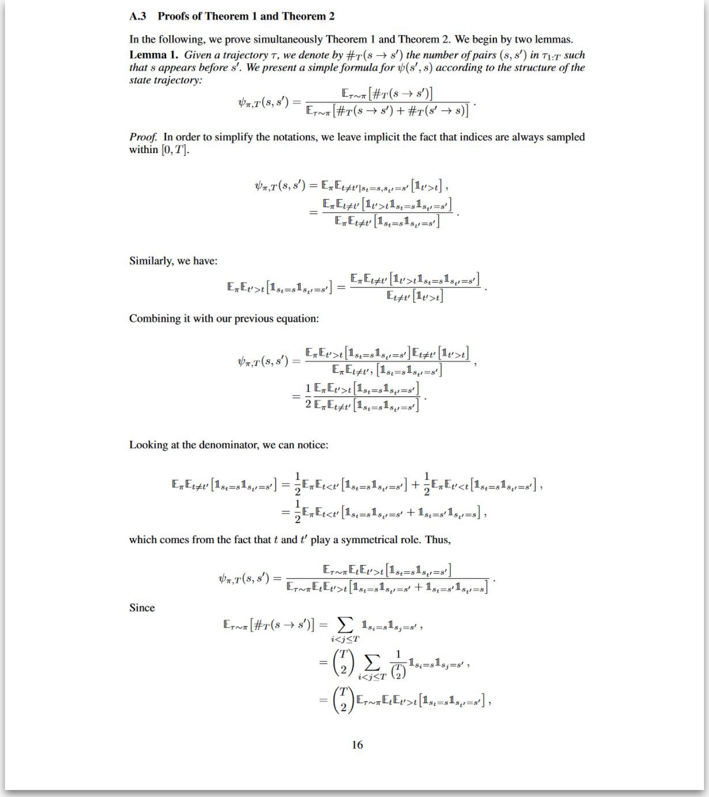
$$\bar{\phi}_\pi(s,a) \geq \frac{\phi_\pi(s,a)}{2}$$



[Nathan et al., *NeurIPS* 2021]

(The Math)

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 $\#_T(s \to s')$ 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:*

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

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

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

Similarly, we have:

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

Combining it with our previous equation:

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

Looking at the denominator, we can notice:

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

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

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

Since

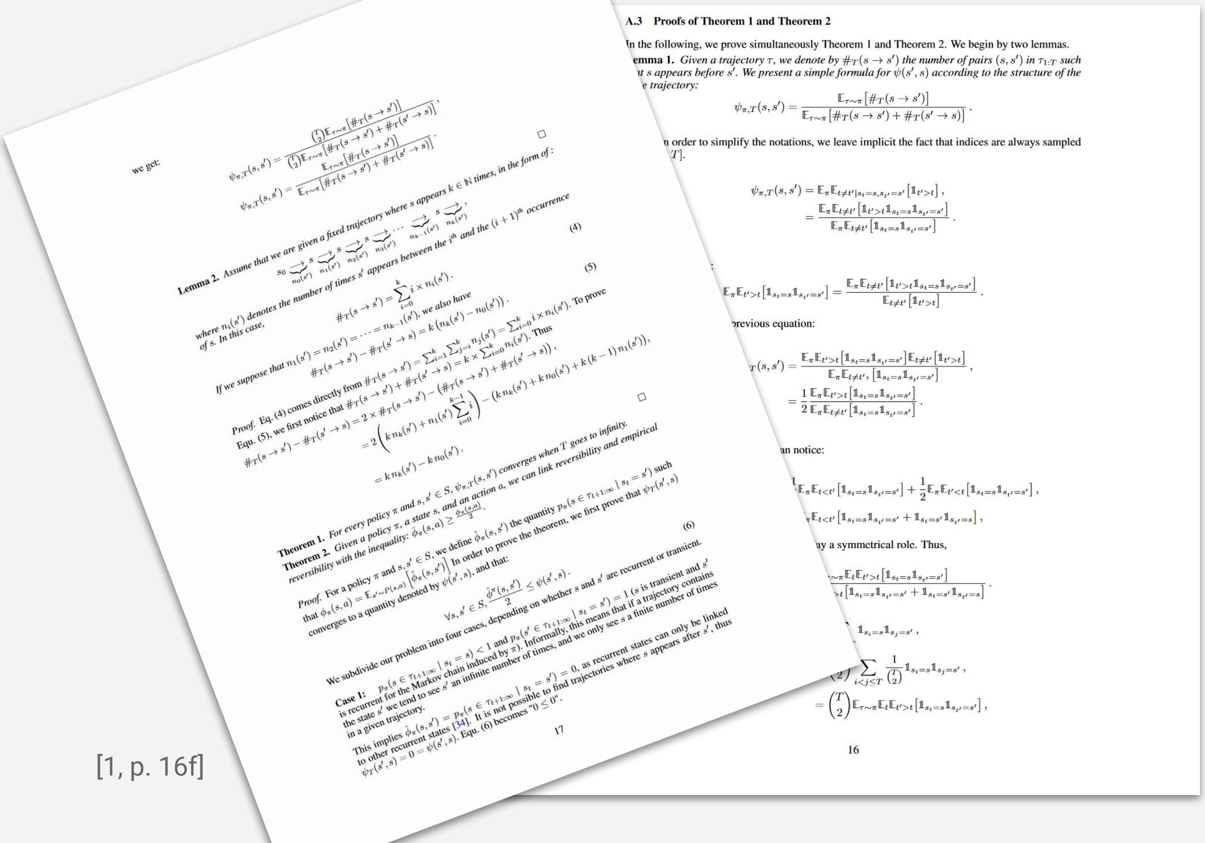
$$\begin{aligned}\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s')\right] &= \sum_{i<j\leq T}\mathbb{1}_{s_i=s}\mathbb{1}_{s_j=s'},\\ &= \binom{T}{2}\sum_{i<j\leq T}\frac{1}{\binom{T}{2}}\mathbb{1}_{s_i=s}\mathbb{1}_{s_j=s'},\\ &= \binom{T}{2}\mathbb{E}_{\tau\sim\pi}\mathbb{E}_t\mathbb{E}_{t'>t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right],\end{aligned}$$

[1, p. 16]

[Nathan et al., *NeurIPS* 2021]

(The Math)

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.

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

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

n order to simplify the notations, we leave implicit the fact that indices are always sampled $T]$.

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

:

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

previous equation:

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

an notice:

$$\mathbb{E}_\pi \mathbb{E}_{t<t'}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right] + \frac{1}{2}\mathbb{E}_\pi \mathbb{E}_{t'<t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right],$$
$$\mathbb{E}_{t<t'}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'} + \mathbb{1}_{s_t=s'}\mathbb{1}_{s_{t'}=s}\right],$$

ay a symmetrical role. Thus,

$$\frac{{}_{\sim\pi}\mathbb{E}_t \mathbb{E}_{t'>t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right]}{{}_{>t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'} + \mathbb{1}_{s_t=s'}\mathbb{1}_{s_{t'}=s}\right]}.$$

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

16

we get:

$$\psi_{\pi,T}(s,s') = \frac{\binom{T}{2}\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s')\right]}{\binom{T}{2}\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]},$$
$$\psi_{\pi,T}(s,s') = \frac{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s')\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]}.$$

□

Lemma 2. *Assume that we are given a fixed trajectory where s appears $k \in \mathbb{N}$ times, in the form of :*

$$s_0 \underbrace{\longrightarrow}_{n_0(s')} s \underbrace{\longrightarrow}_{n_1(s')} s \underbrace{\longrightarrow}_{n_2(s')} s \underbrace{\longrightarrow}_{n_3(s')} \cdots \underbrace{\longrightarrow}_{n_{k-1}(s')} s \underbrace{\longrightarrow}_{n_k(s')}, \tag{4}$$

