Experience Replay with Likelihood-free Importance Weights

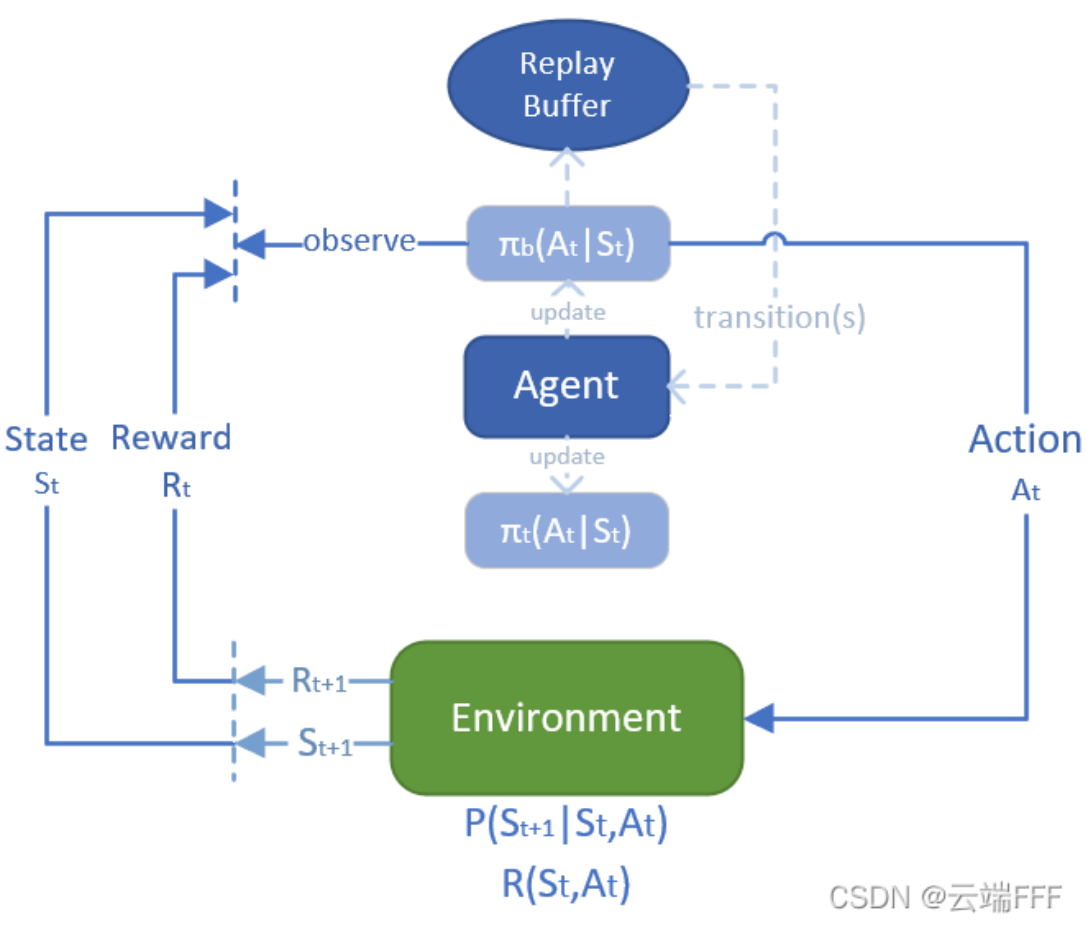
Samarth Sinha*
University of Toronto, Vector Institute
samarth.sinha@mail.utoronto.ca

Jiaming Song*
Stanford University
tsong@cs.stanford.edu

Animesh Garg
University of Toronto, Vector Institute, Nvidia
garg@cs.toronto.edu

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Experience Replay



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

Prioritized Experience Repaly

- 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

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

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

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

Intuition

- In actor-critic methods, the goal is to learn the Q-function induced by the current policy (actor's policy), for a fixed policy, the MDP beacomes a Markov chain

$$d_\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t d_t^\pi(s,a)$$

$$J(\pi) = \mathbb{E}_{d^\pi}[r(s,a)].$$

$$Q^\pi(s,a) := \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) | s_0 = s, a_0 = a]$$

$$\nabla_\phi J(\pi_\phi) = \mathbb{E}_{d^\pi}[\nabla_\phi \log \pi_\phi(a|s) \cdot Q^\pi(s,a)]$$

