A Tighter Convergence Proof of Reverse Experience Replay

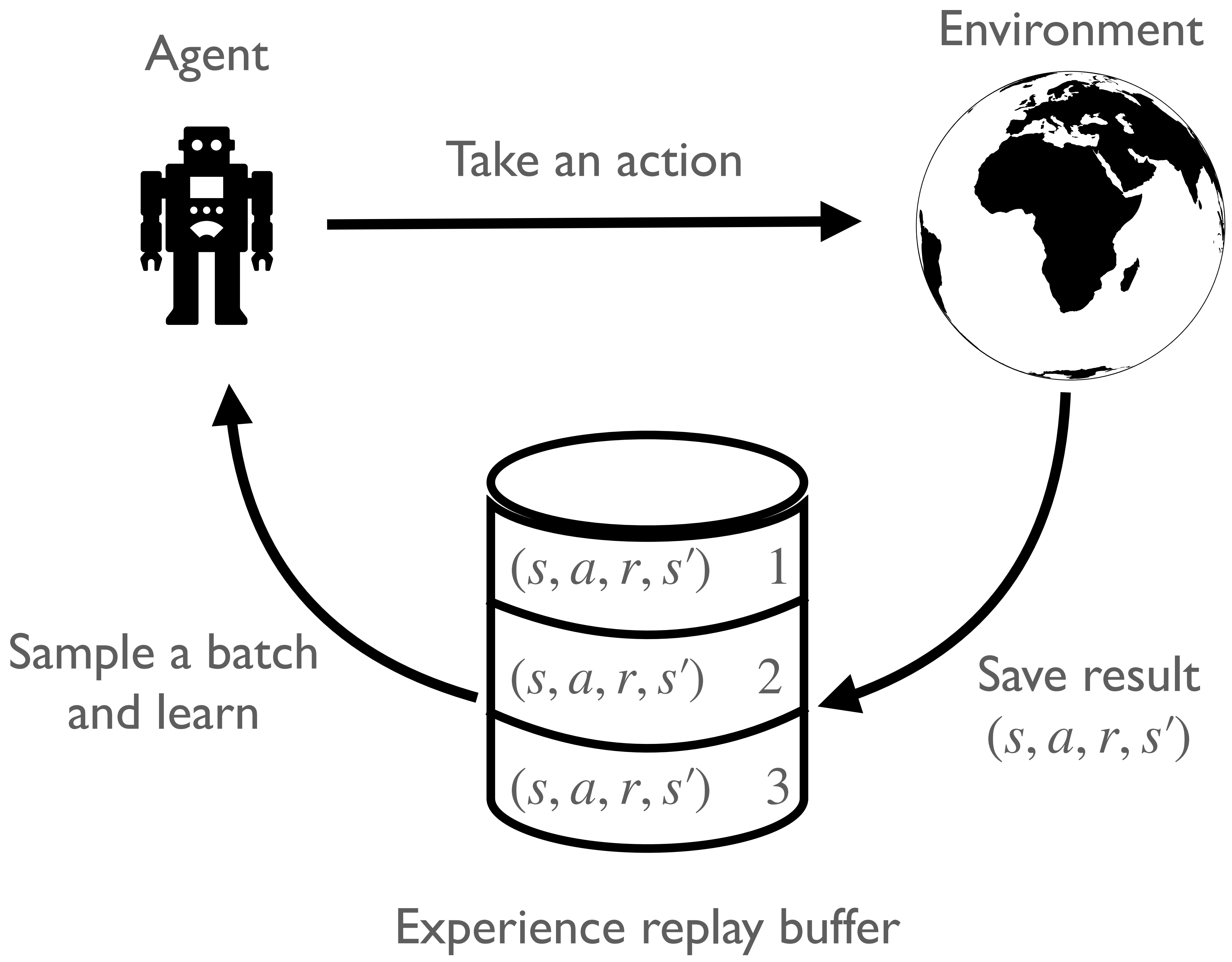
Nan Jiang, Jinzhao Li, Yexiang Xue
Department of Computer Science, Purdue University

Experience Replay

- **Experience Replay (ER)**
  - An agent stores past experiences and randomly samples (replays) transitions during the Q-learning process.
  - Many variants have been proposed.

