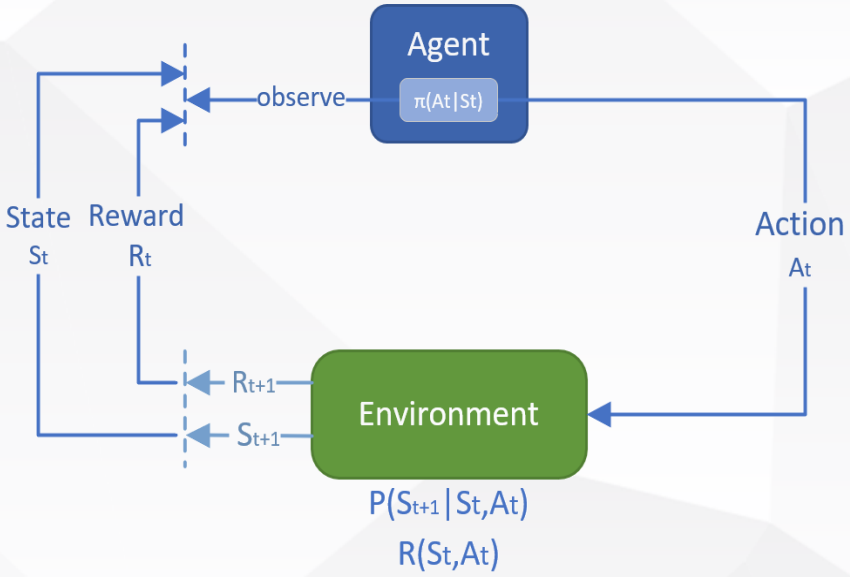
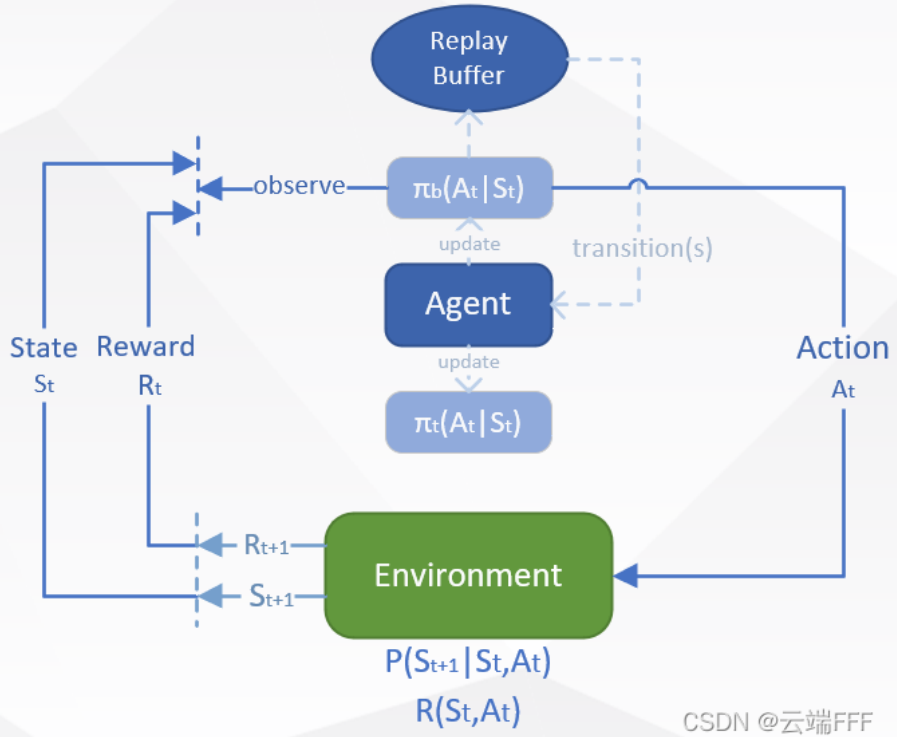
Regret Minimization Experience Replay in Off-Policy Reinforcement Learning

Xu-Hui Liu*, **Zhenghai Xue***, **Jing-Cheng Pang**, **Shengyi Jiang**, **Feng Xu**, **Yang Yu**†
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
liuxh@lamda.nju.edu.cn, xuezh@smail.nju.edu.cn
{pangjc, jiangsy, xufeng}@lamda.nju.edu.cn, yuy@nju.edu.cn

NIPS 2021

Experience Replay





- eliminate circular dependencies
- higher data efficiency
- better data distribution (i.i.d)

Non-uniform Experience Repaly

- Key Idea: RL agent can learn more effectively from some transitions than from others
- Any loss function evaluated with non-uniformly sampled data can be transformed into another uniformly sampled loss function with the same expected gradien

importance sampling ratio

$$\underbrace{\mathbb{E}_{i\sim\mathcal{D}_1}\left[\nabla_Q \mathcal{L}_1(\delta(i))\right]}_{\text{expected gradient of } \mathcal{L}_1 \text{ under } \mathcal{D}_1} = \mathbb{E}_{i\sim\mathcal{D}_2}\left[\frac{p_{\mathcal{D}_1}(i)}{p_{\mathcal{D}_2}(i)}\nabla_Q \mathcal{L}_1(\delta(i))\right].$$

$$\nabla_Q \mathcal{L}_2(\delta(i)) = \frac{p_{\mathcal{D}_1}(i)}{p_{\mathcal{D}_2}(i)}\nabla_Q \mathcal{L}_1(\delta(i))$$