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

$$\#_T(s \to s') = \sum_{i=0}^{k} i \times n_i(s'). \tag{5}$$

If we suppose that $n_1(s') = n_2(s') = \cdots = n_{k-1}(s')$, we also have

$$\#_T(s \to s') - \#_T(s' \to s) = k\left(n_k(s') - n_0(s')\right).$$

Proof. Eq. (4) comes directly from $\#_T(s \to s') = \sum_{i=1}^{k}\sum_{j=i}^{k} n_j(s') = \sum_{i=0}^{k} i \times n_i(s')$. To prove Equ. (5), we first notice that $\#_T(s \to s') + \#_T(s' \to s) = k \times \sum_{i=0}^{k} n_i(s')$. Thus

$$\#_T(s \to s') - \#_T(s' \to s) = 2 \times \#_T(s \to s') - \left(\#_T(s \to s') + \#_T(s' \to s)\right),$$
$$= 2\left(k\,n_k(s') + n_1(s')\sum_{i=0}^{k-1} i\right) - \left(k\,n_k(s') + k\,n_0(s') + k\,(k-1)\,n_1(s')\right),$$
$$= k\,n_k(s') - k\,n_0(s').$$

□

Theorem 1. *For every policy π and $s, s' \in S$, $\psi_{\pi,T}(s,s')$ converges when T goes to infinity.*

Theorem 2. *Given a policy π, a state s, and an action a, we can link reversibility and empirical reversibility with the inequality: $\bar{\phi}_\pi(s,a) \geq \frac{\phi_\pi(s,a)}{2}$.*

Proof. For a policy π and $s, s' \in S$, we define $\hat{\phi}_\pi(s,s')$ the quantity $p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s')$ such that $\phi_\pi(s,a) = \mathbb{E}_{s'\sim P(s,a)}\left[\hat{\phi}_\pi(s,s')\right]$ In order to prove the theorem, we first prove that $\psi_T(s',s)$ converges to a quantity denoted by $\psi(s',s)$, and that:

$$\forall s, s' \in S, \frac{\hat{\phi}^\pi(s,s')}{2} \leq \psi(s',s). \tag{6}$$

We subdivide our problem into four cases, depending on whether s and s' are recurrent or transient.

Case 1: $p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s) < 1$ and $p_\pi(s' \in \tau_{t+1:\infty} \mid s_t = s') = 1$ (s is transient and s' is recurrent for the Markov chain induced by π). Informally, this means that if a trajectory contains the state s' we tend to see s' an infinite number of times, and we only see s a finite number of times in a given trajectory.

This implies $\hat{\phi}_\pi(s,s') = p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s') = 0$, as recurrent states can only be linked to other recurrent states [34]. It is not possible to find trajectories where s appears after s', thus $\psi_T(s',s) = 0 = \psi(s',s)$. Equ. (6) becomes "$0 \leq 0$".

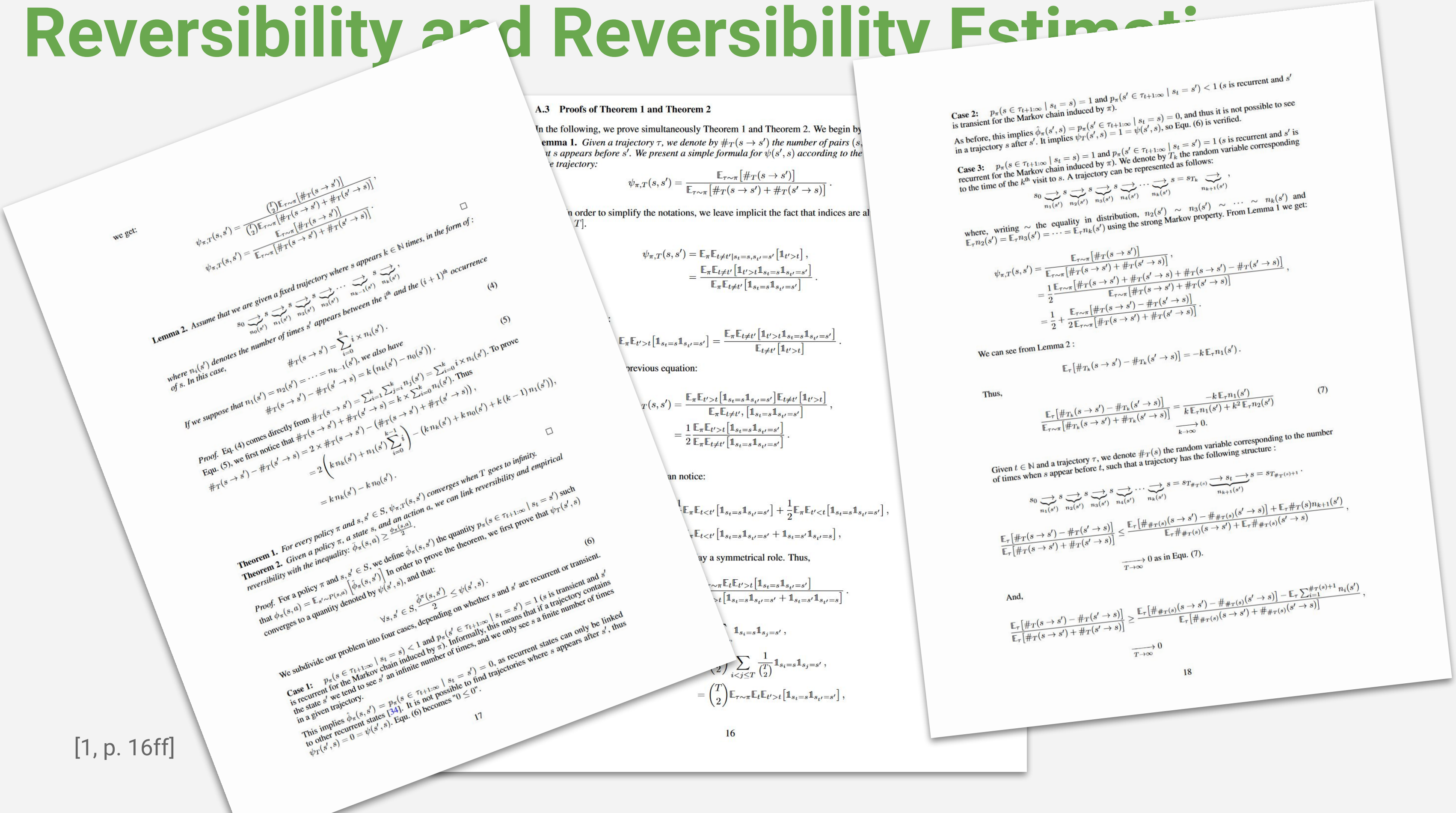
17

[1, p. 16f]

[Nathan et al., *NeurIPS* 2021]

(The Math)

Reversibility and Reversibility Estimati



A.3 Proofs of Theorem 1 and Theorem 2

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

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

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

n order to simplify the notations, we leave implicit the fact that indices are al
$T]$.