- In this case, it might be more beneficial to prioritize the correction of (potentially small) TD errors on frequently encountered states, which are more problematic than in low-frequency ones, as they will negatively impact policy updates more severely

Learn the Q-function

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

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

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

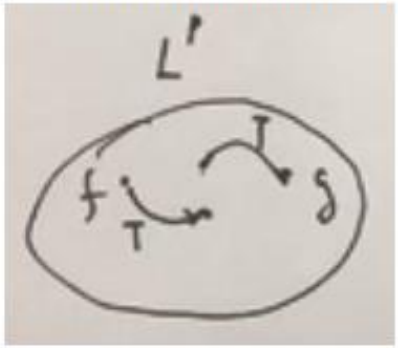
$d^w = d^\pi$

Contraction Mapping

- **收缩映射 Contraction Mapping**：收缩映射 $T: L^p \to L^p$ 是定义在 L_p 空间上的映射，满足 $\forall f, g \in T^p$ 有

$$||T(f) - T(g)||_\rho \leq c||f - g||_\rho, \quad (0 \leq c < 1)$$

其中 $||\cdot||_\rho$ 是 ρ-范数，可以把它看作一种距离度量，**也就是说原先的两个可测函数 f, g 经过收缩映射后距离减小了**



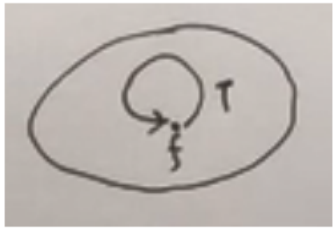
如果其中 T 是微分算子，则称压缩映射 T 是满足 Lipschitz 条件的映射

Contraction Mapping

- **收缩映射定理**：若 T 是 L^p 空间上的收缩映射，则方程

$$(T-I)(f)=0 \Leftrightarrow T(f)=f$$

在 L^p 空间内仅有一个 f 解，称之为 L^p 内 T 的 **不动点**。注意到若 T 是微分算子，则上式为一个常微分方程，因此**收缩映射定理常用于证明常微分方程解的存在性和唯一性。**从几何意义上看，T 将 f 映射回自身



- 压缩映射原理的证明思路如下：

1. 首先**任选** $f_0 \in L^p$，**然后反复使用** T **进行映射得到一个无穷的序列**

$$f_1=T(f_0),\ f_2=T(f_1),\ldots,\ f_n=T(f_{n-1}),\ldots$$

2. 注意到由于来自压缩映射，其中任意相邻两项距离度量越来越近，**即** $\{f\}$ **是一个柯西序列**，由于 L^p 空间具有完备性，该序列必然收敛到 L^p 内部，这**说明不动点** $\lim_{n\to\infty} f_n$ **一定存在**
3. 最后考虑 $T(f_0)$ 是否收敛回 f_0 自身，这只须证明 $\lim_{n\to\infty}||f_n-f_0||=0$ 即可，我们**利用范数的三角不等式，不断向** f_n **和** f_0 **之间插入** f_i，**并结合柯西序列性质进行放缩**，最后即可得证**不动点一定唯一，且为** $\lim_{n\to\infty} f_n=f_0$

Contractive properties of the Bellman operators

- 先考察关于策略 π 的 Bellman 算子 $\mathcal{B}_\pi$，该算子应用于 model-based 的 evaluation 方法 policy evaluation

$$(\mathcal{B}_\pi U)(s) := \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma U(s')]$$

