SOOCHOW UNIVERSITY

A framework of dual replay buffer: balancing forgetting and generalization in reinforcement learning

Linjing Zhang, Zongzhang Zhang , Zhiyuan Pan , Yingfeng Chen, Jiangcheng Zhu, Zhaorong Wang, Meng Wang, Changjie Fan

School of Computer Science and Technology
Soochow University

Motivation

- **Experience replay improves sample efficiency and training stabilization for deep reinforcement learning methods.**
- **Traditional retention method: FIFO**
- **However, this leads to the problem of generalization and forgetting in long-time training.**

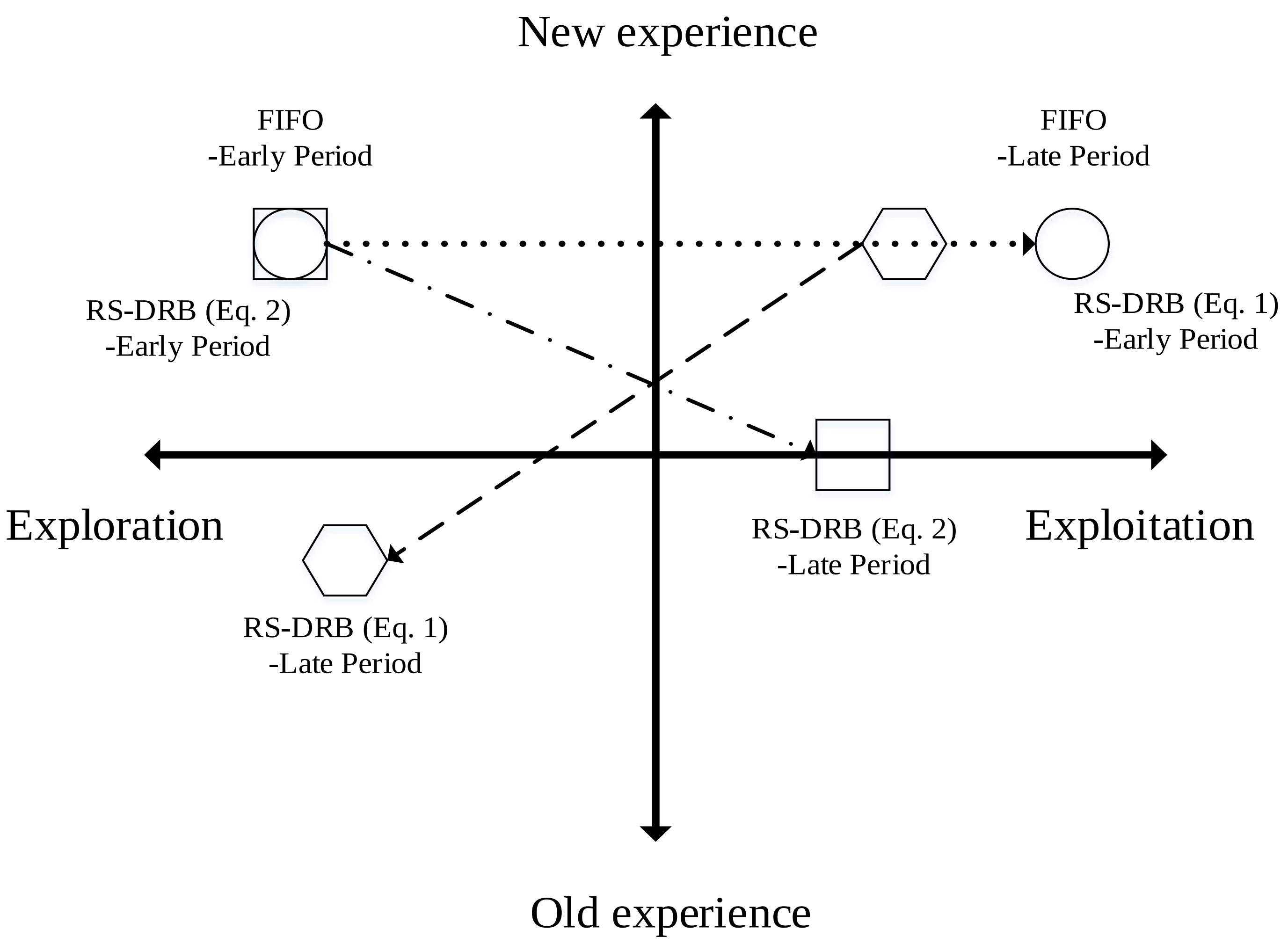
Motivation

- Generalization
  - Stuck in a **small** region of the state space
  - Experiences are **overfitted** and almost the same
- Catastrophic forgetting
  - Forgetting the knowledge obtained previously

