Supplementary Material for "Action Candidate Based Clipped Double Q-learning for Discrete and Continuous Action Tasks"

A. Proof of Theorems on Proposed Estimator

Property 1. *Let a^*_K be the index that maximizes $\hat{\boldsymbol{\mu}}^A$ among $\boldsymbol{\mathcal{M}}_K$: $\hat{\mu}^A_{a^*_K} = \max_{i\in\boldsymbol{\mathcal{M}}_K} \hat{\mu}^A_i$. Then, as the number K decreases, the underestimation decays monotonically, that is $\mathbb{E}\left[\hat{\mu}^B_{a^*_K}\right] \geq \mathbb{E}\left[\hat{\mu}^B_{a^*_{K+1}}\right]$, $1 \leq K < N$.*

Proof. Suppose that $\boldsymbol{\mathcal{M}}_K = \{a_{(1)}, \ldots, a_{(K)}\}$ for $\hat{\mu}^B_{a^*_K}$, where $a_{(i)}$ denotes the index corresponding to the i-th largest value in $\hat{\mu}^B$ and $a^*_K = \arg\max_{j\in\boldsymbol{\mathcal{M}}_K} \hat{\mu}^A_j$. Then, $\boldsymbol{\mathcal{M}}_{K+1} = \boldsymbol{\mathcal{M}}_K \cup \{a_{(K+1)}\}$ for $\hat{\mu}^B_{a^*_{K+1}}$. If $\hat{\mu}^A_{a_{(K+1)}} > \hat{\mu}^A_{a^*_K}$, then $a^*_{K+1} = a_{(K+1)}$. Due to $a_{(K+1)} \notin \boldsymbol{\mathcal{M}}_K$ and $a^*_K \in \boldsymbol{\mathcal{M}}_K$, $\hat{\mu}^B_{a_{(K+1)}} = \hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K}$. Similarly, if $\hat{\mu}^A_{a_{(K+1)}} < \hat{\mu}^A_{a^*_K}$, then a^*_{K+1} is equal to a^*_K. Thus, $\hat{\mu}^B_{a^*_{K+1}} = \hat{\mu}^B_{a^*_K}$. Finally, if $\hat{\mu}^A_{a_{(K+1)}} = \hat{\mu}^A_{a^*_K}$, a^*_{K+1} is either equal to a^*_K or equal to $a_{(K+1)}$. For the former, the estimation value under $K+1$ remain unchanged, that is $\hat{\mu}^B_{a^*_{K+1}} = \hat{\mu}^B_{a^*_K}$. For the latter, $\hat{\mu}^B_{a^*_{K+1}} = \hat{\mu}^B_{a_{(K+1)}} <= \hat{\mu}^B_{a^*_K}$ where the equal sign is established only when there are multiple K-th largest values and $\hat{\mu}^B_{a_{(K)}} = \hat{\mu}^B_{a^*_K}$. Therefore, we can obtain $\hat{\mu}^B_{a^*_{K+1}} \leq \hat{\mu}^B_{a^*_K}$ and $\mathbb{E}\left[\hat{\mu}^B_{a^*_{K+1}}\right] \leq \mathbb{E}\left[\hat{\mu}^B_{a^*_K}\right]$. The inequality is strict if and only if $P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K}\right) > 0$. □

Theorem 1. *As the number K decreases, the underestimation decays monotonically, that is $\mathbb{E}\left[\min\left\{\hat{\mu}^B_{a^*_K}, \hat{\mu}^*_{SE}\right\}\right] \geq \mathbb{E}\left[\min\left\{\hat{\mu}^B_{a^*_{K+1}}, \hat{\mu}^*_{SE}\right\}\right]$, $1 \leq K < N$, where the inequality is strict if and only if $P\left(\hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_K} > \hat{\mu}^B_{a^*_{K+1}}\right) > 0$ or $P\left(\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right) > 0$. Moreover, $\forall K : 1 \leq K \leq N$, $\mathbb{E}\left[\min\left\{\hat{\mu}^B_{a^*_K}, \hat{\mu}^*_{SE}\right\}\right] \geq \mathbb{E}\left[\hat{\mu}^*_{CDE}\right]$.*