$\forall s, s', s'' \in \mathcal{S}, a \in \mathcal{A}$，对于任意两个价值函数 $U_1(s), U_2(s)$，考察映射后二者距离

$$\begin{aligned}
|(\mathcal{B}_\pi U_1)(s) - (\mathcal{B}_\pi U_2)(s)| &= \left| \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\gamma[U_1(s') - U_2(s')] \right| \\
&\le \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \left| U_1(s') - U_2(s') \right| \\
&\le \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \left(\max_{s''} |U_1(s'') - U_2(s'')| \right) \\
&= \gamma \max_{s''} |U_1(s'') - U_2(s'')| \\
&= \gamma ||U_1 - U_2||_\infty
\end{aligned}$$

注意到对于任意 $s \in \mathcal{S}$ 上式都成立，故对 $s = \arg\max_s |(\mathcal{B}_\pi U_1)(s) - (\mathcal{B}_\pi U_2)(s)|$ 也成立，即有

$$||\mathcal{B}_\pi U_1 - \mathcal{B}_\pi U_2||_\infty \le \gamma ||U_1 - U_2||_\infty$$

因此 Bellman 算子是一个压缩映射，根据收缩映射定理，value evaluation 一定能收敛到唯一的价值函数 $V(s)$ 或 $Q(s,a)$

Contractive properties of the Bellman optimal operators

- 进一步考察 Bellman 最优算子 $\mathcal{B}^*$，该算子应用于 model-based 的 evaluation 方法 value iteration

$$(\mathcal{B}^*U)(s,a) := r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} U(s',a')$$

$\forall s, s', s'' \in \mathcal{S}, a, a', a'_1, a'_2 \in \mathcal{A}$，对于任意两个价值函数 $U_1(s,a), U_2(s,a)$，考察映射后二者距离

$$\begin{aligned}
|(\mathcal{B}^*U_1)(s,a) - (\mathcal{B}^*U_2)(s,a)| &= \left|\gamma \sum_{s'} p(s'|s,a)[\max_{a'_1} U_1(s',a'_1) - \max_{a'_2} U_2(s',a'_2)]\right| \\
&\le \gamma \sum_{s'} p(s'|s,a) \left| \max_{a'_1} U_1(s',a'_1) - \max_{a'_2} U_2(s',a'_2)\right| \\
&\le \gamma \sum_{s'} p(s'|s,a) \left| \max_{a'}(U_1(s',a')) - U_2(s',a')\right| \\
&\le \gamma \sum_{s'} p(s'|s,a) \max_{a'} \left|U_1(s',a') - U_2(s',a')\right| \\
&\le \gamma \max_{s'',a''} |U_1(s'',a'') - U_2(s'',a'')| \\
&= \gamma ||U_1 - U_2||_\infty
\end{aligned}$$

注意到对于任意 $s \in \mathcal{S}, a \in \mathcal{A}$ 上式都成立，故对 $s, a = \arg\max_{s,a} |(\mathcal{B}^*U_1)(s,a) - (\mathcal{B}^*U_2)(s,a)|$ 也成立，即有

$$||\mathcal{B}^*U_1 - \mathcal{B}^*U_2||_\infty \le \gamma ||U_1 - U_2||_\infty$$

因此 Bellman optimal operator 也是一个压缩映射，根据收缩映射定理，value iteration 一定能收敛到唯一的价值函数 $V(s)$ 或 $Q(s,a)$

Policy-dependent Norms

$$\|\mathcal{B}^\pi Q - \mathcal{B}^\pi Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

- The l_∞ norm reflects a distance over two Q functions under the worst possible state-action pair, and is independent of the current policy
- If two Q functions are equal everywhere except for a large difference on a single state-action pair that is unlikely under d^π
  - The l_∞ distance between the two Q functions is large
  - In practice, however, this will have little effect over policy updates as it is unlikely for the current policy to sample the pair
- Our goal with the TD updates is to learn Q^π , a distance metric that is related to π is a more suitable one for comparing different Q functions