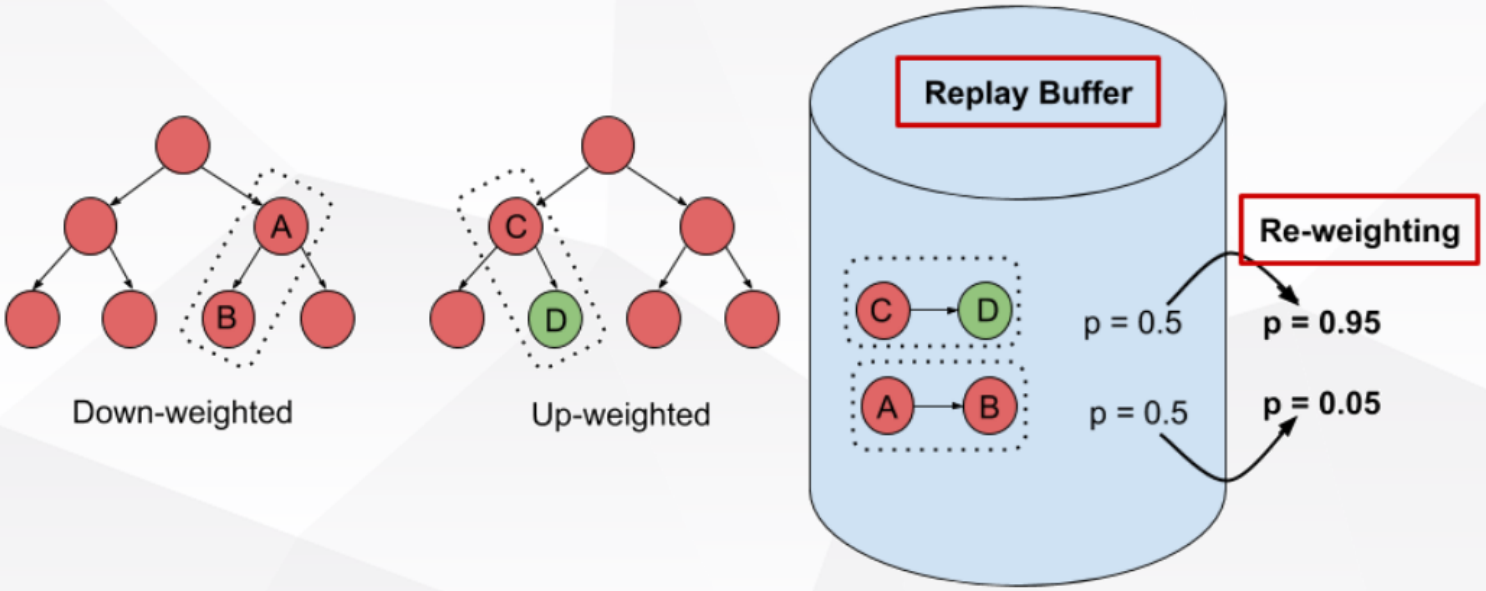
$$\mathbb{E}_{\mathcal{D}_1}\left[\nabla_Q \mathcal{L}_1(\delta(i))\right] = \mathbb{E}_{\mathcal{D}_2}\left[\nabla_Q \mathcal{L}_2(\delta(i))\right]$$

Fujimoto S, Meger D, Precup D. An equivalence between loss functions and non-uniform sampling in experience replay[J]. Advances in neural information processing systems, 2020, 33: 14219-14230.

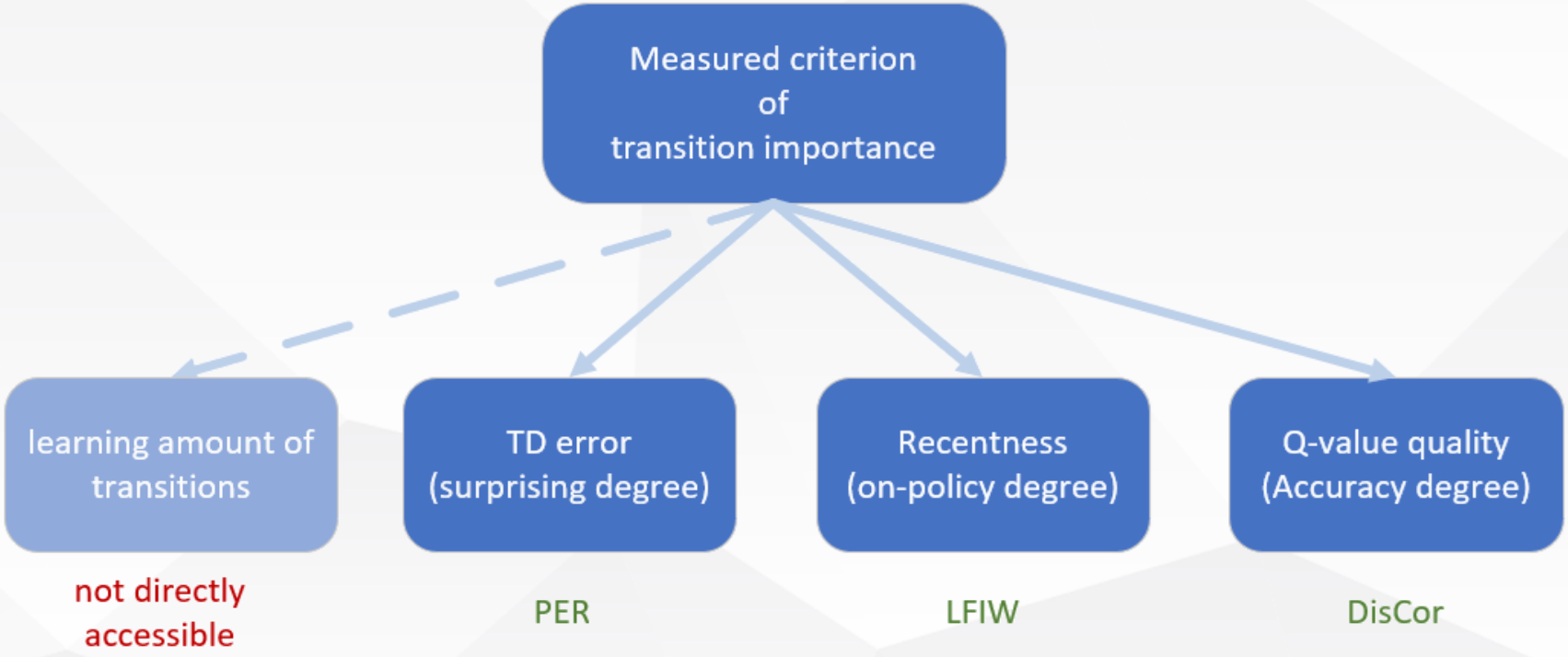
Non-uniform Experience Repaly

$$\theta \leftarrow \arg\min_{\theta} \mathbb{E}_{s,a\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \left(r(s,a) + \gamma \mathbb{E}_{s'|s,a}[\max_{a'} \bar{Q}(s',a')]\right)\right)^2\right]$$

$$Q_k \leftarrow \arg\min_{Q} \frac{1}{N}\sum_{i=1}^{N} w_i(s,a)\cdot\left(Q(s,a) - [r(s,a) + \gamma Q_{k-1}(s',a')]\right)^2$$