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

:

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

previous equation:

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

an notice:

$$\mathbb{E}_\pi \mathbb{E}_{t<t'}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right] + \frac{1}{2}\mathbb{E}_\pi \mathbb{E}_{t'<t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right],$$
$$\mathbb{E}_{t<t'}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'} + \mathbb{1}_{s_t=s'}\mathbb{1}_{s_{t'}=s}\right],$$

ay a symmetrical role. Thus,

$$\frac{\mathbb{E}_{\tau\sim\pi}\mathbb{E}_t\mathbb{E}_{t'>t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'}\right]}{_{>t}\left[\mathbb{1}_{s_t=s}\mathbb{1}_{s_{t'}=s'} + \mathbb{1}_{s_t=s'}\mathbb{1}_{s_{t'}=s}\right]}.$$

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

16

we get:

$$\psi_{\pi,T}(s,s') = \frac{\binom{T}{2}\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s')\right]}{\binom{T}{2}\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]},$$
$$\psi_{\pi,T}(s,s') = \frac{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s')\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]}.$$
□

Lemma 2. *Assume that we are given a fixed trajectory where s appears $k \in \mathbb{N}$ times, in the form of :*

$$s_0 \underbrace{\longrightarrow}_{n_0(s')} s \underbrace{\longrightarrow}_{n_1(s')} s \underbrace{\longrightarrow}_{n_2(s')} s \underbrace{\longrightarrow}_{n_3(s')} \cdots \underbrace{\longrightarrow}_{n_{k-1}(s')} s \underbrace{\longrightarrow}_{n_k(s')},$$

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

$$\#_T(s \to s') = \sum_{i=0}^{k} i \times n_i(s'). \quad (4)$$

If we suppose that $n_1(s') = n_2(s') = \cdots = n_{k-1}(s')$, we also have

$$\#_T(s \to s') - \#_T(s' \to s) = k\left(n_k(s') - n_0(s')\right). \quad (5)$$

Proof. Eq. (4) comes directly from $\#_T(s \to s') = \sum_{i=1}^{k}\sum_{j=i}^{k} n_j(s') = \sum_{i=0}^{k} i \times n_i(s')$. To prove Equ. (5), we first notice that $\#_T(s \to s') + \#_T(s' \to s) = k \times \sum_{i=0}^{k} n_i(s')$. Thus

$$\#_T(s \to s') - \#_T(s' \to s) = 2 \times \#_T(s \to s') - \left(\#_T(s \to s') + \#_T(s' \to s)\right),$$
$$= 2\left(k\,n_k(s') + n_1(s')\sum_{i=0}^{k-1} i\right) - \left(k\,n_k(s') + k\,n_0(s') + k\,(k-1)\,n_1(s')\right),$$
$$= k\,n_k(s') - k\,n_0(s').$$
□

Theorem 1. *For every policy π and $s, s' \in S$, $\psi_{\pi,T}(s,s')$ converges when T goes to infinity.*

Theorem 2. *Given a policy π, a state s, and an action a, we can link reversibility and empirical reversibility with the inequality: $\bar{\phi}_\pi(s,a) \geq \frac{\phi_\pi(s,a)}{2}$.*

Proof. For a policy π and $s, s' \in S$, we define $\hat{\phi}_\pi(s,s')$ the quantity $p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s')$ such that $\phi_\pi(s,a) = \mathbb{E}_{s'\sim P(s,a)}\left[\hat{\phi}_\pi(s,s')\right]$ In order to prove the theorem, we first prove that $\psi_T(s',s)$ converges to a quantity denoted by $\psi(s',s)$, and that:

$$\forall s, s' \in S, \frac{\hat{\phi}^\pi(s,s')}{2} \leq \psi(s',s). \quad (6)$$

We subdivide our problem into four cases, depending on whether s and s' are recurrent or transient.

Case 1: $p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s) < 1$ and $p_\pi(s' \in \tau_{t+1:\infty} \mid s_t = s') = 1$ (s is transient and s' is recurrent for the Markov chain induced by π). Informally, this means that if a trajectory contains the state s' we tend to see s' an infinite number of times, and we only see s a finite number of times in a given trajectory.

This implies $\hat{\phi}_\pi(s,s') = p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s') = 0$, as recurrent states can only be linked to other recurrent states [34]. It is not possible to find trajectories where s appears after s', thus $\psi_T(s',s) = 0 = \psi(s',s)$. Equ. (6) becomes "$0 \leq 0$".

17

Case 2: $p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s) = 1$ and $p_\pi(s' \in \tau_{t+1:\infty} \mid s_t = s') < 1$ (s is recurrent and s' is transient for the Markov chain induced by π).

As before, this implies $\hat{\phi}_\pi(s',s) = p_\pi(s' \in \tau_{t+1:\infty} \mid s_t = s) = 0$, and thus it is not possible to see in a trajectory s after s'. It implies $\psi_T(s',s) = 1 = \psi(s',s)$, so Equ. (6) is verified.

Case 3: $p_\pi(s \in \tau_{t+1:\infty} \mid s_t = s) = 1$ and $p_\pi(s' \in \tau_{t+1:\infty} \mid s_t = s') = 1$ (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:

$$s_0 \underbrace{\longrightarrow}_{n_1(s')} s \underbrace{\longrightarrow}_{n_2(s')} s \underbrace{\longrightarrow}_{n_3(s')} s \underbrace{\longrightarrow}_{n_4(s')} \cdots \underbrace{\longrightarrow}_{n_k(s')} s = s_{T_k} \underbrace{\longrightarrow}_{n_{k+1}(s')},$$

where, writing $\sim$ the equality in distribution, $n_2(s') \sim n_3(s') \sim \cdots \sim n_k(s')$ and $\mathbb{E}_\tau n_2(s') = \mathbb{E}_\tau n_3(s') = \cdots = \mathbb{E}_\tau n_k(s')$ using the strong Markov property. From Lemma 1 we get:

$$\psi_{\pi,T}(s,s') = \frac{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s')\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]},$$
$$= \frac{1}{2}\frac{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s) + \#_T(s \to s') - \#_T(s' \to s)\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]},$$
$$= \frac{1}{2} + \frac{\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') - \#_T(s' \to s)\right]}{2\,\mathbb{E}_{\tau\sim\pi}\left[\#_T(s \to s') + \#_T(s' \to s)\right]}.$$

We can see from Lemma 2 :

$$\mathbb{E}_\tau\left[\#_{T_k}(s \to s') - \#_{T_k}(s' \to s)\right] = -k\,\mathbb{E}_\tau n_1(s').$$

Thus,

$$\frac{\mathbb{E}_\tau\left[\#_{T_k}(s \to s') - \#_{T_k}(s' \to s)\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#_{T_k}(s \to s') + \#_{T_k}(s' \to s)\right]} = \frac{-k\,\mathbb{E}_\tau n_1(s')}{k\,\mathbb{E}_\tau n_1(s') + k^2\,\mathbb{E}_\tau n_2(s')} \xrightarrow[k\to\infty]{} 0. \quad (7)$$

Given $t \in \mathbb{N}$ and a trajectory τ, we denote $\#_T(s)$ the random variable corresponding to the number of times when s appear before t, such that a trajectory has the following structure :

$$s_0 \underbrace{\longrightarrow}_{n_1(s')} s \underbrace{\longrightarrow}_{n_2(s')} s \underbrace{\longrightarrow}_{n_3(s')} s \underbrace{\longrightarrow}_{n_4(s')} \cdots \underbrace{\longrightarrow}_{n_k(s')} s = s_{T_{\#_T(s)}} \underbrace{\longrightarrow s_t \longrightarrow}_{n_{k+1}(s')} s = s_{T_{\#_T(s)+1}}.$$

$$\frac{\mathbb{E}_\tau\left[\#_T(s \to s') - \#_T(s' \to s)\right]}{\mathbb{E}_\tau\left[\#_T(s \to s') + \#_T(s' \to s)\right]} \leq \frac{\mathbb{E}_\tau\left[\#_{\#_T(s)}(s \to s') - \#_{\#_T(s)}(s' \to s)\right] + \mathbb{E}_\tau \#_T(s) n_{k+1}(s')}{\mathbb{E}_\tau \#_{\#_T(s)}(s \to s') + \mathbb{E}_\tau \#_{\#_T(s)}(s' \to s)},$$
$$\xrightarrow[T\to\infty]{} 0 \text{ as in Equ. (7).}$$