    - Prioritized Experience Replay

    - Hindsight Experience Replay



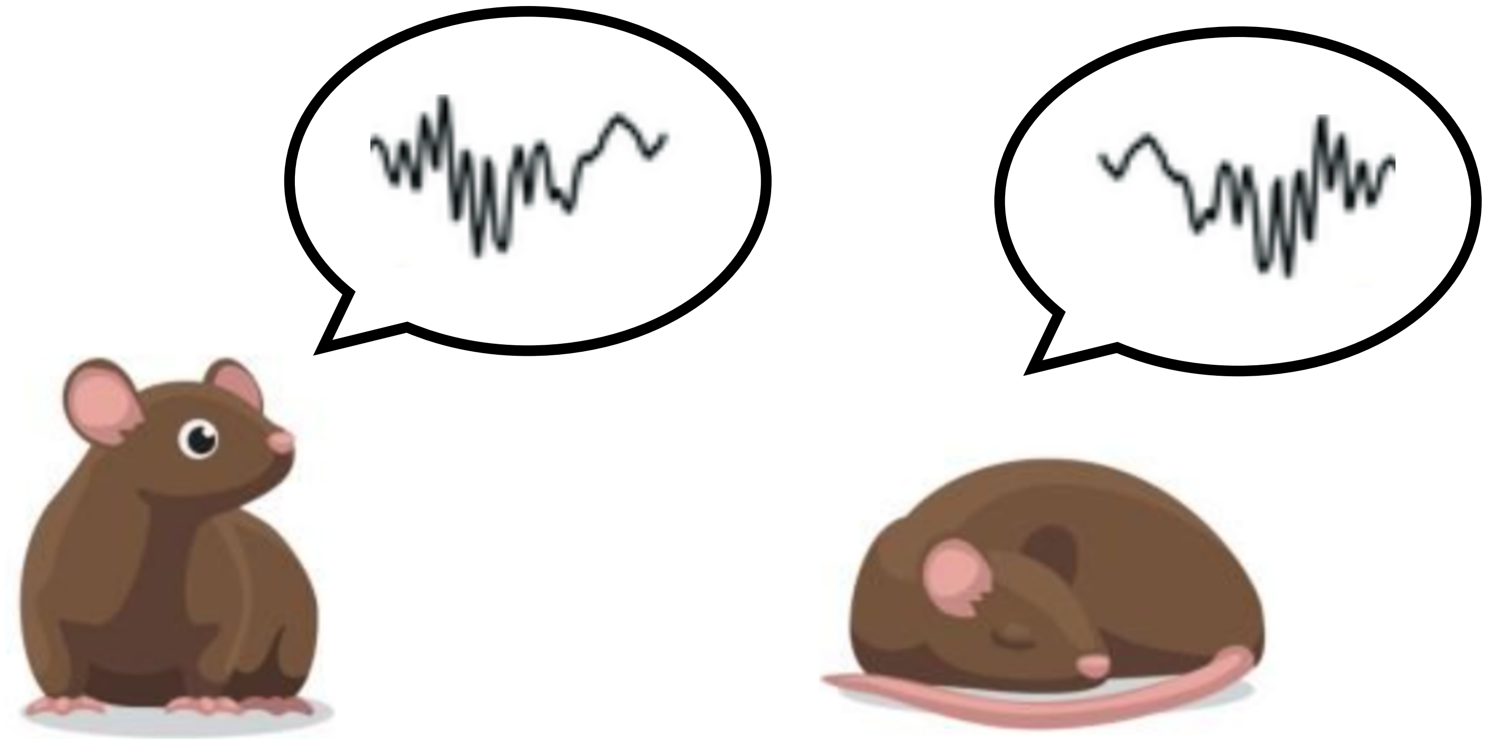
Experience Replay

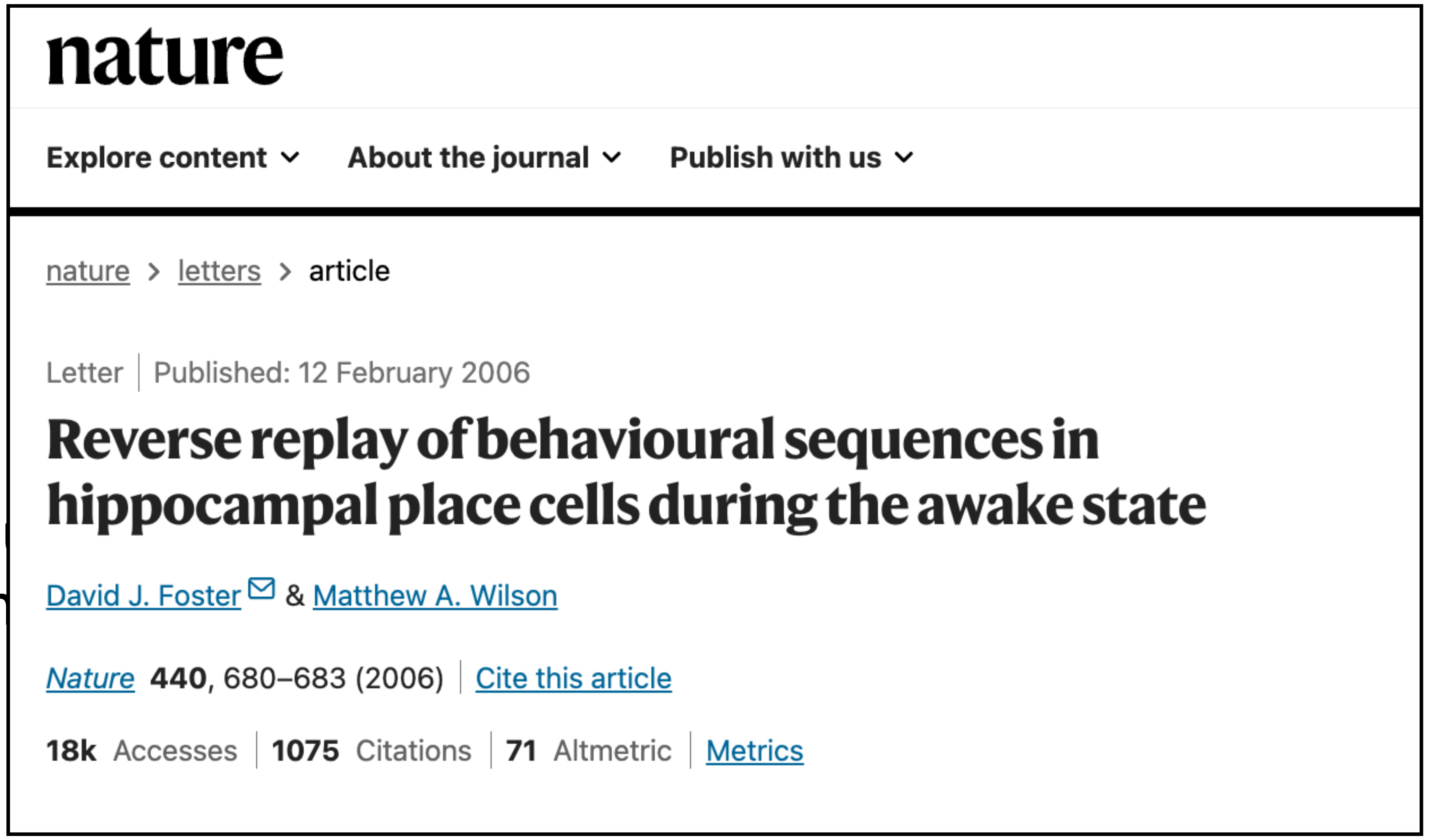
- **Experience Replay (ER)**
  - An agent stores past experiences and randomly samples (replays) transitions during the Q-learning process.
  - Many variants have been proposed.