Non-uniform Experience Repaly



PER

the priority of transition i

$$\begin{cases} p(i) = \dfrac{|\delta(i)|^{\alpha} + \epsilon}{\sum_j (|\delta(j)|^{\alpha} + \epsilon)} \\ p(i) = \dfrac{(\frac{1}{\text{rank(i)}})^{\alpha}}{\sum_j (\frac{1}{\text{rank(j)}})^{\alpha}} \end{cases}$$

between uniform sampling and greedy sampling

power-law distribution with exponent α (more robust)

$$\mathcal{L}_{\text{PER}} = w(i)\mathcal{L}(\delta(i))$$

$$\mathcal{L}_i = \mathbb{E}_{s,a\sim\rho(\cdot)}\left[(y_i - Q(s,a,\theta_i)^2)\right]$$

$$w_i = (\frac{1}{N} \cdot \frac{1}{p(i)})^{\beta}$$ ← annealing to 1

$$w(i) = \frac{w_i}{\max_j w_j}$$ for stability reasons

LFIW

Bellman equation $Q^\pi(s,a) = \mathcal{B}^\pi Q^\pi(s,a)$

Bellman operator

$$\mathcal{B}^\pi Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s',a'}[Q(s',a')]$$

loss for Q-network $L_Q(\theta;\mathcal{D}) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[(Q_\theta(s,a) - \hat{\mathcal{B}}^\pi Q_\theta(s,a))^2\right]$

replay buffer

Tend to $\mathcal{B}^\pi$

introduce prioritization $L_Q(\theta;d,w) = \mathbb{E}_d\left[w(s,a)(Q_\theta(s,a) - \mathcal{B}^\pi Q_\theta(s,a))^2\right]$

sampling distribution (buffer distribution)

objective $\arg\min_\theta L_Q(\theta;d,w) = \arg\min_\theta L_Q(\theta;d^w)$

select a favorable priority distribution $d^w \propto d \cdot w$

$$d^w = d^\pi$$

LFIW

- Estimate the density ratio only rely on samples (e.g. from the replay buffer)

Lemma 1 ([27]). *Assume that f has first order derivatives f' at $[0, +\infty)$. $\forall P, Q \in \mathcal{P}(\mathcal{X})$ such that $P \ll Q$ and $w : \mathcal{X} \to \mathbb{R}^+$,*

$$D_f(P\|Q) \geq \mathbb{E}_P[f'(w(\boldsymbol{x}))] - \mathbb{E}_Q[f^*(f'(w(\boldsymbol{x})))] \tag{9}$$

where f^* denotes the convex conjugate and the equality is achieved when $w = \mathrm{d}P/\mathrm{d}Q$.

$$D_f(p\,\|\,q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

f -divergences

$$w(s,a) := d^{\pi}(s,a)/d^{D}(s,a)$$

- Two types of repaly buffer
  - smaller (faster) replay buffer ⟶ smaller size, more on-policiness
  - regular (slow) replay buffer ⟶ bigger size, more off-policiness

LFIW

- Estimate the density ratio via minimizing the follow objective over network $w_\psi(x)$

$$L_w(\psi) := \mathbb{E}_{\mathcal{D}_\mathrm{s}}[f^*(f'(w_\psi(s,a)))] - \mathbb{E}_{\mathcal{D}_\mathrm{f}}[f'(w_\psi(s,a))]$$

the outputs $w_\psi(s,a)$ are forced to be non-negative via activation functions

- self normalization with temperature hyperparameter T

$$\tilde{w}_\psi(s,a) := \frac{w_\psi(s,a)^{1/T}}{\mathbb{E}_{\mathcal{D}_\mathrm{s}}[w_\psi(s,a)^{1/T}]}$$

- The final objective for TD learning over Q is then

$$L_Q(\theta; d^\pi) \approx L_Q(\theta; \mathcal{D}_\mathrm{s}, \tilde{w}_\psi) := \mathbb{E}_{(s,a)\sim\mathcal{D}_\mathrm{s}}[\tilde{w}_\psi(\boldsymbol{x})(Q_\theta(s,a) - \hat{\mathcal{B}}^\pi Q_\theta(s,a))^2]$$

estimate via MC

DisCor

$$\mathcal{L}(Q) = \mathbb{E}_{s\sim\beta(s),a\sim\pi_k(a|s)}\left[|Q_k(s,a) - Q^*(s,a)|\right].$$

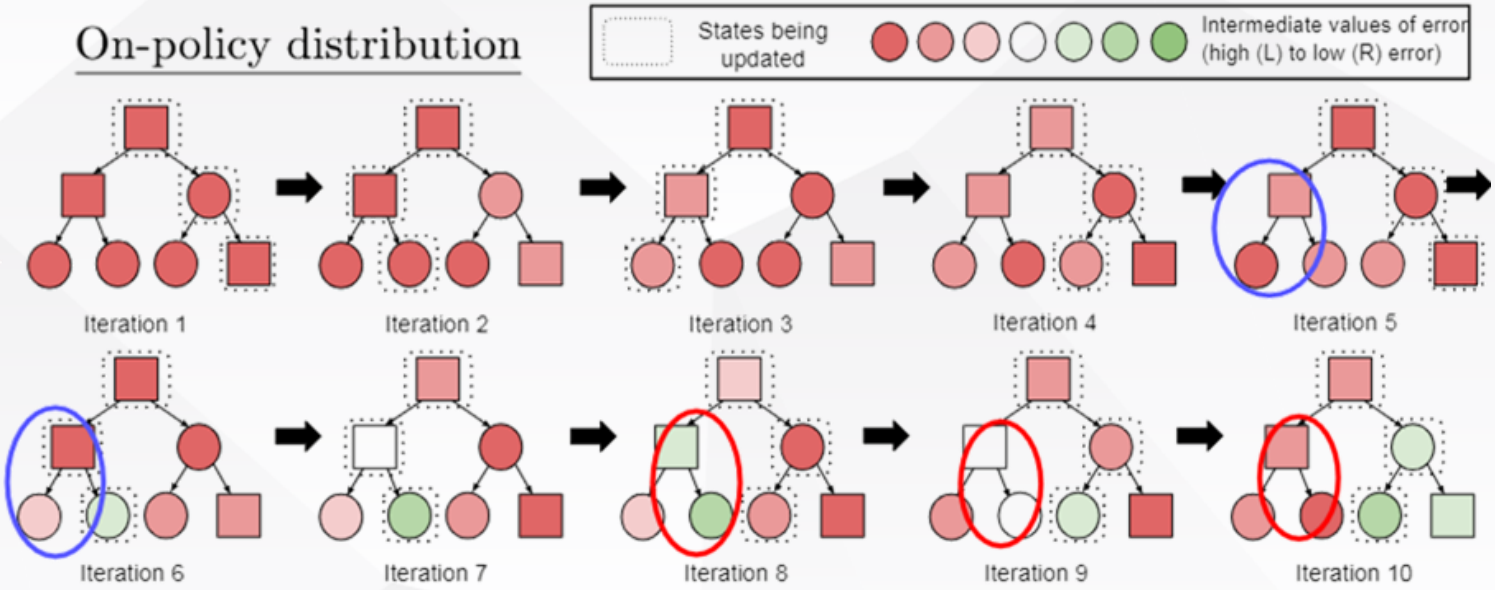
$$\mathcal{L}(Q) = \mathbb{E}_{s\sim\beta(s),a\sim\pi_k(a|s)}\left[|Q_k(s,a) - \mathcal{B}^*Q_k(s,a)|\right]$$

can be a wrong target

precise target

function approximator make things worse

DisCor



- leaf state: rarely visit, provide incorrect TD target
- root state: frequently visit, fit to incorrect target
- state with similar features affect each orther