And,

$$\frac{\mathbb{E}_\tau\left[\#_T(s \to s') - \#_T(s' \to s)\right]}{\mathbb{E}_\tau\left[\#_T(s \to s') + \#_T(s' \to s)\right]} \geq \frac{\mathbb{E}_\tau\left[\#_{\#_T(s)}(s \to s') - \#_{\#_T(s)}(s' \to s)\right] - \mathbb{E}_\tau \sum_{i=1}^{\#_T(s)+1} n_i(s')}{\mathbb{E}_\tau\left[\#_{\#_T(s)}(s \to s') + \#_{\#_T(s)}(s' \to s)\right]},$$
$$\xrightarrow[T\to\infty]{} 0$$

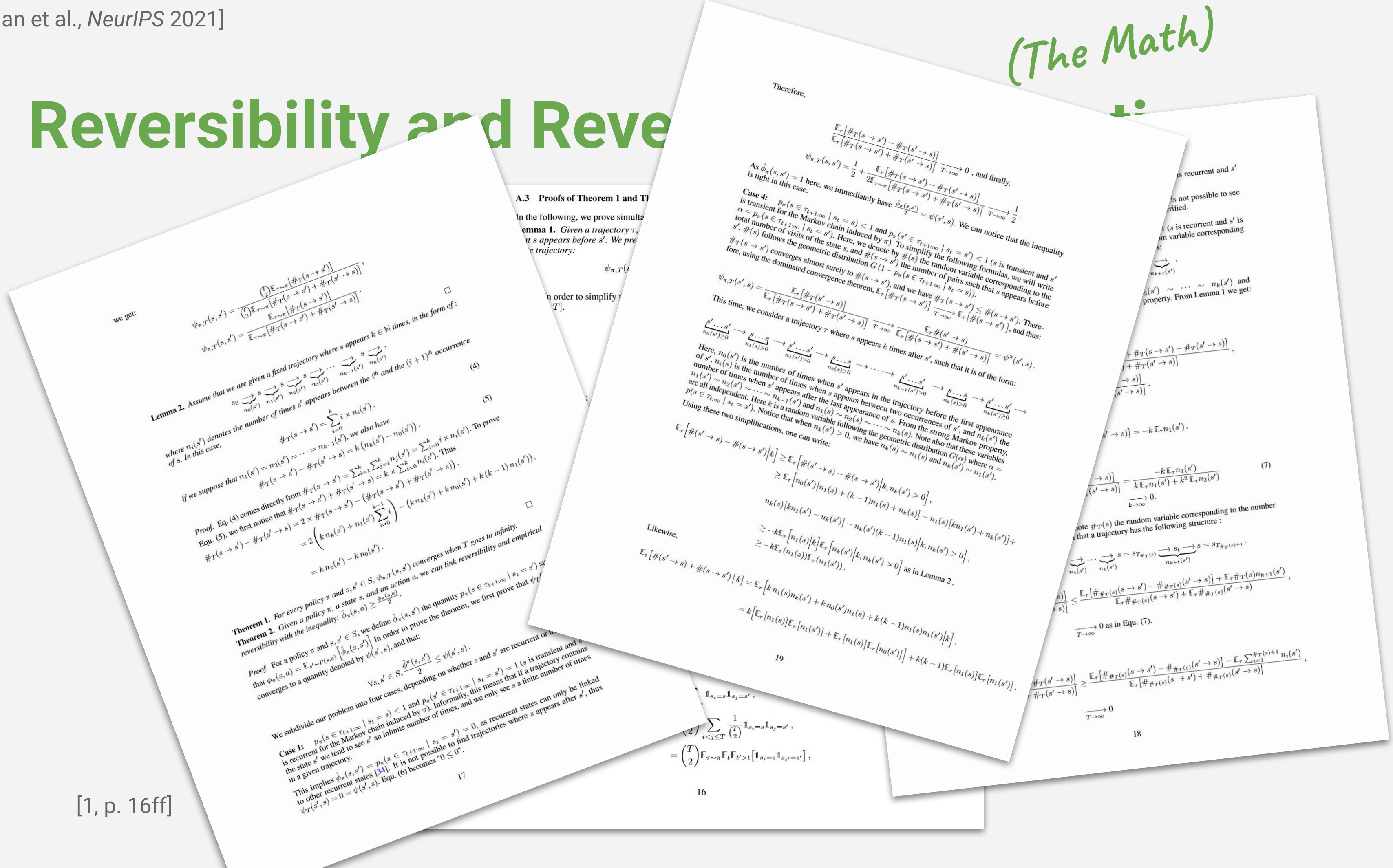
18

[1, p. 16ff]

[Nathan et al., *NeurIPS* 2021]

(The Math)