    - Prioritized Experience Replay

    - Hindsight Experience Replay

- **Reverse Experience Replay (RER)**
  - Inspired by sequential replay occurs in the rat hippocamp [1] — a region of the brain crucial for memory formation





nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > letters > article

Letter | Published: 12 February 2006

Reverse replay of behavioural sequences in hippocampal place cells during the awake state

David J. Foster & Matthew A. Wilson

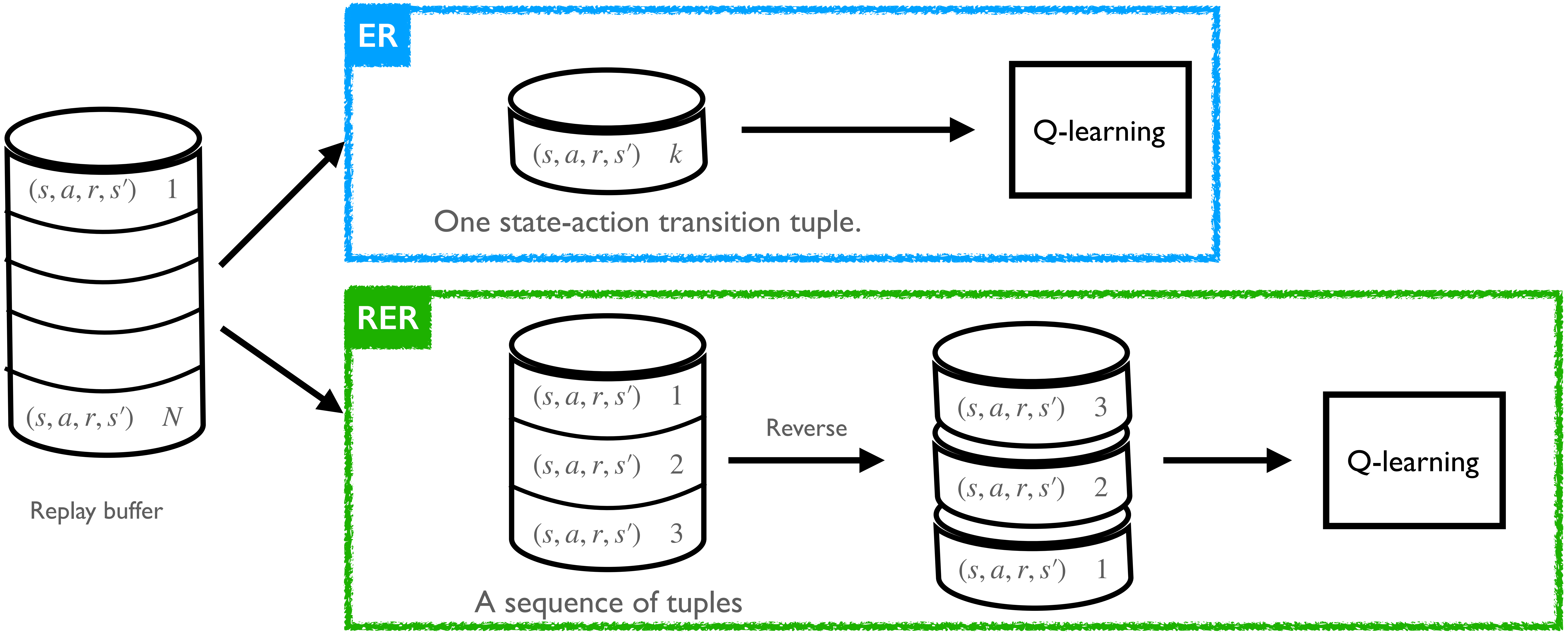
Nature **440**, 680–683 (2006) | Cite this article

18k Accesses | **1075** Citations | **71** Altmetric | Metrics

[1] Foster, David J., and Matthew A. Wilson. "Reverse replay of behavioural sequences in hippocampal place cells during the awake state." Nature (2006)

Reverse Experience Replay - based Q-learning

- Samples ***consecutive sequences*** of transitions (of length L) from the replay buffer.
- Q-learning updates are performed in the ***reverse order*** of the sampled sequences.