DisCor

$$\min_{p_k} \mathbb{E}_{d^{\pi_k}} \left[|Q_k - Q^*| \right]$$

$$\text{s.t. } Q_k = \arg\min_{Q} \mathbb{E}_{p_k} \left[\left(Q - \mathcal{B}^* Q_{k-1} \right)^2 \right], \quad \sum_{s,a} p_k(s,a) = 1, \quad \forall s, a \;\; p_k(s,a) \geq 0$$

↓

$$p_k(s,a) \propto \exp\left(-\left|Q_k - Q^*\right|(s,a)\right) \frac{\left|Q_k - \mathcal{B}^* Q_{k-1}\right|(s,a)}{\lambda^*}$$

$$w_k(s,a) = \frac{p_k(s,a)}{\mu(s,a)}$$ densities $\mu(s,a)$ are unknown

↓

$$q_k^* = \arg\min_{q_k} -\mathbb{E}_{q_k} \left[\log p_k \right] + (\tau) \mathrm{D}_{\mathrm{KL}} \left(q_k \| \mu \right)$$

↓

$$q_k^*(s,a) \propto (\mu_k) \cdot \exp\left(\frac{\log p_k(s,a)}{\tau} \right)$$

$$\therefore \frac{q_k^*}{\mu_k} \propto \exp\left(\frac{-\left|Q_k - Q^*\right|(s,a)}{\tau} \right) \frac{\left|Q_k - \mathcal{B}^* Q_{k-1}\right|(s,a)}{\lambda^*}$$

DisCor

$$\frac{q_k^*}{\mu_k} \propto \exp\left(\frac{-|Q_k - Q^*|(s,a)}{\tau}\right) \frac{|Q_k - \mathcal{B}^* Q_{k-1}|(s,a)}{\lambda^*}$$

$$\Delta_k(s,a) + \sum_{i=1}^{k} \gamma^{k-i} \alpha_i \geq |Q_k - Q^*|(s,a),$$

$$\Delta_k(s,a) = |Q_k(s,a) - (\mathcal{B}^* Q_{k-1})(s,a)| + \gamma (P^{\pi_{k-1}} \Delta_{k-1})(s,a).$$

$$\forall s,a \quad c_1 \leq |Q_k - \mathcal{B}^* Q_{k-1}|(s,a) \leq c_2$$

$$\text{where} \quad c_1 = \min_{s,a} |Q_{k-1} - \mathcal{B}^* Q_{k-2}|,$$

$$c_2 = \max_{s,a} |Q_{k-1} - \mathcal{B}^* Q_{k-2}|$$

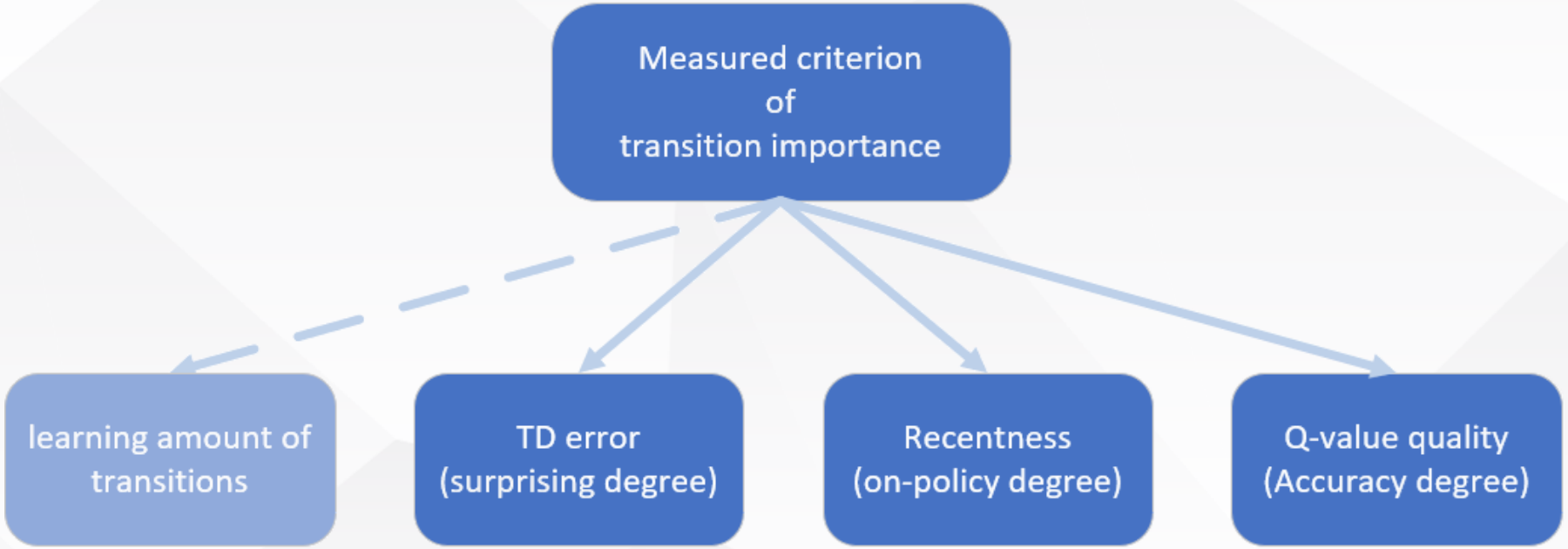
lower bound

$$w_k \propto \exp\left(\frac{-c_2 - \gamma [P^{\pi_{k-1}} \Delta_{k-1}](s,a)}{\tau}\right) \frac{c_1}{\lambda^*}$$

$$w_k(s,a) \propto \exp\left(-\frac{\gamma [P^{\pi_{k-1}} \Delta_{k-1}](s,a)}{\tau}\right).$$

$$P^\pi Q(s,a) := \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(a|s)} \left[Q(s',a')\right]$$