Reversibility and Reve ti

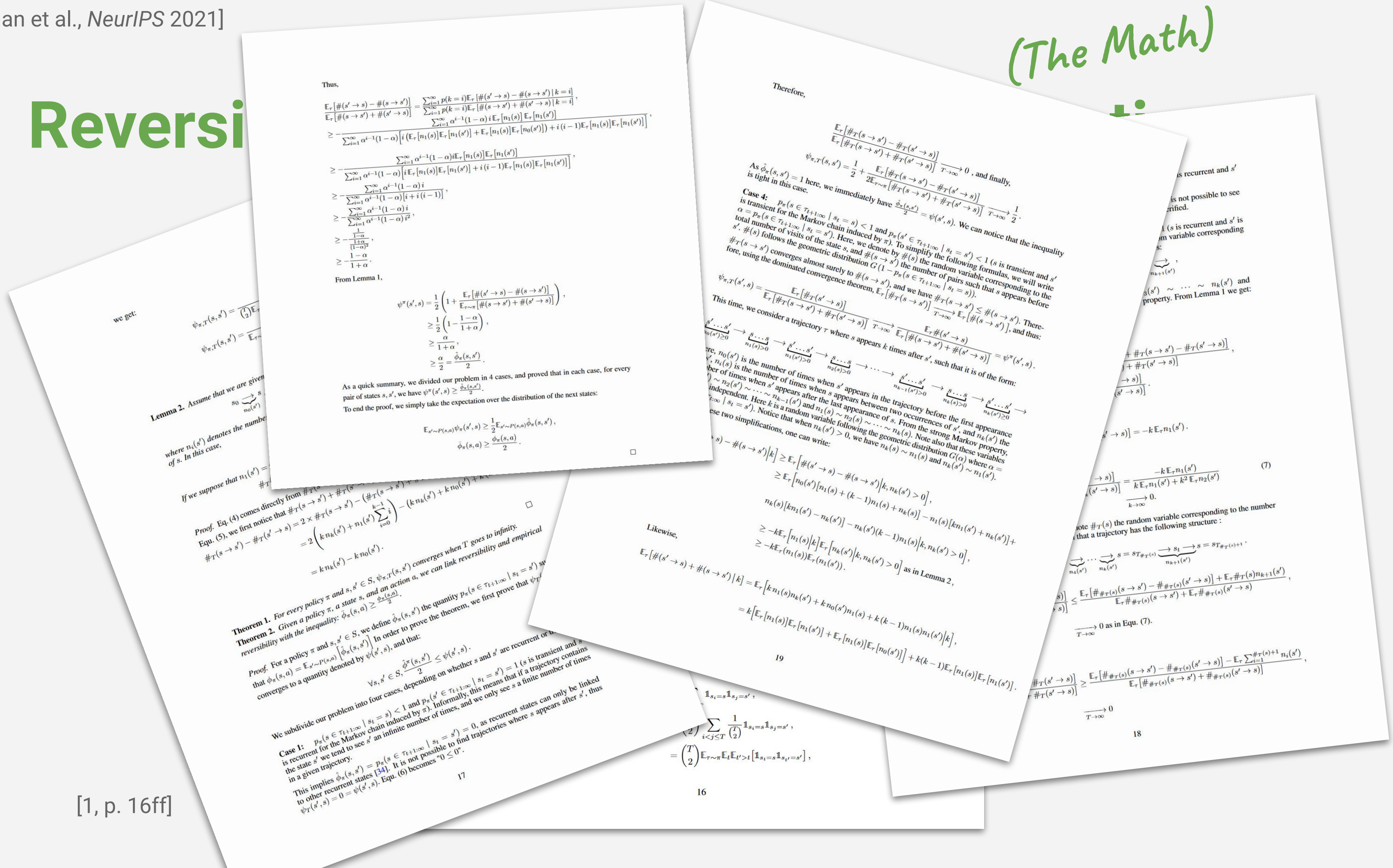


[1, p. 16ff]

[Nathan et al., *NeurIPS* 2021]

Reversi

(The Math)



Thus,

$$\frac{\mathbb{E}_\tau\left[\#(s'\to s)-\#(s\to s')\right]}{\mathbb{E}_\tau\left[\#(s\to s')+\#(s'\to s)\right]}=\frac{\sum_{i=1}^{\infty}p(k=i)\mathbb{E}_\tau\left[\#(s'\to s)-\#(s\to s')\,|\,k=i\right]}{\sum_{i=1}^{\infty}p(k=i)\mathbb{E}_\tau\left[\#(s\to s')+\#(s'\to s)\,|\,k=i\right]},$$

$$\geq-\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i\,\mathbb{E}_\tau[n_1(s)]\,\mathbb{E}_\tau[n_1(s')]}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\Big[i\left(\mathbb{E}_\tau[n_1(s)]\mathbb{E}_\tau[n_1(s')]+\mathbb{E}_\tau[n_1(s)]\mathbb{E}_\tau[n_0(s')]\right)+i\,(i-1)\mathbb{E}_\tau[n_1(s)]\mathbb{E}_\tau[n_1(s')]\Big]},$$

$$\geq-\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)i\mathbb{E}_\tau[n_1(s)]\mathbb{E}_\tau[n_1(s')]}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\Big[i\,\mathbb{E}_\tau[n_1(s)]\mathbb{E}_\tau[n_1(s')]+i\,(i-1)\mathbb{E}_\tau[n_1(s)]\mathbb{E}_\tau[n_1(s')]\Big]},$$

$$\geq-\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)[i+i\,(i-1)]},$$

$$\geq-\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i^2},$$

$$\geq-\frac{\frac{1}{1-\alpha}}{\frac{1+\alpha}{(1-\alpha)^2}},$$

$$\geq-\frac{1-\alpha}{1+\alpha}.$$

From Lemma 1,

$$\psi^\pi(s',s)=\frac{1}{2}\left(1+\frac{\mathbb{E}_\tau\left[\#(s'\to s)-\#(s\to s')\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#(s\to s')+\#(s'\to s)\right]}\right),$$

$$\geq\frac{1}{2}\left(1-\frac{1-\alpha}{1+\alpha}\right),$$

$$\geq\frac{\alpha}{1+\alpha},$$

$$\geq\frac{\alpha}{2}=\frac{\hat{\phi}_\pi(s,s')}{2}.$$

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 $\psi^\pi(s',s)\geq\frac{\hat{\phi}_\pi(s,s')}{2}$.

To end the proof, we simply take the expectation over the distribution of the next states:

$$\mathbb{E}_{s'\sim P(s,a)}\psi_\pi(s',s)\geq\frac{1}{2}\mathbb{E}_{s'\sim P(s,a)}\hat{\phi}_\pi(s,s'),$$

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

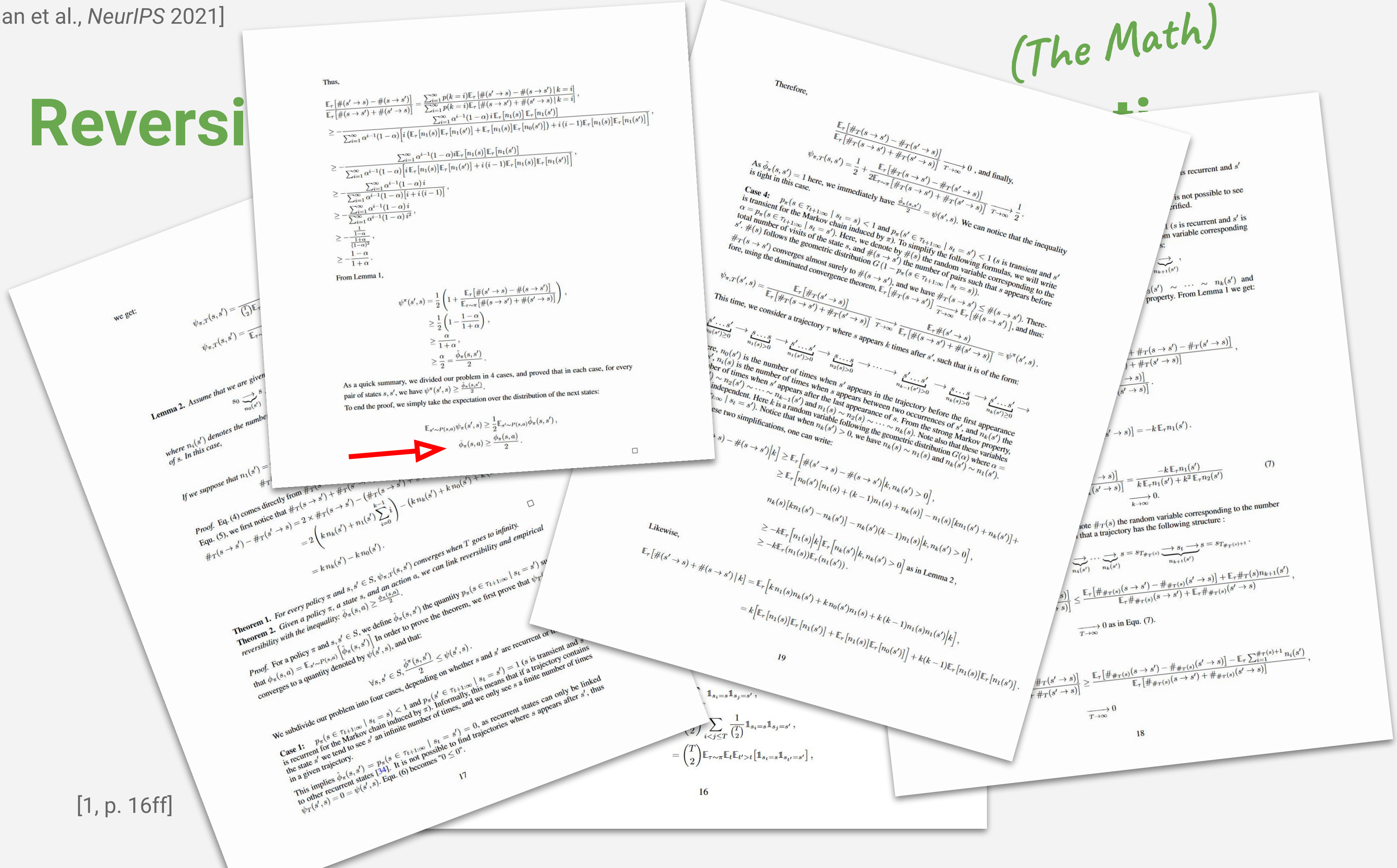
□

[1, p. 16ff]