Our contribution

- RER shows fast convergence speed both empirically [2] and theoretically [3].

- However, the latest theoretical analysis only holds for a small learning rate (η) and short sampled sequences (length L):

$$\eta L < 1/3$$

- We provide a new idea for analyzing RER, offering theoretical support that RER converges with ***a larger learning rate*** and over ***longer sequences***.

[2] Rotinov, Egor. "Reverse experience replay." arXiv:1910.08780 (2019).
[3] Agarwal, Naman, et al. "Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps." ICLR, 2021

Necessary Assumptions of RER

- Linear MDP Assumption:
  - Reward function: can be written as the inner product of the parameter $w \in \mathbb{R}^d$ and the feature function $\phi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$.
  - Transition probability: proportional to its corresponding feature $P(\cdot \mid s, a) \propto \phi(s, a)$.
- The Q function is computed as: $Q(s, a; w) = \langle w, \phi(s, a) \rangle$
- **Convert the error of Q function to the error of learned parameter w by Linear MDP**

$$\varepsilon(s, a) = \hat{Q}(s, a) - Q^*(s, a) \Leftrightarrow \hat{w} - w^*$$

Estimated Q after some iterations

Actual Q

Analysis of the Error

- The error breaks into two parts (Lemma 3):

$$\hat{w} - w^* = \underbrace{\Gamma_L \left(w_1 - w^*\right)}_{\text{Bias term}} + \eta \underbrace{\sum_{l=1}^{L} \varepsilon_l \Gamma_{l-1} \phi_l}_{\text{variance term}} .$$

where $s_1 \xrightarrow{a_1, r_1} s_2 \xrightarrow{a_2, r_2} s_3 \rightarrow \ldots \rightarrow s_L$ is the sampled sequence, and we denote:

$$\Gamma_L = \left(\mathbf{I} - \eta \phi_1 \phi_1^\top\right) \left(\mathbf{I} - \eta \phi_2 \phi_2^\top\right) \ldots \left(\mathbf{I} - \eta \phi_L \phi_L^\top\right), \text{with each } \phi_L = \phi(s_L, a_L)$$

- **The bias term can reduce to zero if $\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right]$ is bounded (Lemma C.2), i.e.,**

Need to prove: $\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] \preceq$ **A proper bound**

Result from previous method

Expanded by definition,

$$\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] = \mathbb{E}_{(s,a)\sim\mu}\left[\left(\mathbf{I} - \eta\phi_L\phi_L^\top\right)\cdots\left(\mathbf{I} - \eta\phi_1\phi_1^\top\right)\left(\mathbf{I} - \eta\phi_1\phi_1^\top\right)\cdots\left(\mathbf{I} - \eta\phi_L\phi_L^\top\right)\right]$$

$$= \mathbf{I} - 2\eta\mathbb{E}_{(s,a)\sim\mu}\left[\sum_{l=1}^{L}\phi_l\phi_l^\top\right] + \mathbb{E}_{(s,a)\sim\mu}\left[\sum_{k=2}^{2L}(-\eta)^k \sum_{l_1,\ldots,l_k}\phi_{l_1}\phi_{l_1}^\top\ldots\phi_{l_k}\phi_{l_k}^\top\right].$$

Requirement: $\eta L < 1/3$

$$\preceq \eta\sum_{l=1}^{L}\mathbb{E}_{(s,a)\sim\mu}\left[\phi_l\phi_l^\top\right]$$