Non-uniform Experience Repaly



not directly accessible

PER

- Non-uniform sampling
- Weighted loss
- Maintain expect gradient
- Bellman Optimal equation

LFIW

- Uniform sampling
- Weighted loss
- Not Maintain expect gradient
- Bellman equation

DisCor

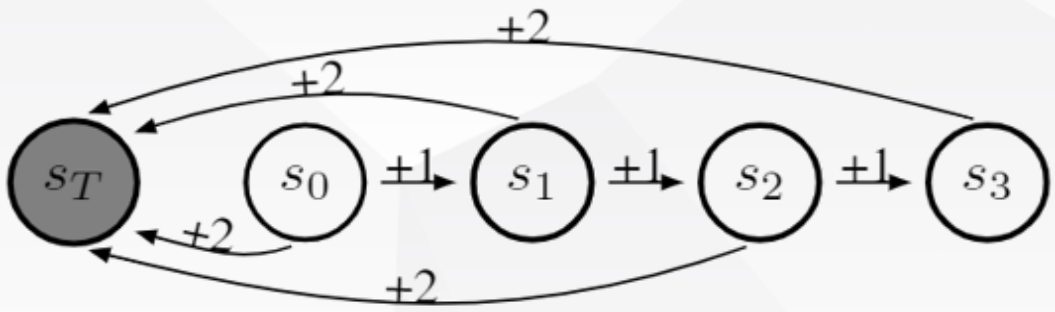
- Uniform sampling
- Weighted loss
- Not Maintain expect gradient
- Bellman Optimal equation & Bellman equation

Regret Minimization Experience Replay in Off-Policy Reinforcement Learning

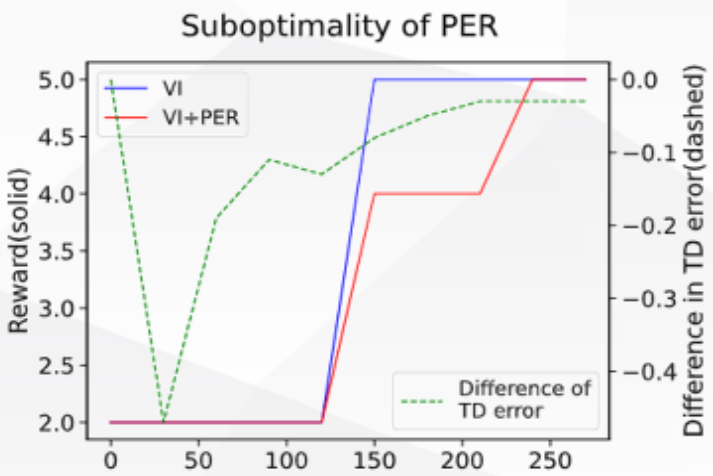
Xu-Hui Liu*, **Zhenghai Xue***, **Jing-Cheng Pang**, **Shengyi Jiang**, **Feng Xu**, **Yang Yu**†
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
liuxh@lamda.nju.edu.cn, xuezh@smail.nju.edu.cn
{pangjc, jiangsy, xufeng}@lamda.nju.edu.cn, yuy@nju.edu.cn

NIPS 2021

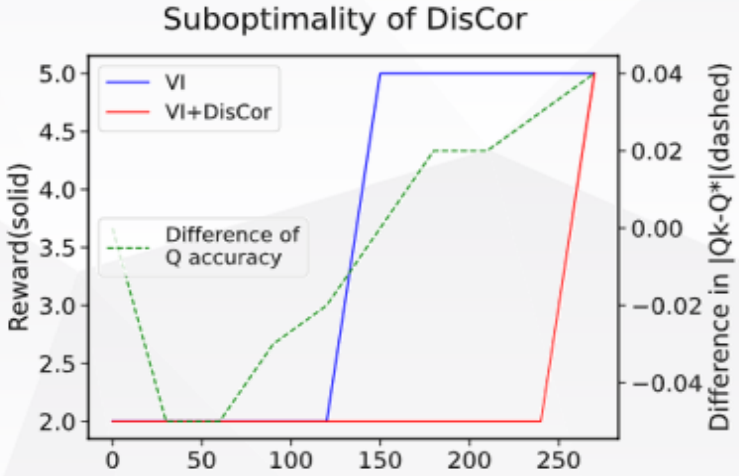
Bad case



(a)



(b)



(c)

Preliminaries

- Expect return of policy

$$\eta(\pi) = \mathbb{E}_{\pi}[\sum_{t\geq 0} \gamma^t r(s_t, a_t)], \text{ 其中 } s_0 \sim \rho_0,\ a_t \sim \pi(\cdot|s_t),\ s_{t+1} \sim T(\cdot|s_t, a_t)$$

$$\lim_{t\to\infty}(\gamma^0 + \gamma^1 + \gamma^2 + \ldots + \gamma^t) r(s,a) = \lim_{t\to\infty} \frac{\gamma^0(\gamma^{t+1} - 1)}{\gamma - 1} r(s,a) = \frac{1}{1-\gamma} r(s,a)$$

$$\eta(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi}(s,a)}[r(s,a)]$$

- Optimal policy

$$\pi^* = \arg\max_{\pi} \eta(\pi)$$

- Regret： the expected loss in return by following policy π instead of the optimal policy

$$\text{Regret}(\pi) = \eta(\pi^*) - \eta(\pi)$$

- Boltzmann exploration policy

$$\pi_k(s) = \frac{\exp(Q_k(s,a))}{\sum_{a'} \exp(Q(s,a'))}$$

Formalize the problem

$$\min_{w_k} \quad \eta(\pi^*) - \eta(\pi_k)$$

$$\text{s.t.} \quad Q_k = \arg\min_{Q \in \mathcal{Q}} \mathbb{E}_\mu[w_k(s,a) \cdot (Q - \mathcal{B}^* Q_{k-1})^2(s,a)],$$

$$\mathbb{E}_\mu[w_k(s,a)] = 1, \quad w_k(s,a) \geq 0,$$

data distribution of the replay buffer

$$\min_{p_k} \mathbb{E}_{d^{\pi_k}} [|Q_k - Q^*|]$$

$$\text{s.t. } Q_k = \arg\min_{Q} \mathbb{E}_{p_k} \left[(Q - \mathcal{B}^* Q_{k-1})^2 \right], \quad \sum_{s,a} p_k(s,a) = 1, \quad \forall s, a \;\; p_k(s,a) \geq 0$$