Policy-dependent Norms

a distribution over state-action pairs

$$\|Q - Q'\|_d^2 := \mathbb{E}_{(s,a)\sim d}[(Q(s,a) - Q'(s,a))^2]$$

- Policy-dependent Norm
- closely tied to the LQ objective

$$L_Q(\theta; d) = \|Q_\theta(s,a) - \mathcal{B}^\pi Q_\theta(s,a)\|_d^2$$

Theorem 1. *The Bellman operator $\mathcal{B}^\pi$ is a γ-contraction with respect to the $\|\cdot\|_d$ norm if and only if $d = d^\pi$ holds almost everywhere, i.e.,*

$$\|\mathcal{B}^\pi Q - \mathcal{B}^\pi Q'\|_d \leq \gamma \|Q - Q'\|_d, \forall Q, Q' \in \mathcal{Q} \iff \boxed{d = d^\pi}, \quad a.e.$$

Policy-dependent

Theorem 1. *The Bellman operator $\mathcal{B}^\pi$ is a γ-contraction with respect to the $\|\cdot\|_d$ norm if and only if $d = d^\pi$ holds almost everywhere, i.e.,*

$$\|\mathcal{B}^\pi Q - \mathcal{B}^\pi Q'\|_d \leq \gamma \|Q - Q'\|_d, \forall Q, Q' \in \mathcal{Q} \iff d = d^\pi, \quad a.e.$$

Proof. From the definitions of $\|\cdot\|_d$ and $\mathcal{B}^\pi$, we have:

$$\begin{aligned}
&\|\mathcal{B}^\pi Q - \mathcal{B}^\pi Q'\|_d^2 && (13)\\
&= \mathbb{E}_{(s,a)\sim d}[(\gamma \mathbb{E}_{s',a'}[Q(s',a')] - \gamma \mathbb{E}_{s',a'}[Q'(s',a')])^2] \\
&= \gamma^2 \mathbb{E}_{(s,a)\sim d}[(\mathbb{E}_{s',a'}[Q(s',a') - Q'(s',a')])^2] \\
&\leq \gamma^2 \mathbb{E}_{(s,a)\sim d}[\mathbb{E}_{s',a'}[(Q(s',a') - Q'(s',a'))^2]] && (14)\\
&= \gamma^2 \mathbb{E}_{(s,a)\sim d'}[(Q(s,a) - Q'(s,a))^2] && (15)\\
&= \gamma^2 \|Q - Q'\|_{d'}^2 && (16)
\end{aligned}$$

where $s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')$ and

$$d'(s',a') = \sum_{s,a} P(s'|s,a)\pi(a'|s')d(s,a)$$

represents the state-action distribution of the next step when the current distribution is d. We use Jensen's inequality over the convex function $(\cdot)^2$ in Eq. 14. Since d^π is the stationary distribution, $d = d' \iff d = d^\pi, a.e.$, so the if direction holds.

Two challenges

estimate d^π

on-policy

need lots of on-policy samples

increase the sample complexity

off-policy (with replay buffer)

hard to estimate importance ratio

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

Likelihood-free density ratio estimation

- 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

Likelihood-free density ratio estimation

- 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

Pseudo Code

Algorithm 1 Actor Critic with Likelihood-free Importance Weighted Experience Replay