Related work

- **Prioritized experience replay (PER)**
  - Focus on the instantaneous utility of experiences and implements the prioritized sampling in replay buffer based on the TD error
- **Synthetic experiences**
  - Two replay buffers with FIFO and a distance-based retention policy
- **Proxies**
  - To guide the retention and sampling of replay buffer via prior knowledge on control problems
- **Hindsight experience replay (HER)**
  - To deal with sparse and binary rewards

Catastrophic forgetting



The stream of state distribution of training batch

Method

- Reservoir sampling

$$P[(s,a,r,s')_i] = \frac{k}{i} \times \prod_{n=1}^{S(\mathcal{D}_a)-i} (1 - \frac{k}{i+n} \times \frac{1}{k})$$
$$= \frac{k}{i} \times \prod_{n=1}^{S(\mathcal{D}_a)-i} (\frac{i+n-1}{i+n})$$
$$= \frac{k}{S(\mathcal{D}_a)}$$

- Double replay buffers
  - Exploration buffer (Reservoir Sampling) and exploitation buffer (FIFO)
- Sampling ratio
  - To sample the experiences of training batch from two buffers
  - Adaptive to the policy update rate

Double replay buffers

- Exploration is necessary to search the **entire** state space
- Discrete action problems (based on DQN)
  - An ϵ-greedy policy to control the magnitude of the exploration
- Continuous action problems(based on DDPG)
  - A noise $\mathcal{N}$ to drive exploration
- A threshold η is to determine whether the action belongs to exploration action a_r or exploitation action a_g.

Sampling ratio

- A training batch size is N_b

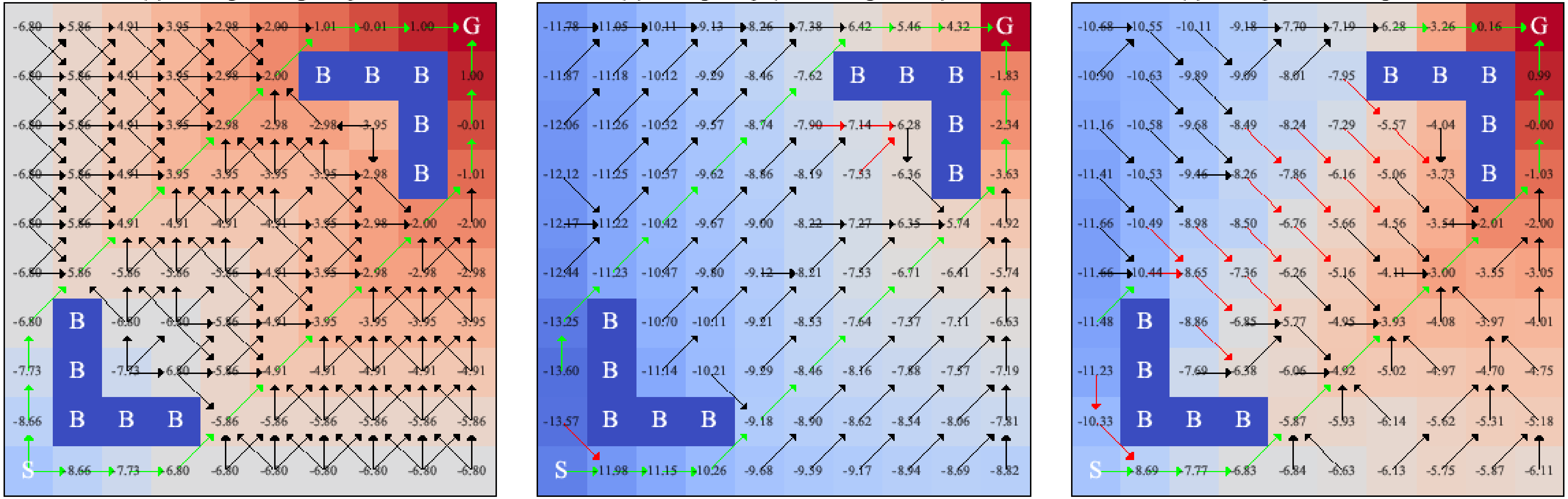
  The number of the same actions n_b from two sets of actions

$$
\begin{cases}
\tau = \dfrac{n_b}{N_b} \times \mathcal{T}_{max} & (1) \\
\tau = \max\{\epsilon, \dfrac{n_b}{N_b} \times \mathcal{T}_{max}\} & (2)
\end{cases}
$$