[Nathan et al., *NeurIPS* 2021]

Reversi

(The Math)



Thus,

$$\frac{\mathbb{E}_\tau\left[\#(s'\to s)-\#(s\to s')\right]}{\mathbb{E}_\tau\left[\#(s\to s')+\#(s'\to s)\right]}=\frac{\sum_{i=1}^{\infty}p(k=i)\mathbb{E}_\tau\left[\#(s'\to s)-\#(s\to s')\,|\,k=i\right]}{\sum_{i=1}^{\infty}p(k=i)\mathbb{E}_\tau\left[\#(s\to s')+\#(s'\to s)\,|\,k=i\right]},$$

$$\geq -\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i\,\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_1(s')\right]}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\left[i\left(\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_1(s')\right]+\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_0(s')\right]\right)+i\,(i-1)\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_1(s')\right]\right]},$$

$$\geq -\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)i\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_1(s')\right]}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\left[i\,\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_1(s')\right]+i\,(i-1)\mathbb{E}_\tau\left[n_1(s)\right]\mathbb{E}_\tau\left[n_1(s')\right]\right]},$$

$$\geq -\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\left[i+i\,(i-1)\right]},$$

$$\geq -\frac{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i}{\sum_{i=1}^{\infty}\alpha^{i-1}(1-\alpha)\,i^2},$$

$$\geq -\frac{\frac{1}{1-\alpha}}{\frac{1+\alpha}{(1-\alpha)^2}},$$

$$\geq -\frac{1-\alpha}{1+\alpha}.$$

From Lemma 1,

$$\psi^\pi(s',s)=\frac{1}{2}\left(1+\frac{\mathbb{E}_\tau\left[\#(s'\to s)-\#(s\to s')\right]}{\mathbb{E}_{\tau\sim\pi}\left[\#(s\to s')+\#(s'\to s)\right]}\right),$$

$$\geq\frac{1}{2}\left(1-\frac{1-\alpha}{1+\alpha}\right),$$

$$\geq\frac{\alpha}{1+\alpha},$$

$$\geq\frac{\alpha}{2}=\frac{\hat{\phi}_\pi(s,s')}{2}.$$

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 $\psi^\pi(s',s)\geq\frac{\hat{\phi}_\pi(s,s')}{2}$.

To end the proof, we simply take the expectation over the distribution of the next states:

$$\mathbb{E}_{s'\sim P(s,a)}\psi_\pi(s',s)\geq\frac{1}{2}\mathbb{E}_{s'\sim P(s,a)}\hat{\phi}_\pi(s,s'),$$

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

□

[1, p. 16ff]

[Nathan et al., *NeurIPS* 2021]

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

Relation of reversibility and empirical reversibility

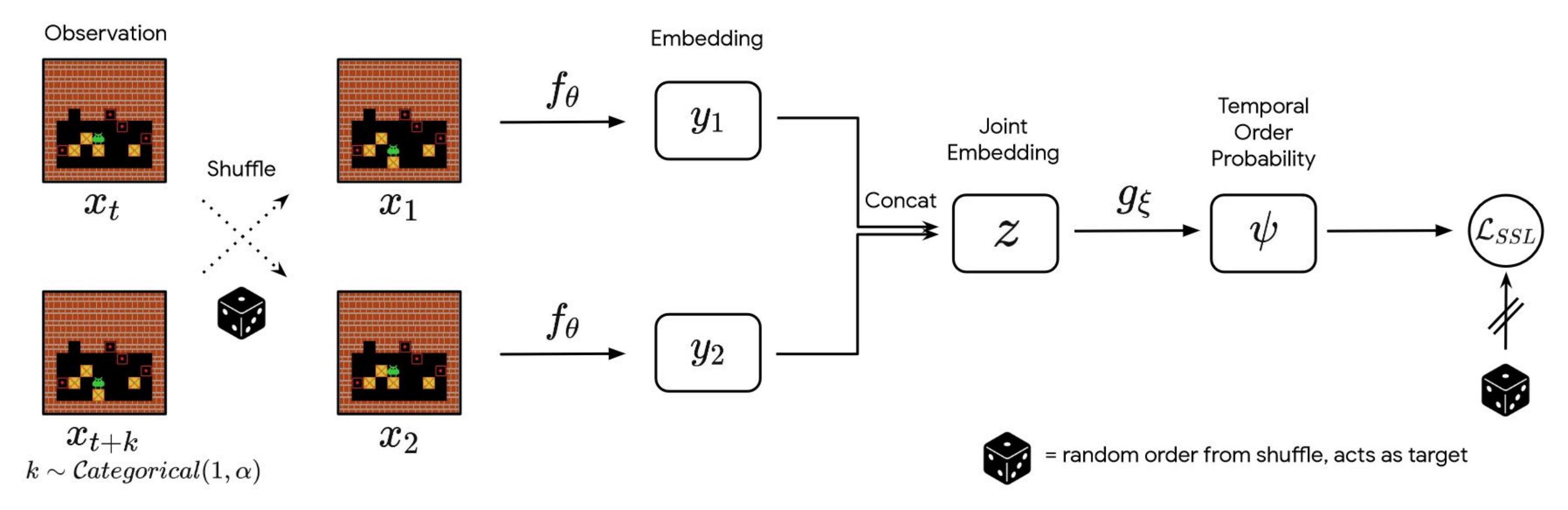
$$\overline{\phi}_\pi(s,a) \geq \frac{\phi_\pi(s,a)}{2}$$

[6]

[Nathan et al., *NeurIPS* 2021]

(How to RL)