Proof. For simplicity, we set $G(K) = \min\left\{\hat{\mu}^B_{a^*_K}, \hat{\mu}^*_{SE}\right\} - \min\left\{\hat{\mu}^B_{a^*_{K+1}}, \hat{\mu}^*_{SE}\right\}$. First, we have

$$\mathbb{E}\left[G\left(K\right)\right] = P\left(\hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\mathbb{E}\left[G\left(K\right) \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right] + P\left(\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right)\mathbb{E}\left[G\left(K\right) \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right]. \tag{1}$$

Then, from Property 1, we have $\hat{\mu}^B_{a^*_{K+1}} \leq \hat{\mu}^B_{a^*_K}$, $1 \leq K < N$. Hence, the expected value of $G(K)$ under the condition $\hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}$ can be obtained as below:

$$\begin{aligned}
&\mathbb{E}\left[G\left(K\right) \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right] = \mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right] \\
=&P\left(\hat{\mu}^B_{a^*_K} > \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\underbrace{\mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right]}_{>0} \\
&+ P\left(\hat{\mu}^B_{a^*_K} = \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\underbrace{\mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_{K+1}} = \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right]}_{=0} \\
=&P\left(\hat{\mu}^B_{a^*_K} > \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\underbrace{\mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right]}_{>0}.
\end{aligned} \tag{2}$$

Thus, the first item in Eq. 1 can be rewritten as below:

$$\begin{aligned}
&P\left(\hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\mathbb{E}\left[G\left(K\right) \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right] \\
=&P\left(\hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)P\left(\hat{\mu}^B_{a^*_K} > \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right] \\
=&P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right)\mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right]
\end{aligned} \tag{3}$$

Moreover, the expected value of $G(K)$ under the condition $\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}$ under the condition $\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}$ can be obtained as below:

$$\begin{aligned}
&\mathbb{E}\left[G\left(K\right) \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right] = \mathbb{E}\left[\hat{\mu}^*_{SE} - \min\left\{\hat{\mu}^B_{a^*_{K+1}}, \hat{\mu}^*_{SE}\right\} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right] \\
=&P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^*_{SE} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right) \underbrace{\mathbb{E}\left[\hat{\mu}^*_{SE} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right]}_{>0} \\
&+ P\left(\hat{\mu}^B_{a^*_{K+1}} \geq \hat{\mu}^*_{SE} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right) \underbrace{\mathbb{E}\left[\hat{\mu}^*_{SE} - \hat{\mu}^*_{SE} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^B_{a^*_{K+1}} \geq \hat{\mu}^*_{SE}\right]}_{=0} \\
=&P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^*_{SE} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right) \mathbb{E}\left[\hat{\mu}^*_{SE} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right].
\end{aligned} \tag{4}$$

Therefore, the second item in Eq. 1 can be rewritten as below:

$$\begin{aligned}
&P\left(\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right) \mathbb{E}\left[G\left(K\right) \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right] \\
=&P\left(\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right) P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^*_{SE} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE}\right) \mathbb{E}\left[\hat{\mu}^*_{SE} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right] \\
=&P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^*_{SE} \leq \hat{\mu}^B_{a^*_K}\right) \mathbb{E}\left[\hat{\mu}^*_{SE} - \hat{\mu}^B_{a^*_{K+1}} \mid \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right].
\end{aligned} \tag{5}$$

