On the Pessimism in Offline Reinforcement Learning

Xuhui Liu

LAMDA, NJUAI

July 22, 2020

- Have a fixed dataset rather than gains information from its interaction with the environment.
- Performs pure exploitation rather than concerns both exploration and exploitation.

Outline

1. Background
2. Marginalized Behavior Support Algorithms (MBS)
   - Key Ideas
   - Analysis
   - Experiment
3. Expected-Max Q-Learning Operator (EMaQ)
   - Key Ideas
   - Analysis
   - Algorithms

Table of Contents

Related Work

Batch Constrained Q-Learning (BCQ)

$$\pi_\theta(a|s) = \underset{a_i+\xi_\theta(s,a_i)}{\arg\max}\ Q_\phi(s, a_i + \xi_\theta(s, a_i))$$

$$a_i \sim \mu(a|s), i = 1, \cdots, N$$

- Q_ϕ is learned, $\mu(a|s)$ is the behavior policy.
- ξ_θ is an MLP and is bounded with a range $[\Phi, \Phi]$.

Bootstrapping Error Accumulation Reduction (BEAR)

$$\max_{\pi} \mathbb{E}_{s\sim\mathcal{D}}\mathbb{E}_{\pi(a|s)}[\min_{j=1,\cdots,K} Q_j(s,a)]$$

$$\text{s.t.} \quad \mathbb{E}_{s\sim\mathcal{D}}[\text{MMD}(\mu(\cdot|s), \pi(\cdots|s))] \leq \epsilon$$

- Constrain the support of learned policies to match the support of $\mu(a|s)$.

Related Work

- SPIBB
  Follows the behavior policy in less explored state-action pairs while attempting improvement everywhere else.

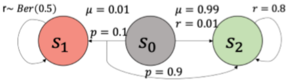
Challenges for Existing Algorithms

- Concentrability coefficient (Algorithm 1)
  Let $\rho_{\mathcal{D}}$ be the state action distribution for dataset $\mathcal{D}$, ρ_π be the distribution for policy π, then the concentrability coefficient is

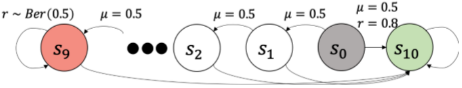
$$C = \left|\left|\frac{\rho_\pi}{\rho_{\mathcal{D}}}\right|\right|_\infty$$

  - Strong assumption
  - Hard to verify
- Hyperparameter (Algorithm 2)
  Hyperparameter is hard to choose in these algorithms.

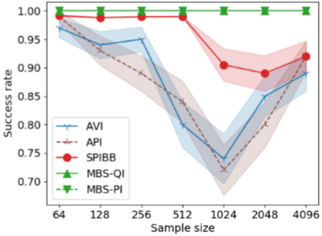
Related Work



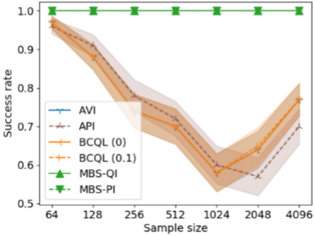
(a) MDP with a rare transition



(b) Combination lock



(a) MDP with a rare transition



(b) 2-arm combination lock

Related Work

- BCQL and BEAR based on just the action probability, even if the state in question itself is less explored.
- In SPIBB, estimating behavior policy is dangerous from rare transitions.

Notation

- Markov Decision Process $M =< \mathcal{S}, \mathcal{A}, P, R, \gamma, \rho >$.
- Policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$.
- The state action distribution of the dataset D is $\mu(s, a)$, and state distribution $\mu(s) = \sum_{a \in \mathcal{A}} \mu(s, a)$.
- Value function $V^\pi(s) = \mathbb{E}[\sum_{h=0}^{\infty} \gamma^h r_h]$ and Q-function $Q^\pi(s, a)$.
- v^π is the expectation of $V^\pi(s)$ under initial state distribution.
- The function we aim to fit a Q-function:

$$f : \mathcal{S} \times \mathcal{A} \rightarrow [0, V_{\max}]$$

.