Analyze computationally

Theorem 1 (Informal). *Under mild conditions, the solution w_k to a relaxation of the optimization problem 1 in MDPs with discrete action spaces is*

$$w_k(s,a) = \frac{1}{Z_1^*}\left(E_k(s,a) + \epsilon_{k,1}(s,a)\right). \tag{2}$$

In MDPs with continuous action spaces, the solution is

$$w_k(s,a) = \frac{1}{Z_2^*}\left(F_k(s,a) + \epsilon_{k,2}(s,a)\right). \tag{3}$$

where

$$E_k(s,a) = \underbrace{\frac{d^{\pi_k}(s,a)}{\mu(s,a)}}_{(a)} \underbrace{\left(2-\pi_k(a|s)\right)}_{(b)} \underbrace{\exp\left(-\left|Q_k - Q^*\right|(s,a)\right)}_{(c)} \underbrace{\left|Q_k - \mathcal{B}^* Q_{k-1}\right|(s,a)}_{(d)}$$

$$F_k(s,a) = 2\underbrace{\frac{d^{\pi_k}(s,a)}{\mu(s,a)}}_{(a)} \underbrace{\exp\left(-\left|Q_k - Q^*\right|(s,a)\right)}_{(c)} \underbrace{\left|Q_k - \mathcal{B}^* Q_{k-1}\right|(s,a)}_{(d)},$$

Z_1^*, Z_2^* are normalization factors, $\epsilon_{k,1}(s,a)$ and $\epsilon_{k,2}(s,a)$ satisfy $\max\left\{\frac{\epsilon_{k,1}(s,a)}{E_k(s,a)}, \frac{\epsilon_{k,2}(s,a)}{F_k(s,a)}\right\} \leq \epsilon_{\pi_k}$.

DisCor

$$\frac{q_k^*}{\mu_k} \propto \exp\left(\frac{-|Q_k - Q^*|(s,a)}{\tau}\right) \frac{|Q_k - \mathcal{B}^* Q_{k-1}|(s,a)}{\lambda^*}$$

ReMERN

$$w_k(s,a) \propto \frac{d^{\pi_k}(s,a)}{\mu(s,a)} \exp\left(-\gamma \left[P^{\pi_{k-1}} \Delta_{k-1}\right](s,a)\right)$$

Algorithm 1 ReMERN

Initialize Q-values $Q_\theta(s,a)$, a replay buffer μ, an error model $\Delta_\phi(s,a)$, and a weight model κ_ψ.

for step k in $\{1,\ldots,N\}$ **do**

  Collect M samples using π_k, add them to replay buffer μ, sample $\{(s_i,a_i)\}_{i=1}^N \sim \mu$.

  Evaluate $Q_\theta(s,a)$, $\Delta_\phi(s,a)$ and $\kappa_\psi(s,a)$ on samples (s_i,a_i).

  Compute target values for Q and Δ on samples:

  $y_i = r_i + \gamma \max_{a'} Q_{k-1}(s_i',a')$.

  $\hat{a}_i = \arg\max_a Q_{k-1}(s_i',a)$.

  $\hat{\Delta} = |Q_\theta(s,a) - y_i| + \gamma \Delta_{k-1}(s_i',\hat{a}_i)$.

  Optimize κ_ψ using

$$L_\kappa(\psi) := \mathbb{E}_{\mathcal{D}_s}\left[f^*\left(f'\left(\kappa_\psi(s,a)\right)\right)\right] - \mathbb{E}_{\mathcal{D}_f}\left[f'\left(\kappa_\psi(s,a)\right)\right].$$

  Compute w_k using

$$w_k(s,a) \propto \frac{d^{\pi_k}(s,a)}{\mu(s,a)} \exp\left(-\gamma \left[P^{\pi^{w_{k-1}}} \Delta_{k-1}\right](s,a)\right).$$

  Minimize Bellman error for Q_θ weighted by w_k.

  $\theta_{k+1} \leftarrow \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_i^N w_k(s_i,a_i)\left(Q_\theta(s_i,a_i) - y_i\right)^2$.

  Minimize ADP error for training ϕ.

  $\phi_{k+1} \leftarrow \underset{\phi}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N \left(\Delta_\phi(s_i,a_i) - \hat{\Delta}_i\right)^2$.

end for

ReMERT

- Estimate the upper bound of $|Q_k - Q^*|$ by temporal Structure

$$|Q_k - Q^*| \leq |Q_k - \mathcal{B}^* Q_{k-1}| + |\mathcal{B}^* Q_{k-1} - Q^*|$$

projection error

error accumulate through the MDP

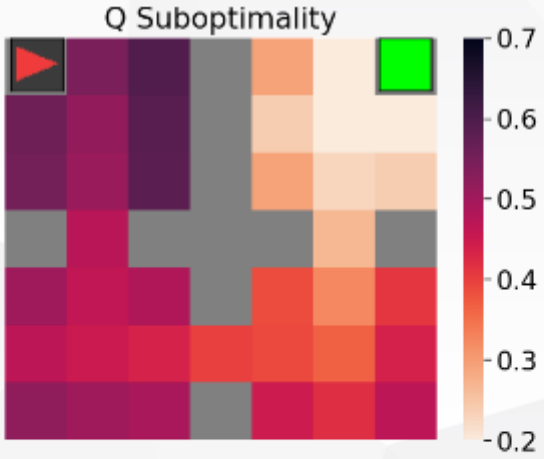


Figure 2: The visualized error of target Q value in a GridWorld Environment. The Q error is visualized by the color of the grid.

ReMERT

Definition 2 (Distance to End) ：给定 MDP $\mathcal{M}$，策略 π 在 $\mathcal{M}$ 中交互得到轨迹 $\tau = \{s_t, a_t\}_{t=0}^{T}$，定义该轨迹中 (s_t, a_t) 的 “到终点的距离” 为

$$h_\tau(s_t, a_t) = T - t$$

作者的直觉表明 $|Q_k - Q^*|$ 的值和 “到终点的距离” 有关

Theorem 2 (Informal). *Under mild conditions, with probability at least* $1-\delta$, *we have*

$$\begin{aligned} &|Q_k(s,a) - Q^*(s,a)| \\ &\leq \mathbb{E}_\tau\left(f(h_\tau^{\pi_k}(s,a))\left(L_{Q_{k-1}} + c\right) + \gamma^{h_\tau^{\pi_k}(s,a)+1}c\right) + g(k,\delta) \end{aligned} \tag{8}$$

where $c = \max_{s,a}\left(Q^*(s,a^*) - Q^*(s,a)\right)$, $f(t) = \frac{\gamma-\gamma^t}{1-\gamma}$, $L_{Q_{k-1}} = \mathbb{E}[|Q_{k-1} - \mathcal{B}^*Q_{k-2}|]$ *and* $g(k,\delta)$ *decreases exponentially as* k *increases.*