Reversibility-Aware Exploration / Control

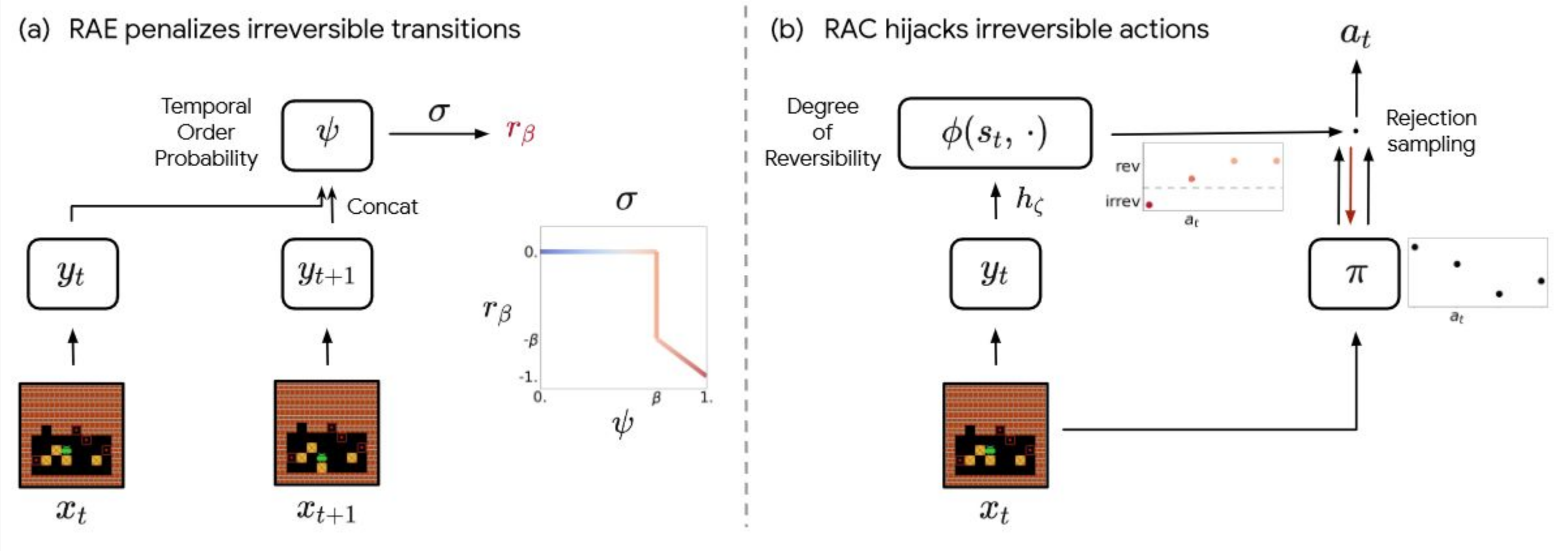


[1, p. 5]

[Nathan et al., *NeurIPS* 2021]

(How to RL)

Reversibility-Aware Exploration / Control

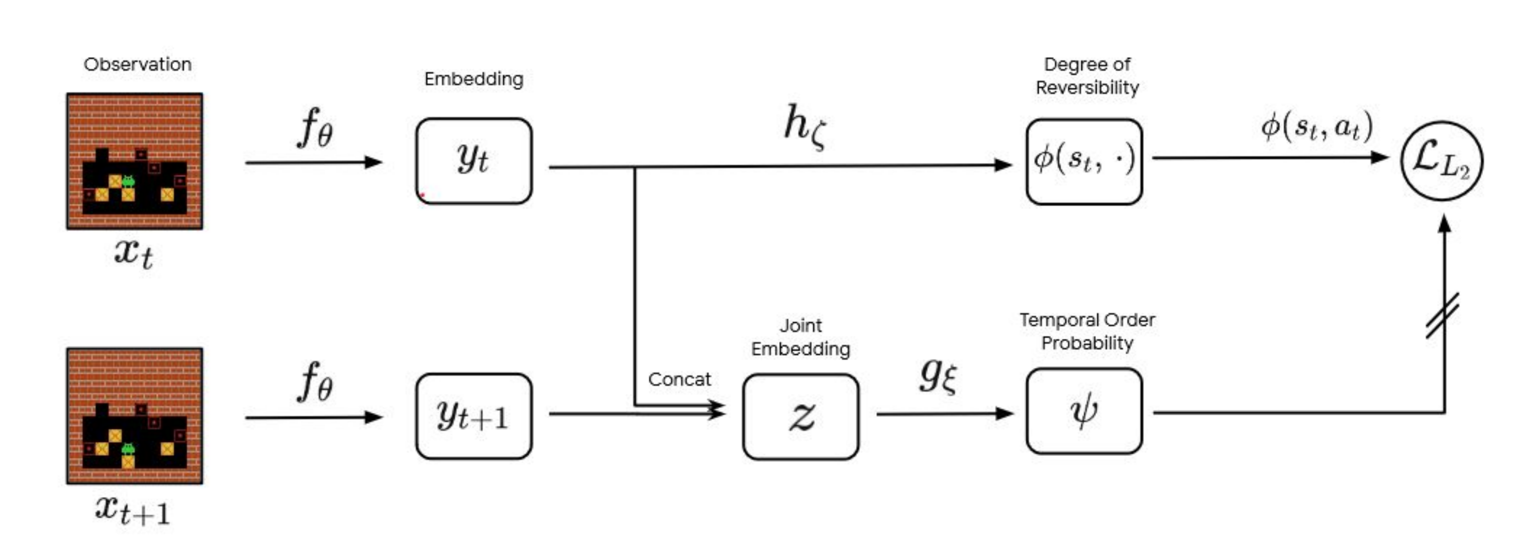


[1, p. 6]

[Nathan et al., *NeurIPS* 2021]

(How to RL)

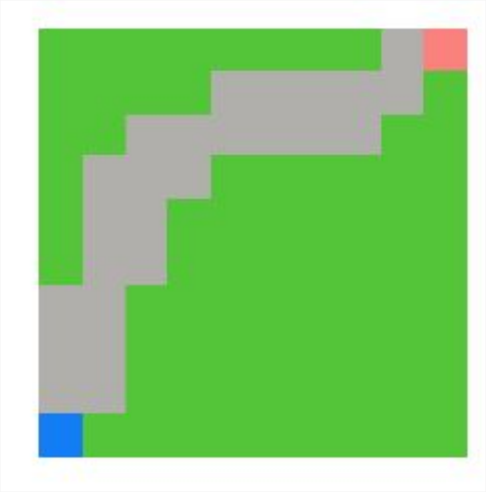
Reversibility-Aware Exploration / Control



[1, p. 5]

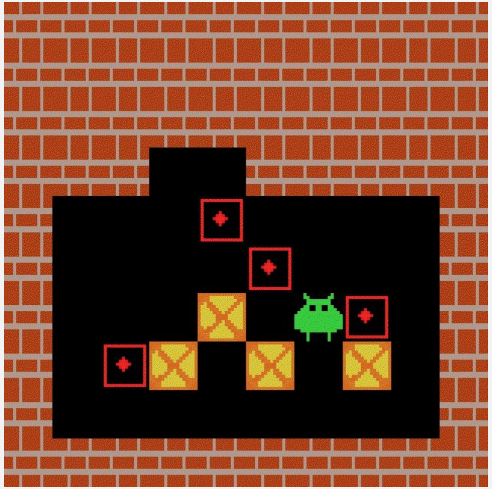
[Nathan et al., *NeurIPS* 2021]

Tests and results



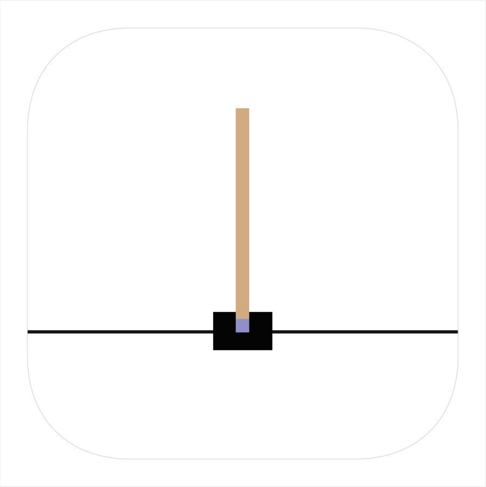
[1, p. 8]

Turf-Environment



[7]