Notation

- Bellman optimality operators $\mathcal{T}$

$$(\mathcal{T}f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'}[\max_{a'} f(s',a')].$$

- Bellman evaluation operators $\tilde{\mathcal{T}}$:

$$(\mathcal{T}^{\pi}f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'}\mathbb{E}_{a'\sim\pi} f(s',a').$$

- Empirical Bellman optimality/evaluation operators $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}}^{\pi}$.

Table of Contents

Algorithms

- We assume we have a density function $\hat{\mu}$ which is an approximate estimate of μ.
- Given $\hat{\mu}$ and a threshold b we define a filter function:

$$\zeta(s, a; \hat{\mu}, b) = \mathbb{I}(\hat{\mu}(s, a) \geq b).$$

- Write $\zeta(s, a; \hat{\mu}, b)$ as $\zeta(s, a)$ and define $\zeta \circ f(s, a) := \zeta(s, a) f(s, a)$.
- Define $\zeta - \textit{constrained Bellman evaluation operator } \tilde{\mathcal{T}}^{\pi}$ as

$$(\tilde{\mathcal{T}}^{\pi}) f(s, a) = r(s, a) + \gamma \mathbb{E}_{s'} \sum_{a' \in \mathcal{A}} [\pi(a'|s') \zeta \circ f(s', a')].$$

Algorithms

- Empirical loss of f given f' and policy π:

$$\mathcal{L}_D(f; f', \pi) := \mathbb{E}_D\Big(f(s,a) - r - \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s')\zeta \circ f'(s', a')\Big)^2.$$

- Similarly, for Bellman optimality operator

$$(\tilde{\mathcal{T}} f)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'}[\max_{a'} \zeta \circ f(s', a')].$$

$$\mathcal{L}_D(f; f') := \mathbb{E}_D\Big(f(s,a) - r - \gamma \max_{a' \in \mathcal{A}} \zeta \circ f'(s', a')\Big)^2.$$

Algorithms

Algorithm 1 MBS Policy Iteration (MBS-PI)

Input: D, $\mathcal{F}$, Π, $\widehat{\mu}$, b
Output: $\widehat{\pi}_T$
for $t = 0$ **to** $T - 1$ **do**
  for $k = 0$ **to** K **do**
    $f_{t,k+1} \leftarrow \arg\min_{f \in \mathcal{F}} \mathcal{L}_D(f, f_{t,k}; \widehat{\pi}_t)$
  end for
  $\widehat{\pi}_{t+1} \leftarrow \arg\max_{\pi \in \Pi} \mathbb{E}_D[\mathbb{E}_\pi [\zeta \circ f_{t,K+1}]]$
end for

Algorithm 2 MBS Q Iteration (MBS-QI)

Input: D, $\mathcal{F}$, $\widehat{\mu}$, b
Output: $\widehat{\pi}_T$
for $t = 0$ **to** $T - 1$ **do**
  $f_{t+1} \leftarrow \arg\min_{f \in \mathcal{F}} \mathcal{L}_D(f; f_t)$
  $\widehat{\pi}_{t+1}(s) \leftarrow \arg\max_{a \in \mathcal{A}} \zeta \circ f_{t+1}(s, a)$
end for

Assumption

Let $\eta_h^\pi(s) := \Pr[s_h = s|\pi]$, $\eta_h^\pi(s,a) = \eta_h^\pi(s)\pi(a|s)$, and
$\eta^\pi(s,a) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\eta_h^\pi(s,a)$.

Assumption 1 (Bounded densities). *For any non-stationary policy* π *and* $h \geq 0$, $\eta_h^\pi(s,a) \leq U$.

Assumption 2 (Density estimation error). *With probability at least* $1-\delta$, $\|\widehat{\mu} - \mu\|_{TV} \leq \epsilon_\mu$.

Assumption 3 (Completeness under $\widetilde{\mathcal{T}}^\pi$). $\forall \pi \in \Pi$, $\max_{f\in\mathcal{F}} \min_{g\in\mathcal{F}} \|g - \widetilde{\mathcal{T}}^\pi f\|_{2,\mu}^2 \leq \epsilon_{\mathcal{F}}$.

Assumption 4 (Π Completeness). $\forall f \in \mathcal{F}$, $\min_{\pi\in\Pi} \|\mathbb{E}_\pi[\zeta \circ f(s,a)] - \max_a \zeta \circ f(s,a)\|_{1,\mu} \leq \epsilon_\Pi$.