So previous method upper bounds the large summation with a strong assumption.

$$\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] \preceq \mathbf{I} - \eta\sum_{l=1}^{L}\mathbb{E}_{(s,a)\sim\mu}\left[\phi_l\phi_l^\top\right] \preceq \left(1 - \frac{\eta L}{\kappa}\right)\mathbf{I}$$

Our Result

The expansion is:

$$\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] = \mathbb{E}_{(s,a)\sim\mu}\left[\left(\mathbf{I}-\eta\phi_L\phi_L^\top\right)\cdots\left(\mathbf{I}-\eta\phi_1\phi_1^\top\right)\left(\mathbf{I}-\eta\phi_1\phi_1^\top\right)\cdots\left(\mathbf{I}-\eta\phi_L\phi_L^\top\right)\right]$$

$$= \mathbf{I} - 2\eta\mathbb{E}_{(s,a)\sim\mu}\left[\sum_{l=1}^{L}\phi_l\phi_l^\top\right] + \boxed{\mathbb{E}_{(s,a)\sim\mu}\left[\sum_{k=2}^{2L}(-\eta)^k\sum_{l_1,\ldots,l_k}\phi_{l_1}\phi_{l_1}^\top\ldots\phi_{l_k}\phi_{l_k}^\top\right]}.$$

Requirement: $\cancel{\eta L < 1/3}$ $\quad 0 < \eta \leq 1$

$\leq \eta \cancel{\sum_{l=1}^{L}\mathbb{E}_{(s,a)\sim\mu}\left[\phi_l\phi_l^\top\right]}$ A tighter bound

We show it can be upper bound with a weaker assumption using the proposed combinatorial counting.
The upper bound becomes:

$$\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] \leq \left(1 - \frac{\eta(4-2L)L + L - (1-\eta)^{L-1}L - \eta^2 L}{\kappa}\right)\mathbf{I},$$

Our idea: combinatorially counting the big summation

To tackle: $\mathbb{E}_{(s,a)\sim\mu}\left[\sum_{k=2}^{2L}(-\eta)^k \sum_{l_1,\ldots,l_k} \phi_{l_1}\phi_{l_1}^\top \ldots \phi_{l_k}\phi_{l_k}^\top\right]$

- Lemma 1: for non-zero vector $\mathbf{x}$:

$$|\mathbf{x}^\top \phi_{l_1}\phi_{l_1}^\top \ldots \phi_{l_k}\phi_{l_k}^\top \mathbf{x}| \leq \frac{1}{2}\mathbf{x}^\top \left(\phi_{l_1}\phi_{l_1}^\top + \phi_{l_k}\phi_{l_k}^\top\right)\mathbf{x}$$

  - It is a relaxation: only depend on l_1 and l_k.

- The summation containing a combinatorial number of elements becomes:

$$\sum_{l_1,\ldots,l_k} \phi_{l_1}\phi_{l_1}^\top \ldots \phi_{l_k}\phi_{l_k}^\top \qquad \preceq \sum_{(l_1,l_k)} \frac{1}{2}\left(\phi_{l_1}\phi_{l_1}^\top + \phi_{l_k}\phi_{l_k}^\top\right) = \sum_{l=1}^{L} C_l \cdot \phi_l\phi_l^\top$$

Represents the number of combinations $\phi_{l_1}\phi_{l_1}^\top \ldots \phi_{l_k}\phi_{l_k}^\top$ that start/end with ϕ_l

Our idea: combinatorially counting the big summation

$$\sum_{(l_1,l_k)} \frac{1}{2}\left(\phi_{l_1}\phi_{l_1}^\top + \phi_{l_k}\phi_{l_k}^\top\right) \qquad \Rightarrow \sum_{l=1}^{L} C_l \cdot \phi_l \phi_l^\top$$

Count how many cases of picking valid l_1 and l_k at each possible position in the consecutive sequence of state-action-reward tuples.

"////" indicates this slot cannot be chosen.