Finally, the expected value of $G(K)$ can be expressed as:

$$\begin{aligned}
\mathbb{E}\left[G\left(K\right)\right] =&P\left(\hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right) \underbrace{\mathbb{E}\left[\hat{\mu}^B_{a^*_K} - \hat{\mu}^B_{a^*_{K+1}} | \hat{\mu}^B_{a^*_{K+1}} < \hat{\mu}^B_{a^*_K} < \hat{\mu}^*_{SE}\right]}_{>0} \\
&+ P\left(\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right) \underbrace{\mathbb{E}\left[\hat{\mu}^*_{SE} - \hat{\mu}^B_{a^*_{K+1}} | \hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right]}_{>0} \geq 0,
\end{aligned} \tag{6}$$

where the inequality is strict if and only if $P\left(\hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_K} > \hat{\mu}^B_{a^*_{K+1}}\right) > 0$ or $P\left(\hat{\mu}^B_{a^*_K} \geq \hat{\mu}^*_{SE} > \hat{\mu}^B_{a^*_{K+1}}\right) > 0$.

Further, due to the monotonicity of the expected value of $\min\left\{\hat{\mu}^B_{a^*_K}, \hat{\mu}^*_{SE}\right\}$ with regard to K ($1 \leq K \leq N$), we can know that the minimum value is at $K = N$, that is $\mathbb{E}\left[\min\left\{\hat{\mu}^B_{a^*_K}, \hat{\mu}^*_{SE}\right\}\right] \geq \mathbb{E}\left[\min\left\{\hat{\mu}^B_{a^*_N}, \hat{\mu}^*_{SE}\right\}\right]$. Since $\hat{\mu}^B_{a^*_N}$ means that we choose the index corresponding to the largest value in $\hat{\boldsymbol{\mu}}^A$ among the all indexs, which is equal to the double estimator, we can further have $\hat{\mu}^B_{a^*_N} = \hat{\mu}^*_{DE}$. Hence, $\mathbb{E}\left[\min\left\{\hat{\mu}^B_{a^*_K}, \hat{\mu}^*_{SE}\right\}\right] \geq \mathbb{E}\left[\min\left\{\hat{\mu}^*_{DE}, \hat{\mu}^*_{SE}\right\}\right] = \mathbb{E}\left[\hat{\mu}^*_{CDE}\right], 1 \leq K \leq N$. □

B. Proof of Convergence of Action Candidate Based Clipped Double Q-learning

For current variants of Double Q-learning, there are two main updating methods including random updating and simultaneous updating. In former method, only one Q-function is updated while in latter method, we update both them with the same target value. In this section, we prove that our action candidate based clipped Double Q-learning can converge to the optimal action value for both updaing methods under finite MDP setting.

B.1 Convergence Analysis on Random Updating

In our action candidate based clipped Double Q-learning (see Algorithm 1 in the paper), we randomly choose one Q-function to update its action value in each time step. Specifically, with collected experience $\langle s_t, a_t, r_t, s_{t+1}\rangle$, if we update Q^A, the updating formula is shown as below:

$$Q^A_{t+1}\left(s_t, a_t\right) \leftarrow Q^A_t\left(s_t, a_t\right) + \alpha_t\left(s_t, a_t\right)\left(r_t + \gamma \min\left\{Q^B_t\left(s_{t+1}, a^*_K\right), Q^A_t\left(s_{t+1}, a^*\right)\right\} - Q^A_t\left(s_t, a_t\right)\right), \tag{7}$$

where $a^*_K = \arg\max_{a \in \mathcal{M}_K} Q^A_t\left(s_{t+1}, a\right)$ with $\mathcal{M}_K = \left\{a_i \mid Q^B_t\left(s_{t+1}, a_i\right) \in \text{top } K \text{ values in } Q^B_t\left(s_{t+1}, \cdot\right)\right\}$ and $a^* = \arg\max_a Q^A\left(s_{t+1}, a\right)$. Instead, if we update Q^B_t, the updating formula is:

$$Q^B_{t+1}\left(s_t, a_t\right) \leftarrow Q^B_t\left(s_t, a_t\right) + \alpha_t\left(s_t, a_t\right)\left(r_t + \gamma \min\left\{Q^A_t\left(s_{t+1}, b^*_K\right), Q^B_t\left(s_{t+1}, b^*\right)\right\} - Q^B_t\left(s_t, a_t\right)\right), \tag{8}$$