Main Result

Definition 1 (ζ-constrained policy set). *Let Π_C^{all} be the set of policies $\mathcal{S} \to \Delta(\mathcal{A})$ such that $\Pr(\zeta(s,a) = 0|\pi) \leq \epsilon_\zeta$. That is, $\mathbb{E}_{s,a\sim\eta^\pi}\left[\mathbb{1}\left(\zeta(s,a)=0\right)\right] \leq \epsilon_\zeta$.*

Theorem 1. *Given an MDP M, a dataset $D = \{(s,a,r,s')\}$ with n samples drawn i.i.d. from $\mu \times R \times P$, and a Q-function class $\mathcal{F}$ and a policy class Π satisfying Assumption 3 and 4, $\widehat{\pi}_t$ from Algorithm 1 satisfies that w. p. at least $1-3\delta$,*

$$v_M^{\widetilde{\pi}} - v_M^{\widehat{\pi}_t} \leq \mathcal{O}\left(\frac{C\sqrt{V_{\max}^2 \ln(|\mathcal{F}||\Pi|/\delta)}}{(1-\gamma)^3\sqrt{n}}\right) + \frac{8C\sqrt{\epsilon_\mathcal{F}} + 6CV_{\max}\epsilon_\mu}{(1-\gamma)^3} + \frac{2C\epsilon_\Pi + 3\gamma^{K-1}V_{\max}}{(1-\gamma)^2} + \frac{V_{\max}\epsilon_\zeta}{1-\gamma},$$

for any policy $\widetilde{\pi} \in \Pi_C^{all}$ under Assumptions 1 and 2 and any $t \geq K$. $C = U/b$. K is the number of policy evaluation iterations (inner loop) and t is the number of policy improvement steps.

Proof Sketch

- Define an auxiliary MDP $M' =< \mathcal{S}', \mathcal{A}', R', P', \gamma, \rho >$, where $\mathcal{S}' = \mathcal{S} \cup \{s_{abs}\}$, $\mathcal{A}' = \mathcal{A} \cup \{a_{abs}\}$.
- $R'(s_{abs}, a_{abs}) = 0$, $P'(s_{abs}, a) = s_{abs}$ and $P'(s, a_{abs}) = s_{abs}$.
- Define $(\Xi\pi)(a|s) = \zeta(s,a)\pi(a|s)$ if $a \in \mathcal{A}$, $(\Xi\pi(a|s) = \sum_{a' \in \mathcal{A}'} \pi(a'|s)(1 - \xi(s,a'))$ if $a = a_{abs}$.
- Then $Q^{\Xi(\pi)}$ is the fixed point of $\tilde{\mathcal{T}}^{\pi}$.
- For any policy $\pi \in \Pi_C^{all}$, $v_M^{\pi} \leq v_{M'}^{\Xi(\pi)} + \frac{\epsilon_\zeta V_{\max}}{1-\gamma}$.

Lemma Proof

Lemma 6. *For any policy $\pi : \mathcal{S}' \to \Delta(\mathcal{A}')$, the fixed point solution of $\widetilde{\mathcal{T}}^{\pi}$ is equal to $Q^{\Xi(\pi)}$ on $\mathcal{S} \times \mathcal{A}$.*

Proof. By definition $Q^{\Xi(\pi)}$ is the fixed point of the standard Bellman evaluation operator on M': $\mathcal{T}_{M'}^{\Xi(\pi)}$. So for any $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$
\begin{aligned}
& Q^{\Xi(\pi)}(s, a) && (25)\\
&= (\mathcal{T}_{M'}^{\Xi(\pi)} Q^{\Xi(\pi)})(s, a) && (26)\\
&= r(s, a) + \gamma \mathbb{E}_{s'} \left[\sum_{a' \in \mathcal{A}'} \Xi(\pi)(a'|s') Q^{\Xi(\pi)}(s', a') \right] && (27)\\
&= r(s, a) + \gamma \mathbb{E}_{s'} \left[\Xi(\pi)(a_{\text{abs}}|s') Q^{\Xi(\pi)}(s', a_{\text{abs}}) + \sum_{a' \in \mathcal{A}} \Xi(\pi)(a'|s') Q^{\Xi(\pi)}(s', a') \right] && (28)\\
&= r(s, a) + \gamma \mathbb{E}_{s'} \left[\sum_{a' \in \mathcal{A}} \Xi(\pi)(a'|s') Q^{\Xi(\pi)}(s', a') \right] && (29)\\
&= r(s, a) + \gamma \mathbb{E}_{s'} \left[\sum_{a' \in \mathcal{A}} \pi(a'|s') \zeta(s', a') Q^{\Xi(\pi)}(s', a') \right] && (30)\\
&= (\widetilde{\mathcal{T}}^{\pi} Q^{\Xi(\pi)})(s, a) && (31)
\end{aligned}
$$