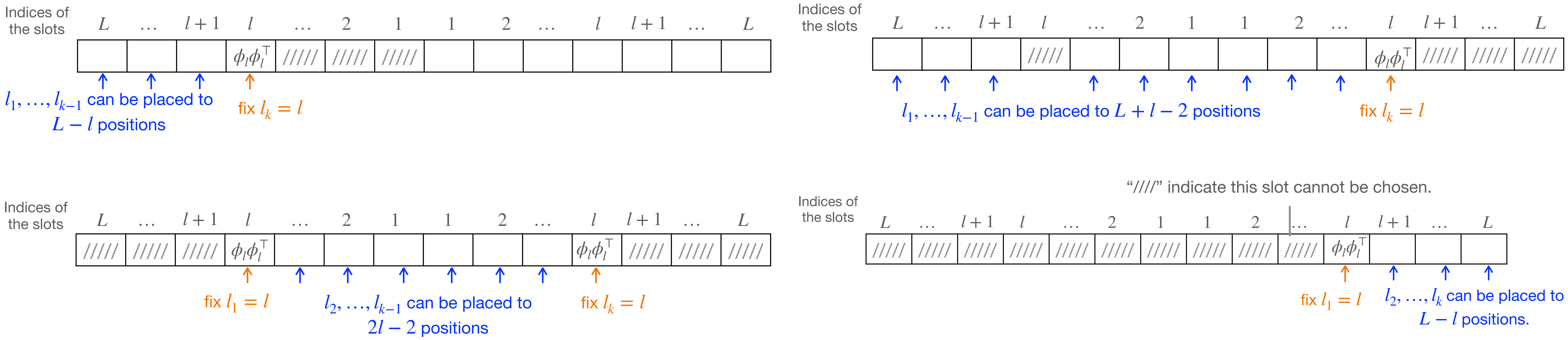
Indices of the slots	L	$\dots$	$l+1$	l	$\dots$	2	1	1	2	$\dots$	l	$l+1$	$\dots$	L
	/////	/////	/////	$\phi_l\phi_l^\top$							/////			

fix $l_1 = l$

$l_2, \dots, l_k$ can be placed to $L + l - 2$ many positions.

In this example, the count is: $\displaystyle\sum_{l=1}^{L} \binom{L+l-2}{k-1} \phi_l \phi_l^\top$

The rest cases (omitted)



- Finally:

$$\underbrace{\sum_{l_1,\dots,l_k}}_{\text{Sum over extensive terms}} \phi_{l_1}\phi_{l_1}^\top \dots \phi_{l_k}\phi_{l_k}^\top \preceq \sum_{l=1}^{L} \Bigg(\underbrace{\binom{L+l-2}{k-1} + \binom{L-l}{k-1} + \binom{2l-2}{k-2}}_{\text{Re-weighted sum}} \Bigg) \phi_l\phi_l^\top$$

Main convergence is improved

Theorem 2. *For Linear MDP, assume the reward function, as well as the feature, is bounded $R(s,a) \in [0,1]$, $\|\phi(s,a)\|_2 \leq 1$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Let T be the maximum episodes, N be the frequency of the target network update, η be the learning rate and L be the length of sequence for RER described in Algorithm 1. When $\eta \in (0,1), L \geq 1$, with sample complexity*

$$\mathcal{O}\left(\frac{\gamma^{T/N}}{1-\gamma} + \sqrt{\frac{T\kappa}{N\delta(1-\gamma)^4}}\exp\left(-\frac{N(\eta(4-2L)L + L - \eta^2 L)}{\kappa}\right) + \sqrt{\frac{\eta\log(\frac{T}{N\delta})}{(1-\gamma)^4}}\right),$$

$\|Q_T(s,a) - Q^*(s,a)\|_\infty \leq \varepsilon$ *holds with probability at least* $1-\delta$.