where $b^*_K = \arg\max_{a \in \mathcal{M}_K} Q^B_t\left(s_{t+1}, a\right)$ with $\mathcal{M}_K = \left\{a_i \mid Q^A_t\left(s_{t+1}, a_i\right) \in \text{top } K \text{ values in } Q^A_t\left(s_{t+1}, \cdot\right)\right\}$ and $b^* = \arg\max_a Q^B_t\left(s_{t+1}, a\right)$. Next, we prove that our clipped Double Q-learning can converge to the optimal Q-function $Q^*\left(s, a\right)$ under the updating method above.

Lemma 1. *Consider a stochastic process (ζ_t, Δ_t, F_t), $t \geq 0$, where ζ_t, Δ_t, $F_t : X \to \mathbb{R}$ satisfy the equations:*

$$\Delta_{t+1}(x_t) = (1 - \zeta_t(x_t))\,\Delta_t(x_t) + \zeta_t(x_t)\,F_t(x_t), \tag{9}$$

where $x_t \in X$ and $t = 0, 1, 2, \ldots$. Let P_t be a sequence of increasing σ-fields such that ζ_0 and Δ_0 are P_0-measurable and ζ_t, Δ_t and F_{t-1} are P_t-measure, $t = 1, 2, \ldots$. Assume that the following hold:

1) The set X is finite.

2) $\zeta_t(x_t) \in [0, 1]$, $\sum_t \zeta_t(x_t) = \infty$, $\sum_t (\zeta_t(x_t))^2 < \infty$ with probability 1 and $\forall x \neq x_t : \zeta_t(x) = 0$.

3) $\|\mathbb{E}[F_t \mid P_t]\| \leq \kappa \|\Delta_t\| + c_t$, where $\kappa \in [0, 1)$ and c_t converges to zero with probability 1.

4) $\mathrm{Var}[F_t(x_t) \mid P_t] \leq K(1 + \kappa ||\Delta_t||)^2$, where K is some constant. Here $\|\cdot\|$ denotes a maximum norm.

Then Δ_t converges to zero with probability 1.

Theorem 2. *Given the following conditions:*

1) Each state action pair is sampled an infinite number of times.

2) The MDP is finite, that is $|S \times A| < \infty$.

3) $\gamma \in [0, 1)$.

4) Q values are stored in a lookup table.

5) Both Q^A and Q^B receive an infinite number of updates.

6) The learning rates satisfy $\alpha_t(s, a) \in [0, 1], \sum_t \alpha_t(s, a) = \infty, \sum_t (\alpha_t(s, a))^2 < \infty$ with probability 1 and $\alpha_t(s, a) = 0, \forall (s, a) \neq (s_t, a_t)$.

7) $\mathrm{Var}[r(s, a)] < \infty, \forall s, a$.

Then, our proposed action candidate based clipped Double Q-learning under random updating will converge to the optimal value function Q^* with probability 1.

Proof. We apply Lemma 1 with $P_t = \{Q_0^A, Q_0^B, s_0, a_0, \alpha_0, r_1, s_1, \ldots, s_t, a_t\}$, $X = S \times A$, $\Delta_t = Q_t^A - Q^*$, $\zeta_t = \alpha_t$ and $F_t(s_t, a_t) = r_t + \gamma \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} - Q^*(s_t, a_t)$, where $a^* = \arg\max_a Q_t^A(s_{t+1}, a)$ and $a_K^* = \arg\max_{\mathrm{a} \in \boldsymbol{\mathcal{M}}_\mathrm{K}} \mathrm{Q^A}(\mathrm{s_{t+1}}, \mathrm{a})$. The conditions 1 and 4 of Lemma 1 can hold by the conditions 2 and 7 of Theorem 1, respectively. Condition 2 in Lemma 1 holds by the condition 6 in Theorem 2 along with our selection of $\zeta_t = \alpha_t$.