So we proved that $Q^{\Xi(\pi)}$ is also the fixed-point solution of $\widetilde{\mathcal{T}}^{\pi}$ constrained on $\mathcal{S} \times \mathcal{A}$. □

Lemma 3. *For any policy* $\pi \in \Pi_C^{all}$, $v_M^\pi \leq v_{M'}^{\Xi(\pi)} + \frac{\epsilon_\zeta V_{\max}}{1-\gamma}$

Proof. Since π only takes action in $\mathcal{A}$, by Lemma 1, we have that $v_M^\pi = v_{M'}^\pi$. Since $\pi \in \Pi_C^{all}$, we have that $\Pr(\zeta(s,a)=0|\pi) \leq \epsilon_\zeta$, which means that:

$$(1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s\sim\eta_h^\pi}\left[\mathbb{1}\left(\zeta(s,a)=0\right)\right] \leq \epsilon_\zeta \tag{9}$$

Thus:

$$\begin{aligned}
v^{\Xi(\pi)} - v^\pi &= \sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s\sim\eta_h^\pi}\left[V^{\Xi(\pi)}(s) - \sum_{a\in\mathcal{A}'}\pi(a|s)Q^{\Xi(\pi)}(s,a)\right] && (10)\\
&= \sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s\sim\eta_h^\pi}\left[V^{\Xi(\pi)}(s) - \sum_{a\in\mathcal{A}'}\pi(a|s)\zeta(s,a)Q^{\Xi(\pi)}(s,a)\right] && (11)\\
&\quad - \sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s,a\sim\eta_h^\pi}\left[\mathbb{1}\left(\zeta(s,a)=0\right)Q^{\Xi(\pi)}(s,a)\right] && (12)\\
&\geq \sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s\sim\eta_h^\pi}\left[V^{\Xi(\pi)}(s) - \sum_{a\in\mathcal{A}'}\pi(a|s)\zeta(s,a)Q^{\Xi(\pi)}(s,a)\right] && (13)\\
&\quad - V_{\max}\sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s,a\sim\eta_h^\pi}\left[\mathbb{1}\left(\zeta(s,a)=0\right)\right] && (14)\\
&\geq \sum_{h=0}^{\infty}\gamma^h \mathbb{E}_{s\sim\eta_h^\pi}\left[V^{\Xi(\pi)}(s) - \sum \pi(a|s)\zeta(s,a)Q^{\Xi(\pi)}(s,a)\right] - \frac{V_{\max}\epsilon_\zeta}{1-\gamma} && (15)
\end{aligned}$$

Analysis

Corollary 1. *If there exists an* $\pi^\star$ *on* M *such that* $\Pr(\mu(s,a) \le 2b|\pi^\star) \le \epsilon$. *then under the assumptions of Theorem 1,* $\widehat{\pi}_t$ *from Algorithm 1 satisfies that with probability at least* $1-3\delta$,

$$v_M^{\pi^\star} - v_M^{\pi_t} \le \frac{4C}{(1-\gamma)^3}\left(\sqrt{\frac{419V_{\max}^2 \ln\frac{|\mathcal{F}||\Pi|}{\delta}}{3n}} + 2\sqrt{\epsilon_{\mathcal{F}}}\right) + \frac{6CV_{\max}\epsilon_\mu}{(1-\gamma)^3} + \frac{2C\epsilon_\Pi + 3\gamma^{K-1}V_{\max}}{(1-\gamma)^2} + \frac{V_{\max}(\epsilon + C\epsilon_\mu)}{1-\gamma}$$