- τN_b experiences are sampled from exploration buffer D_r, and the rest ones are sampled from exploitation buffer D_g

Experiment

- GridWorld (10×10)
  - Eight actions
  - Reward -1 every state except terminate state



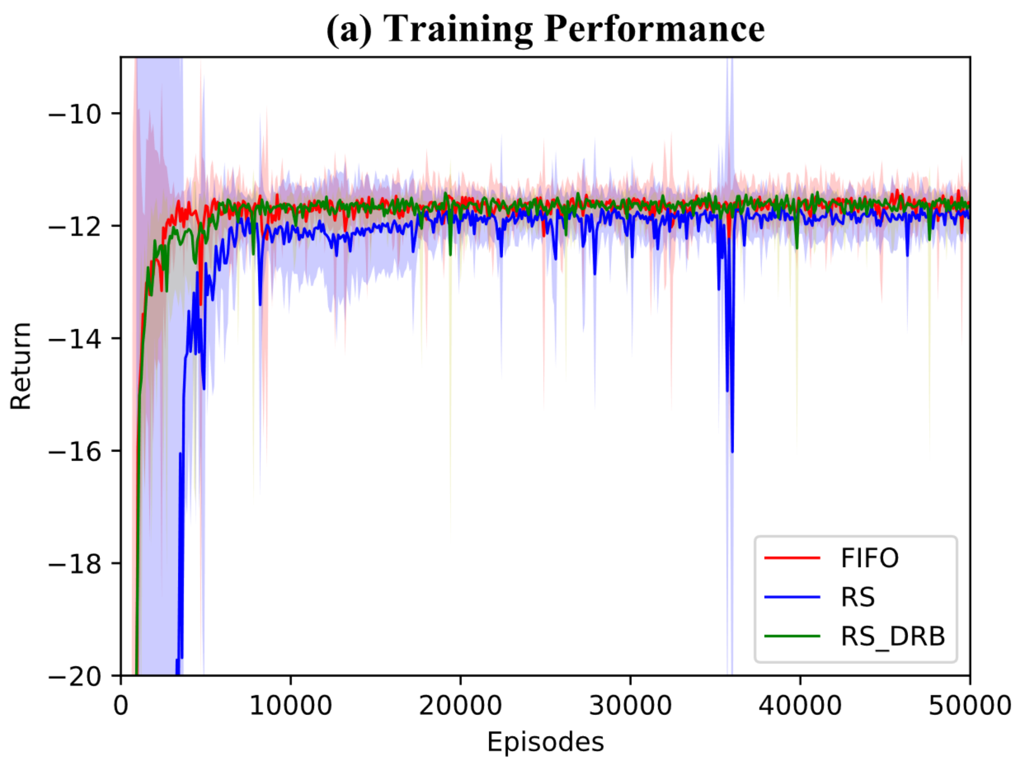
(a) The optimal policy

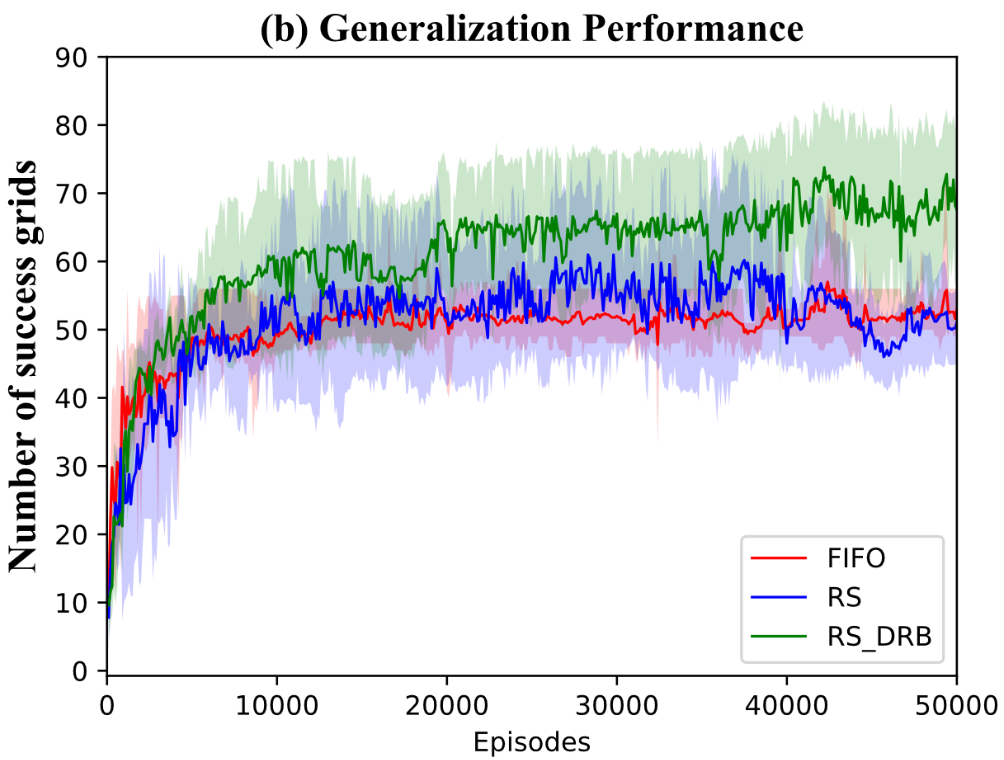
(b) Best policy (at 2000 episodes)