$|Q_k - Q^*|$ increases with $h_\tau^{\pi_k}$, and the remaining terms are negligible as k increases.

$$\begin{aligned} |\,Q_k(s,a) - Q^*(s,a)\,| &\approx \mathbb{E}_\tau\,\mathrm{TCE}_c(s,a) \\ &= \mathbb{E}_\tau\left(f\left(h_\tau^{\pi_{k-1}}(s,a)\right)\left(L_{Q_{k-1}} + c\right) + \gamma^{h_\tau^{\pi_{k-1}}(s,a)+1}c\right) \end{aligned}$$

ReMERT

$$w_k(s,a) \propto \frac{d^{\pi_k}(s,a)}{\mu(s,a)} \exp\left(-\mathbb{E}_\tau \operatorname{TCE}_c(s,a)\right)$$

Algorithm 2 ReMERT

1: Initialize Q-values $Q_\theta(s,a)$, a replay buffer μ, and a weight model κ_ψ.
2: **for** step k in $\{1,\ldots,N\}$ **do**
3: Collect M samples using π_k, add them to replay buffer μ, sample $\{(s_i,a_i)\}_{i=1}^N \sim \mu$.
4: Evaluate $Q_\theta(s,a)$ and $\kappa_\psi(s,a)$ on samples (s_i,a_i).
5: Compute target values for Q on samples:
$y_i = r_i + \gamma \max_{a'} Q_{k-1}(s_i',a')$.
$\hat{a}_i = \arg\max_a Q_{k-1}(s_i',a)$.
6: Optimize κ_ψ using

$$L_\kappa(\psi) := \mathbb{E}_{\mathcal{D}_\mathrm{s}}\left[f^*\left(f'\left(\kappa_\psi(s,a)\right)\right)\right] - \mathbb{E}_{\mathcal{D}_\mathrm{f}}\left[f'\left(\kappa_\psi(s,a)\right)\right].$$

7: Compute w_k using

$$w_k(s,a) \propto \frac{d^{\pi_k}(s,a)}{\mu(s,a)} \exp\Big(-\mathbb{E}_{q_{k-1}(\tau)} \operatorname{TCE}_c(s,a)\Big).$$

8: Minimize Bellman error for Q_θ weighted by w_k.
$\theta_{k+1} \leftarrow \operatorname*{argmin}_\theta \frac{1}{N} \sum_i^N w_k\left(s_i,a_i\right)\left(Q_\theta\left(s_i,a_i\right) - y_i\right)^2$.
9: **end for**

Experiments

Deterministic env (MoJoCo & DMC)

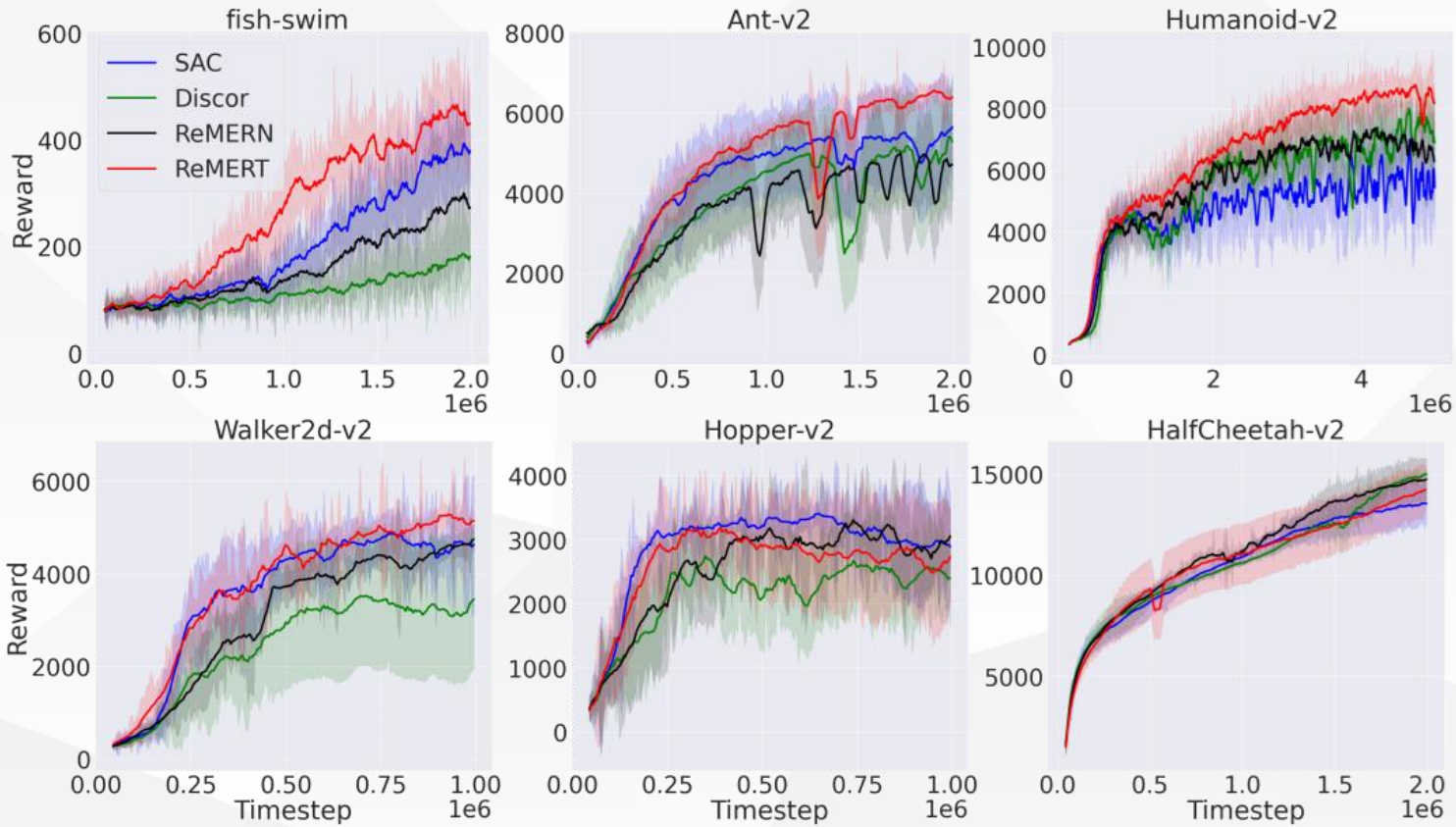
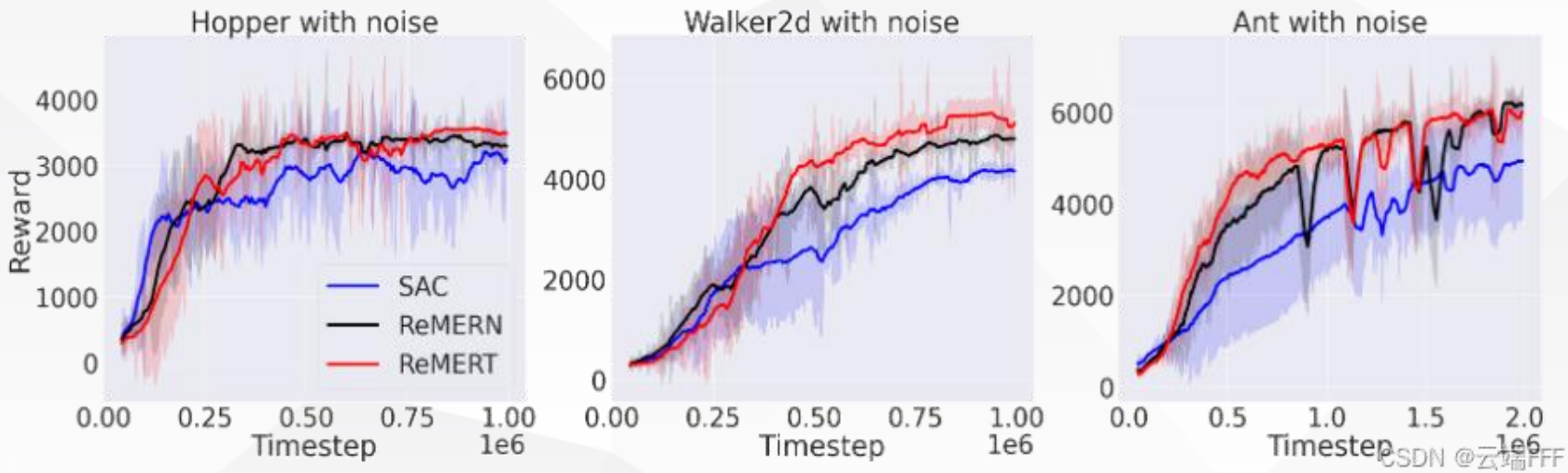