Proof. Given the condition of $\pi^\star$,

$$\begin{aligned}
\Pr\left(\widehat{\mu}(s,a) \le b \middle| \pi^\star\right) &\le \Pr\left(\mu(s,a) \le 2b|\pi^\star\right) + \Pr\left(|\mu(s,a) - \widehat{\mu}(s,a)| \ge b|\pi^\star\right) && (93)\\
&\le \epsilon + \Pr\left(|\mu(s,a) - \widehat{\mu}(s,a)| \ge b|\pi^\star\right) && (94)\\
&\le \epsilon + \frac{\mathbb{E}_{\eta^{\pi^\star}}\left[|\mu(s,a) - \widehat{\mu}(s,a)|\right]}{b} && (95)\\
&\le \epsilon + \frac{U d_{\mathrm{TV}}(\mu(s,a), \widehat{\mu}(s,a))}{b} && (96)\\
&\le \epsilon + C\epsilon_\mu && (97)
\end{aligned}$$

Then $\pi^\star \in \Pi_C^{all}$ with $\epsilon_\zeta = \epsilon + C\epsilon_\mu$, and applying Theorem 1 finished the proof. □

In many scenarios we aim to have a policy improvement that is guaranteed to be no worse than the data collection policy, which is called safe policy improvement.
When the state-action space is finite, there must exist an minimum value for all non-zero $\mu(s,a)$'s. Let $\mu_{\min} = \min_{s,a,s.t.\mu(s,a)>0} \mu(s,a)$.

Corollary 2 (Safe policy improvement – discrete state space). *For finite state action spaces and $b \leq \mu_{\min}$, under the same assumptions as Theorem 1, there exist function sets $\mathcal{F}$ and Π (specified in the proof) such that $\widehat{\pi}_t$ from Algorithm 1 satisfies that with probability at least $1-3\delta$,*

$$v_M^{\hat{\pi}_t} \geq v_M^{\mu} - \widetilde{\mathcal{O}}\left(\frac{V_{\max}}{b(1-\gamma)^3}\left(\frac{|\mathcal{S}||\mathcal{A}|}{n} + \sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{n}}\right) + \frac{\gamma^K V_{\max}}{(1-\gamma)^2}\right)$$

Experiment

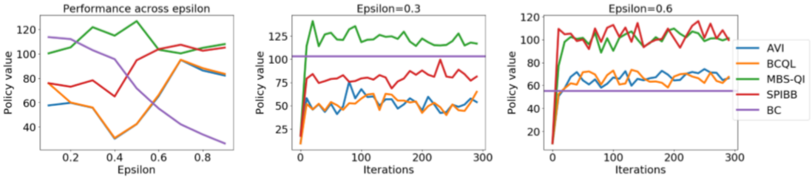


Figure 3: CartPole-v0. Left: convergent policy value across different (ϵ-greedy) behavior policies. Middle and Right: learning curves when $\epsilon = 0.3, 0.6$. We allow non-zero threshold for BCQL to subsume the tabular algorithm of BEAR [17].

Optimization Algorithms

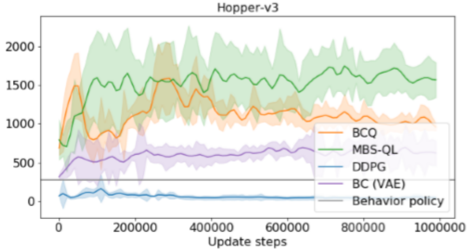


Figure 4: MuJoCo Hopper-v3 domain. Averaged over 5 random seeds and the shadow region in plot shows the standard deviation.

Table of Contents

Expected Max-Q Operator

- Q-Evaluation for policy μ

$$\mathcal{T}_\mu Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s'} \mathbb{E}_{a' \sim \mu}[Q(s',a')]$$

- Q-Learning

$$\mathcal{T}^* Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s'}[\max_{a'} Q(s',a')]$$

- Expected Max-Q Operator

$$\mathcal{T}_\mu^N Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s'} \mathbb{E}_{\{a'\}^N \sim \mu}[\max_{a'' \in \{a'\}^N} Q(s',a'')]$$

Theorem 3.1. *In the tabular setting, for any $N \in \mathbb{N}$, $\mathcal{T}_{\mu}^{N}$ is a contraction operator in the $\mathcal{L}_{\infty}$ norm. Hence, with repeated applications of the $\mathcal{T}_{\mu}^{N}$, any initial Q function converges to a unique fixed point.*