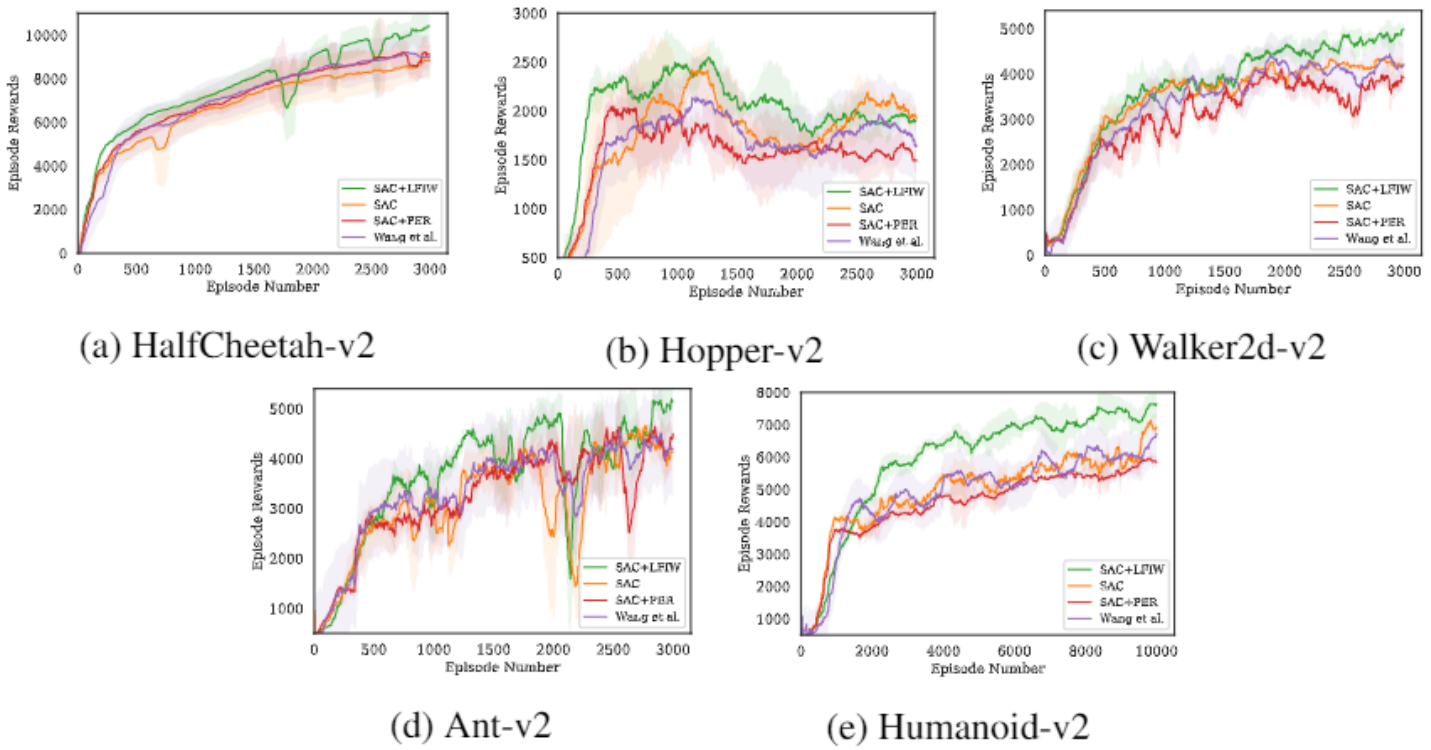
repeat
  for each environment step **do**
    gather new transition tuples (s, a, r, s')
    update (s, a, s, s') to $\mathcal{D}_\mathrm{s}$ (slow replay buffer) and $\mathcal{D}_\mathrm{f}$ (fast replay buffer)
  end for
  remove stale experiences in $\mathcal{D}_\mathrm{s}, \mathcal{D}_\mathrm{f}$ ($|\mathcal{D}_\mathrm{f}| < |\mathcal{D}_\mathrm{s}|$)
  if $|\mathcal{D}_\mathrm{s}|$ exceeds some threshold **then**
    obtain samples from $\mathcal{D}_\mathrm{s}$ and $\mathcal{D}_\mathrm{f}$
    update w_ψ with loss function $L_w(\psi)$ (Eq. 10)
    assign $\tilde{w}_\psi$ according to Eq. 11
  else
    $\tilde{w}_\psi = 1$ (no re-weighting)
  end if
  obtain estimates for $B^\pi Q_\theta$ with base algorithm
  update Q_θ with loss function $L_Q(\theta; \mathcal{D}_\mathrm{s}, \tilde{w})$ (Eq. 12)
  update π_ϕ and value network (if available) with base algorithm
until Stopping criterion
return Q_θ, π_ϕ