Figure 3: Performance of ReMERT, ReMERN with SAC and DisCor as baselines on continuous control tasks.

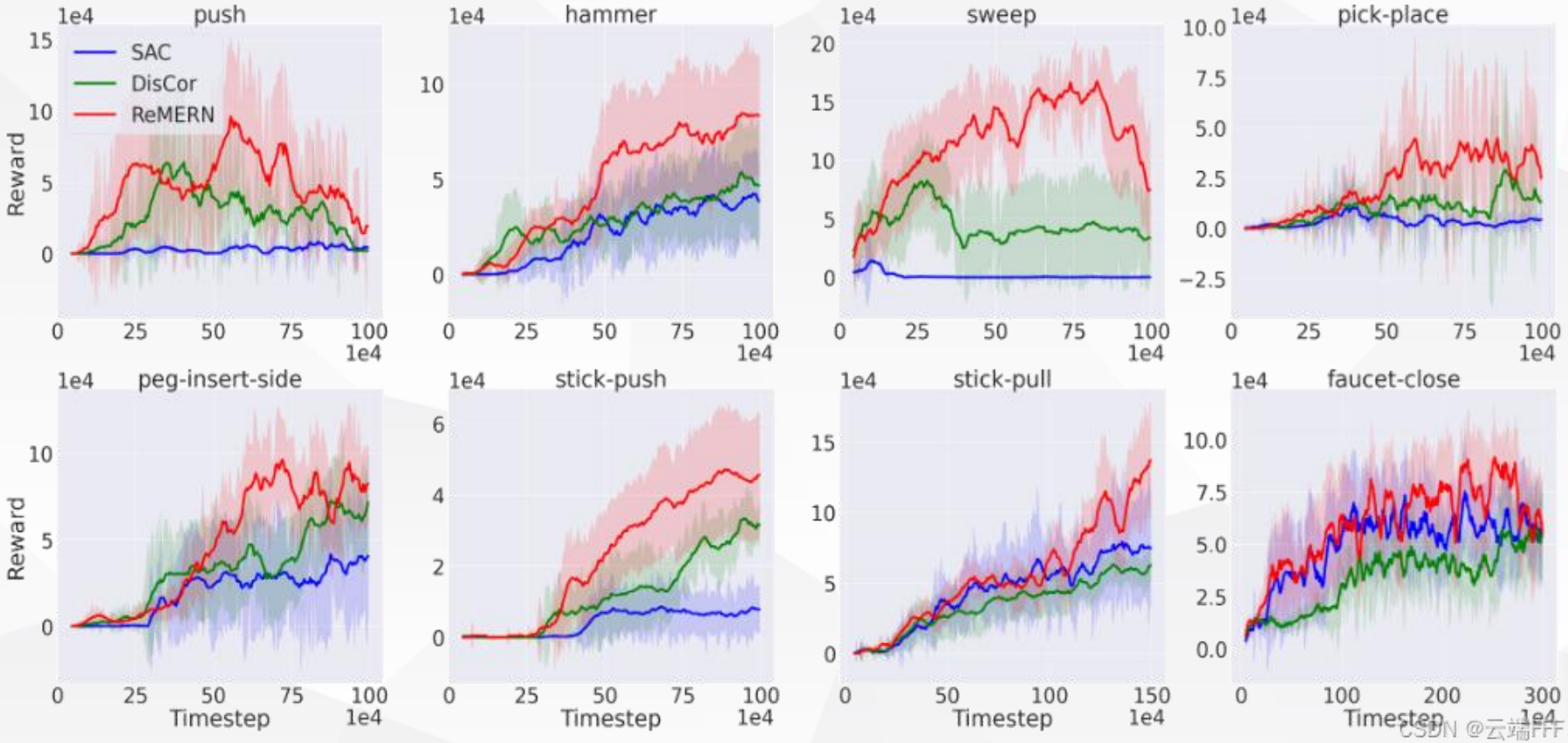
Experiments

MoJoCo with noisy reward



Experiments

Meta-World with high randomness



The End

thanks