Theorem 3.2. *Let Q_{μ}^{N} denote the unique fixed point achieved in Theorem 3.1, and let $\pi_{\mu}^{N}(a|s)$ denote the policy that samples N actions from $\mu(a|s)$, $\{a_i\}^N$, and chooses the action with the maximum Q_{μ}^{N}. Then Q_{μ}^{N} is the Q-value function corresponding to $\pi_{\mu}^{N}(a|s)$.*

Theorem 3.3. *Let π^*_μ denote the optimal policy from the class of policies whose actions are restricted to lie within the support of the policy $\mu(a|s)$. Let Q^*_μ denote the Q-value function corresponding to π^*_μ. Furthermore, let Q_μ denote the Q-value function of the policy $\mu(a|s)$. Let $\mu^*(s) := \int_{Support(\pi^*_\mu(a|s))} \mu(a|s)$ denote the probability of optimal actions under $\mu(a|s)$. Under the assumption that $\inf_s \mu^*(s) > 0$ and $r(s,a)$ is bounded, we have that,*

$$Q^1_\mu = Q_\mu \qquad \text{and} \qquad \lim_{N\to\infty} Q^N_\mu = Q^*_\mu$$

Theorem 3.4. *For all $N, M \in \mathbb{N}$, where $N > M$, we have that $\forall s \in \mathcal{S}, \forall a \in \text{Support}(\mu(\cdot|s))$, $Q^N_\mu(s,a) \geq Q^M_\mu(s,a)$. Hence, $\pi^N_\mu(a|s)$ is at least as good of a policy as $\pi^M_\mu(a|s)$.*

Theorem 3.5. *For $s \in \mathcal{S}$ let,*

$$\Delta(s) = \max_{a \in \text{Support}(\mu(\cdot|s))} Q^*_\mu(s,a) - \mathbb{E}_{\{a_i\}^N \sim \mu(\cdot|s)}[\max_{b \in \{a_i\}^N} Q^*_\mu(s,b)]$$

The suboptimality of Q^N_μ can be upperbounded as follows,

$$\left\| Q^N_\mu - Q^*_\mu \right\|_\infty \leq \frac{\gamma}{1-\gamma} \max_{s,a} \mathbb{E}_{s'} \left[\Delta(s') \right] \leq \frac{\gamma}{1-\gamma} \max_{s} \Delta(s) \tag{2}$$

The same also holds when Q^*_μ is replaced with Q^N_μ in the definition of Δ.

Analysis

Algorithm 1: Full EMaQ Training Algorithm

Offline dataset $\mathcal{D}$, Pretrain $\mu(a|s)$ on $\mathcal{D}$

Initialize K Q functions with parameters θ_i, and K target Q functions with parameters θ_i^{target}

`Ensemble` parameter λ, Exponential moving average parameter α

Function `Ensemble(`*values*`):`
- **return** $\lambda \cdot \min(values) + (1-\lambda) \cdot \max(values)$

Function $y_{\text{target}}(s, a, s', r, t)$:
- $\{a'_i\}^N \sim \mu(a'|s')$
- $Qvalues \leftarrow [\ \ ]$
- **for** $k \leftarrow 1$ **to** N **do**
  - `/* Estimate the value of action` a'_k `*/`
  - $Qvalues$.append$\Big($`Ensemble`$\big([Q_i^{target}(s', a'_k)$ for all $i]\big)\Big)$
- **return** $r + (1-t) \cdot \gamma \max(Qvalues)$

while *not converged* **do**
- Sample a batch $\{(s_m, a_m, s'_m, r_m, t_m)\}^M \sim \mathcal{D}$
- **for** $i = 1, ..., K$ **do**
  - $\mathcal{L}(\theta_i) = \sum_m \Big(Q_i(s_m, a_m) - y_{\text{target}}(s_m, a_m, s'_m, r_m, t_m)\Big)^2$
  - $\theta_i \leftarrow \theta_i -$ `AdamUpdate`$\Big(\mathcal{L}(\theta_i), \theta_i\Big)$
  - $\theta_i^{\text{target}} \leftarrow \alpha \cdot \theta_i^{\text{target}} + (1-\alpha) \cdot \theta_i$

Thanks!