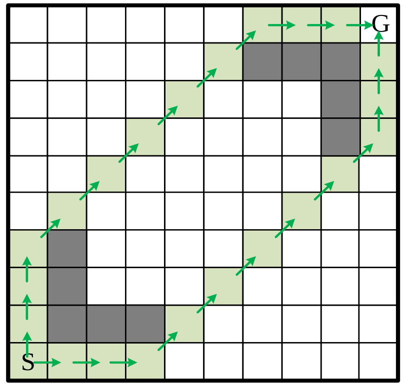
(c) Policy at 30000 episodes

Experiment

- Discrete Problem: A Barriered GridWorld



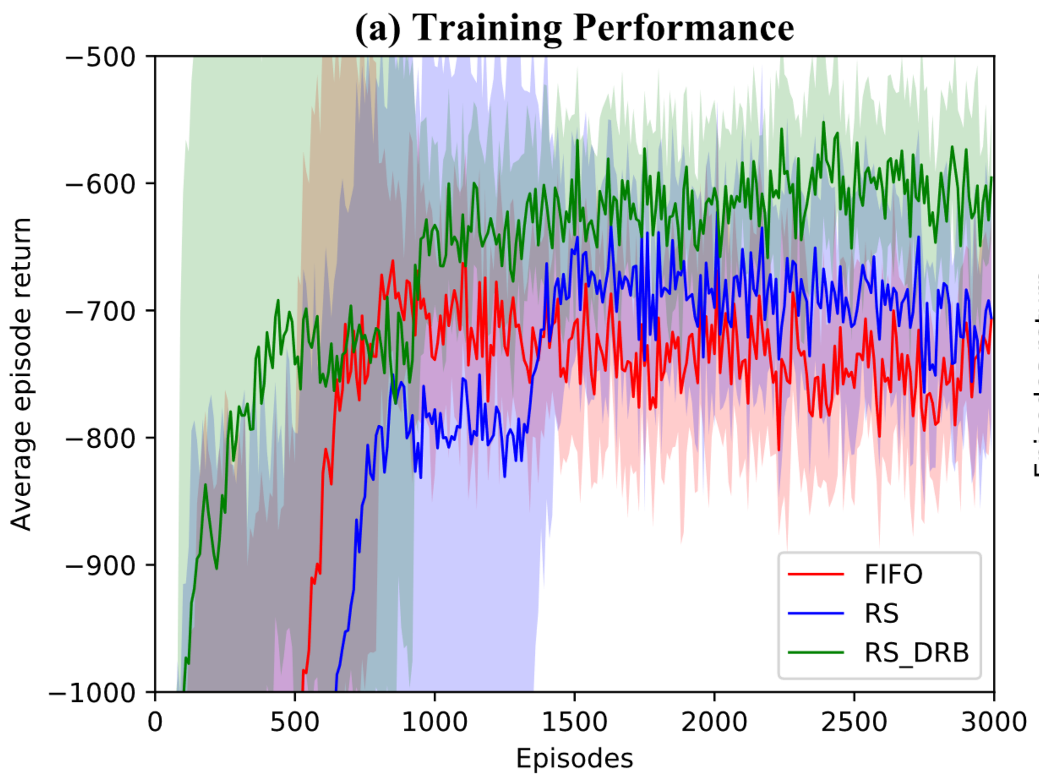


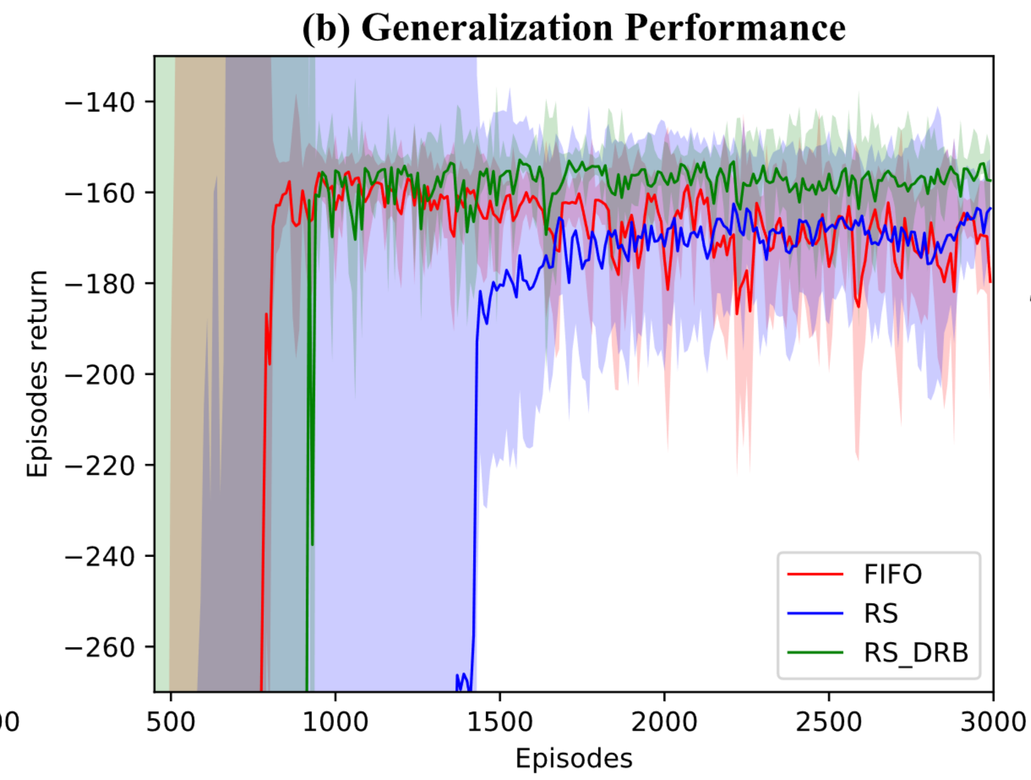


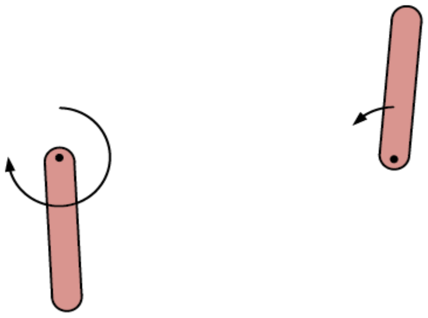
A Barriered GridWorld

Experiment

- Continuous problem: Pendulum







Pendulum

Conclusion

- Our paper presented a new RS-DRB framework to retain the experiences in the replay buffer.
- The exploration buffer with the reservoir sampling helps to maintain the coverage of the entire state space.
- The adaptive sampling ratio balances the experiences sampled from these two buffers according to the change of the policy.

Thanks!