Summary

- We tighten the convergence analysis using combination-counting, which is particularly well-suited for RER.
- With the new bound: $\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] \preceq \left(1 - \frac{\eta(4-2L)L + L - (1-\eta)^{L-1}L - \eta^2 L}{\kappa}\right)\mathbf{I},$
  - When learning rate η and sequence length L satisfies $\eta L < 1/3$, it provides a tighter bound on the Q-learning error. (Theorem 2 and Lemma 4)
  - The bound is applicable for boarder cases (only needs $0 < \eta < 1$).
- We believe that RER has great potential and warrants further study.

Thank You

- Contact:
  - Nan Jiang: jiang631@purdue.edu
  - Jinzhao Li: li4255@purdue.edu
  - Yexiang Xue: yexiang@purdue.edu

PURDUE UNIVERSITY®

References

- Rotinov, Egor. "Reverse experience replay." arXiv preprint arXiv:1910.08780 (2019).
- Techniques to Improve the Performance of a DQN Agent, https://towardsdatascience.com/techniques-to-improve-the-performance-of-a-dqn-agent-29da8a7a0a7e
- Agarwal, Naman, et al. "Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps." arXiv preprint arXiv:2110.08440 (2021).
- Zanette, Andrea, et al. "Learning near optimal policies with low inherent bellman error." International Conference on Machine Learning. PMLR, 2020.

Sketch of Pipeline

- Convert the error of Q function to the error of learned parameter w (Linear MDP)

$$\varepsilon(s,a) = \hat{Q}(s,a) - Q^*(s,a) \Leftrightarrow \hat{w} - w^*$$

- The error breaks into two parts (Lemma 3):

$$\hat{w} - w^* = \underbrace{\Gamma_L \left(w_1 - w^*\right)}_{\text{Bias term}} + \eta \underbrace{\sum_{l=1}^{L} \varepsilon_l \Gamma_{l-1} \phi_l}_{\text{variance term}} .$$

- The bias term reduces to zero if $\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right]$ is bounded (Lemma C.2).

$$\mathbb{E}_{(s,a)\sim\mu}\left[\Gamma_L^\top \Gamma_L\right] \preceq \text{Some upper bound}$$

Details

Figure 1: Case 1 in the propose combinatorial counting procedure. The task is to count how many terms $\phi_{l_1}\phi_{l_1}^\top \dots \phi_{l_k}\phi_{l_k}^\top$ can be "reduced to" $\phi_l\phi_l^\top$ for a fixed l using Lemma 1, for $1 \le l \le L$. When we let l_1 pick the left l-th slot, l_k cannot choose the left terms with indices $L, \dots, l+1$. Because of the sequential ordering constraint l_i should be on the right of l_{i-1}. To avoid double counting, we also disallow assigning the right l-th slot to l_k. There are $2L - (L - (l+1)) - 1 = L + l - 2$ many slots to assign the rest sequences $l_2, \dots, l_k$ of length $k-1$. Therefore, we obtain $\binom{L+l-2}{k-1}$ many terms for the first case. See all the rest cases in Figure 2 in the appendix.

Bounds for Bias and variance terms are improved

The convergence requirement is relaxed from

$$\eta * L < 1/3$$

to

$$0 < \eta \leq 1$$

Lemma 4 (Bound on the bias term)**.** *Let* $\mathbf{x} \in \mathbb{R}^d$ *be a non-zero vector and* N *is the frequency for the target network to be updated. For* $\eta \in (0,1), L \in \mathbb{N}$ *and* $L > 1$*, the following matrix's positive semi-definite inequality holds with probability at least* $1-\delta$*:*

$$\mathbb{E}\left\|\prod_{j=N}^{1} \Gamma_L \mathbf{x}\right\|_{\phi}^{2} \leq \exp\left(-\frac{N(\eta(4-2L)L + L - \eta^2 L)}{\kappa}\right)\sqrt{\tfrac{\kappa}{\delta}}\, \|\mathbf{x}\|_{\phi}\,.$$

The ϕ*-based norm is defined in Definition 1.*