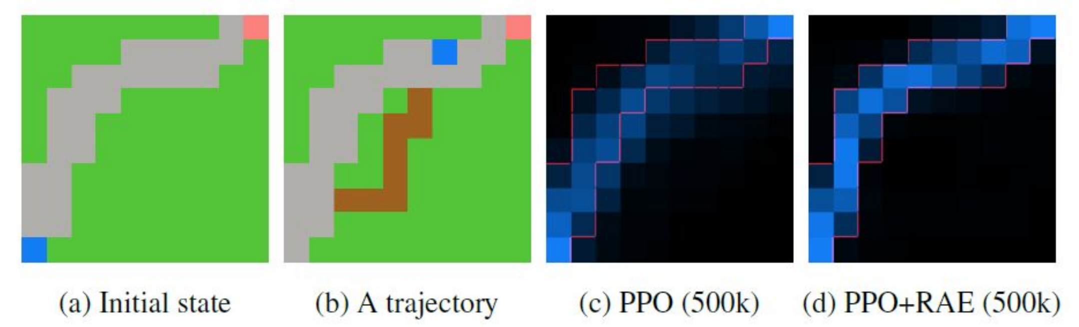
Sokoban



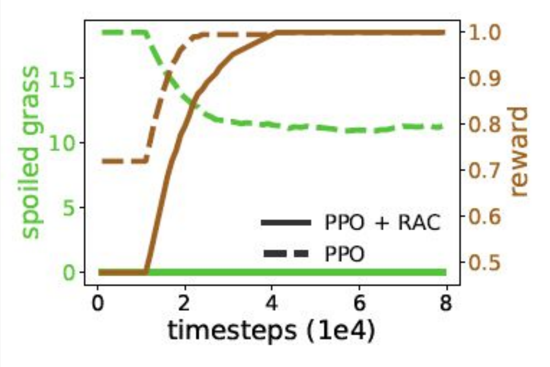
[8]

Cartpole

Tests and results – Turf



[1, p. 8]

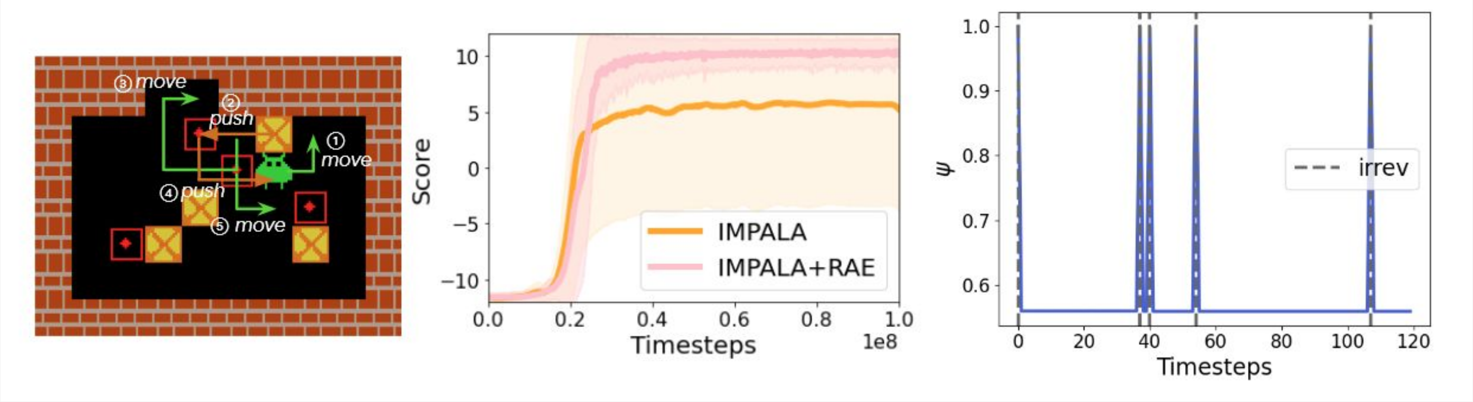


[1, p. 9]

- No irreversible actions
- Slower learning

[Nathan et al., *NeurIPS* 2021]

Tests and results - Sokoban



[1, p. 8]

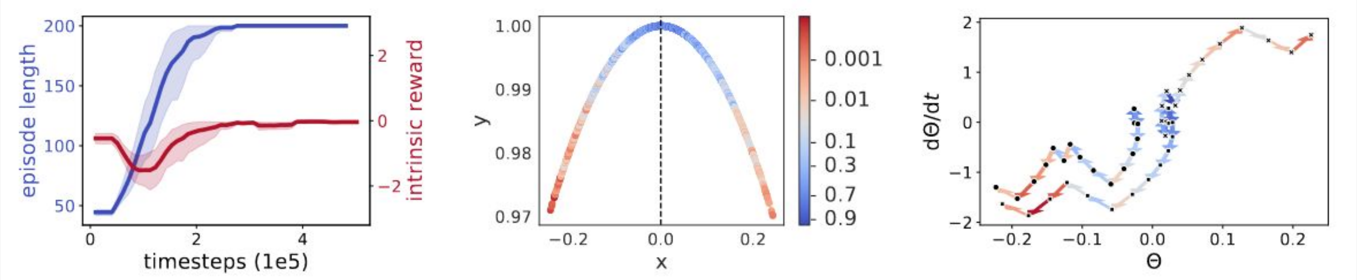
- Very challenging environment
- Sparse irreversible actions
- Better and more consistent performance with RAE

[Nathan et al., *NeurIPS* 2021]

Tests and results - Cartpole

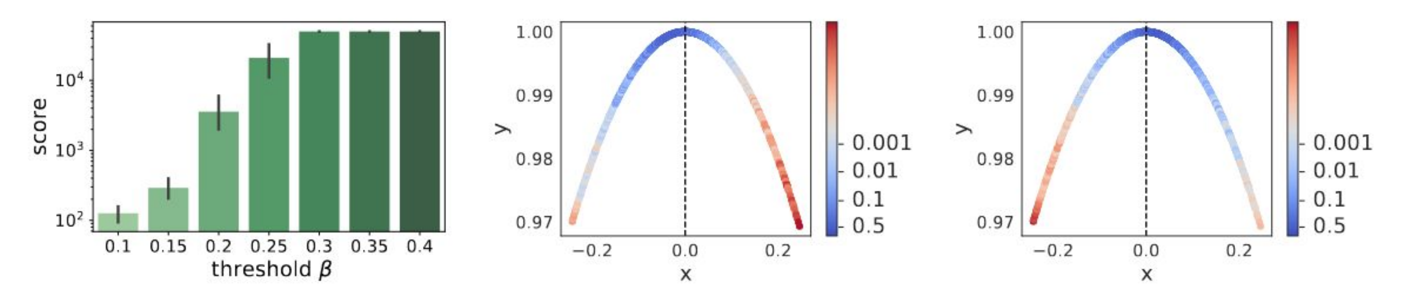
Reward-free cartpole (RAE)

[1, p. 7]



Cartpole+ (RAC)

Note: Color indicates estimated reversibility values



[1, p. 9]

Action 0 (move left)

Action 1 (move right)

[Nathan et al., *NeurIPS* 2021]

Summary

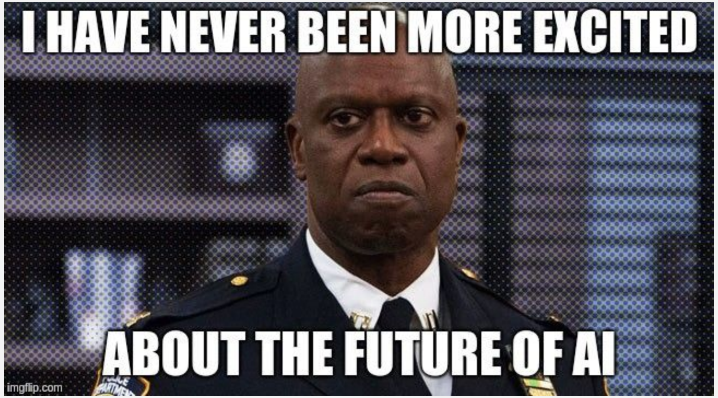
- Safety = Reversibility
- Definition of reversibility via precedence
- Precedence classification as surrogate task
- Reversibility-Aware Exploration / Control
- Turf / Sakoban / Cartpole

[9]

[Nathan et al., *NeurIPS* 2021]

Conclusion and critique

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



[10]

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=HHwWhsCwner407QoAAAA
  - [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?