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

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

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

Experiments

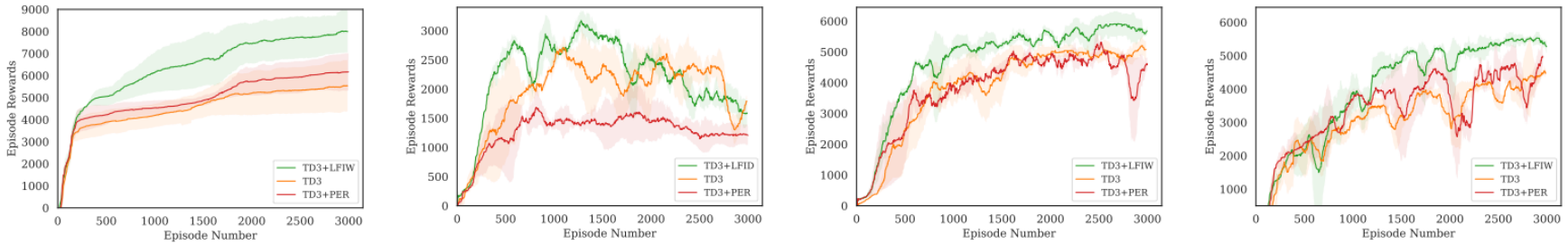


(a) HalfCheetah-v2 (b) Hopper-v2 (c) Walker2d-v2

(d) Ant-v2 (e) Humanoid-v2

Figure 2: Learning curvers for the OpenAI gym continuous control tasks using SAC [15]. The shaded region represents the standard deviation of the average evaluation over 5 trials.

Experiments



(a) HalfCheetah-v2 (b) Hopper-v2 (c) Walker2d-v2 (d) Ant-v2

Figure 3: Learning curvers for the OpenAI gym continuous control tasks using TD3 [13]. The shaded region represents the standard deviation of the average evaluation over 5 trials. We did not include Humanoid as the original TD3 algorithm fails to learn successfully.

Table 1: Max-performance attained by a given environment timestep for the Mujoco control tasks. We report the mean maximum attained performance over 5 random seeds and standrad deviation.

Env	Hopper-v2	Walker-v2	Cheetah-v2	Ant-v2	Humanoid-v2
Timesteps	1M	1M	1M	1M	5M
SAC [15]	2004 ± 356	**3862** ± 106	6548 ± 635	3138 ± 283	5515 ± 329
SAC + PER [32]	1853 ± 106	3210 ± 418	6816 ± 531	2853 ± 132	4650 ± 315
SAC + ERE [42]	1759 ± 234	3601 ± 485	6666 ± 589	3346 ± 116	5586 ± 705
SAC + LFIW	**2395** ± 212	**3855** ± 224	**7037** ± 629	**3857** ± 221	**6436** ± 254
TD3 [13]	2486 ± 125	4212 ± 456	4295 ± 523	2969 ± 202	-
TD3 + PER [32]	1704 ± 228	4268 ± 278	4766 ± 325	3679 ± 156	-
TD3 + LFIW	**3003** ± 261	**5159** ± 189	**6184** ± 876	**3864** ± 205	-

What's more

fitted value iteration algorithm:

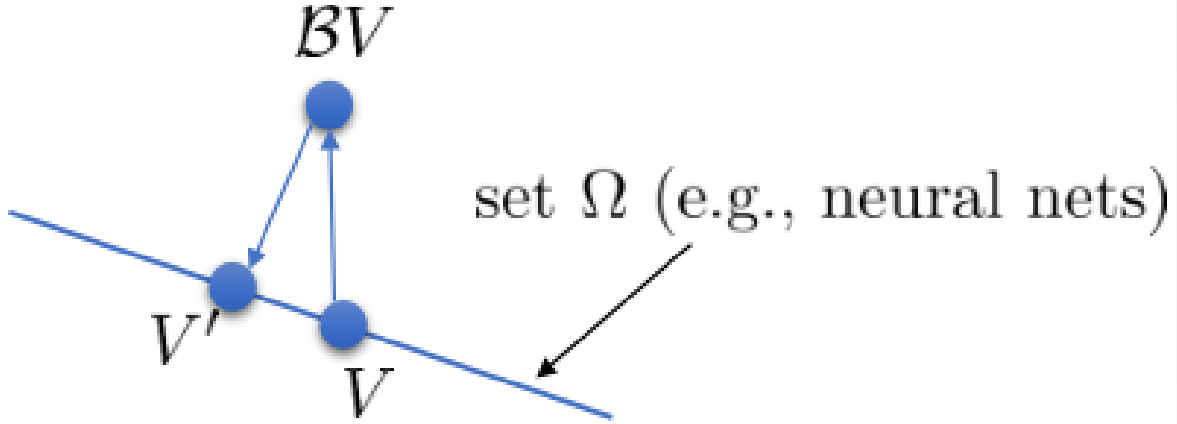
1. set $\mathbf{y}_i \leftarrow \max_{\mathbf{a}_i}(r(\mathbf{s}_i, \mathbf{a}_i) + \gamma E[V_\phi(\mathbf{s}'_i)])$
2. set $\phi \leftarrow \arg\min_\phi \frac{1}{2} \sum_i \|V_\phi(\mathbf{s}_i) - \mathbf{y}_i\|^2$

updated value function

$$V' \leftarrow \arg\min_{V' \in \Omega} \frac{1}{2} \sum \|V'(\mathbf{s}) - (\mathcal{B}V)(\mathbf{s})\|^2$$

all value functions represented by, e.g., neural nets

What's more



fitted value iteration algorithm (using $\mathcal{B}$ and Π):

1. $V \leftarrow \Pi\mathcal{B}V$

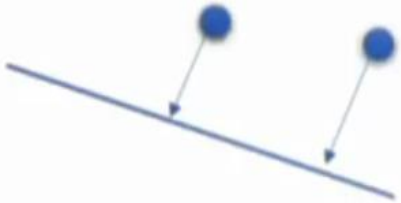
define new operator Π: $\Pi V = \arg\min_{V' \in \Omega} \frac{1}{2} \sum \|V'(\mathbf{s}) - V(\mathbf{s})\|^2$

Π is a *projection* onto Ω (in terms of ℓ_2 norm)

What's more

fitted value iteration algorithm (using $\mathcal{B}$ and Π):

1. $V \leftarrow \Pi\mathcal{B}V$



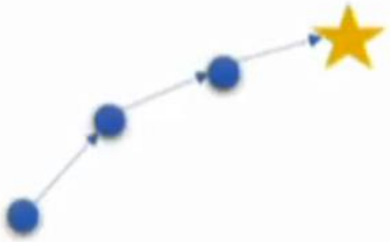
$\mathcal{B}$ is a contraction w.r.t. ∞-norm ("max" norm)

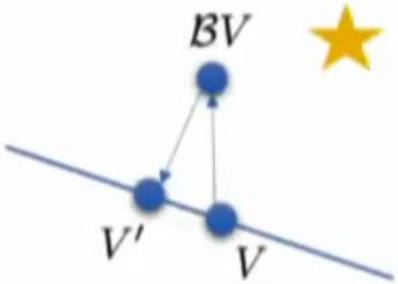
$$\|\mathcal{B}V - \mathcal{B}\bar{V}\|_\infty \leq \gamma\|V - \bar{V}\|_\infty$$

Π is a contraction w.r.t. ℓ_2-norm (Euclidean distance)

$$\|\Pi V - \Pi\bar{V}\|^2 \leq \|V - \bar{V}\|^2$$

but... $\Pi\mathcal{B}$ is not a contraction of any kind





The End

thanks