Then, we just need verify the condition 3 on the expected condition of F_t holds. We can write:

$$\begin{aligned} F_t(s_t, a_t) &= r_t + \gamma \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} - Q^*(s_t, a_t) + \gamma Q_t^A(s_t, a_t) - \gamma Q_t^A(s_t, a_t) \\ &= r_t + \gamma Q_t^A(s_t, a_t) - Q^*(s_t, a_t) + \gamma \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} - \gamma Q_t^A(s_t, a_t) \\ &= F_t^Q(s_t, a_t) + \gamma \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} - \gamma Q_t^A(s_t, a_t), \end{aligned} \tag{10}$$

where $F_t^Q = r_t + \gamma Q_t^A(s_t, a_t) - Q^*(s_t, a_t)$ is the value of F_t if normal Q-learning would be under consideration. It is wll-known that $\mathbb{E}\left[F_t^Q | P_t\right] \leq \gamma \|\Delta_t\|$, so in order to apply the lemma we identify $c_t = \gamma \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} - \gamma Q_t^A(s_t, a_t)$ and it suffices to show that $\Delta_t^{BA} = Q_t^B - Q_t^A$ converges to to zero. Depending on whether Q^B or Q^A is updated, the update of Δ_t^{BA} at time t is either

$$\begin{aligned} \Delta_{t+1}^{BA}(s_t, a_t) &= \Delta_t^{BA}(s_t, a_t) + \alpha_t(s_t, a_t)\,F_t^B(s_t, a_t), \text{ or} \\ \Delta_{t+1}^{BA}(s_t, a_t) &= \Delta_t^{BA}(s_t, a_t) - \alpha_t(s_t, a_t)\,F_t^A(s_t, a_t), \end{aligned} \tag{11}$$

where $F_t^A(s_t, a_t) = r_t + \gamma \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} - Q_t^A(s_t, a_t)$ and $F_t^B(s_t, a_t) = r_t + \gamma \min\{Q_t^A(s_{t+1}, b_K^*), Q_t^B(s_{t+1}, b^*)\} - Q_t^B(s_t, a_t)$. We define $\zeta_t^{BA} = \frac{1}{2}\alpha_t$. Then,

$$\begin{aligned} \mathbb{E}\left[\Delta_{t+1}^{BA}(s_t, a_t) \mid P_t\right] &= \Delta_t^{BA}(s_t, a_t) + \mathbb{E}\left[\alpha_t(s_t, a_t)\,F_t^B(s_t, a_t) - \alpha_t(s_t, a_t)\,F_t^A(s_t, a_t) \mid P_t\right] \\ &= \left(1 - \zeta_t^{BA}(s_t, a_t)\right)\Delta_t^{BA}(s_t, a_t) + \zeta_t^{BA}(s_t, a_t)\,\mathbb{E}\left[F_t^{BA}(s_t, a_t) \mid P_t\right], \end{aligned} \tag{12}$$

where $\mathbb{E}\left[F_t^{BA}(s_t, a_t) \mid P_t\right] = \gamma \mathbb{E}\left[\min\{Q_t^A(s_{t+1}, b_K^*), Q_t^B(s_{t+1}, b^*)\} - \min\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\} \mid P_t\right]$. For this step it is important that the selection whether to update Q^A or Q^B is independent on the sample (e.g. random).

Assume $\mathbb{E}\left[\min\left\{Q_t^A(s_{t+1},b_K^*),Q_t^B(s_{t+1},b^*)\right\}\right] \geq \mathbb{E}\left[\min\left\{Q_t^B(s_{t+1},a_K^*),Q_t^A(s_{t+1},a^*)\right\} \mid P_t\right]$. Then,

$$
\begin{aligned}
&\left|\mathbb{E}\left[F_t^{BA}\left(s_t,a_t\right) \mid P_t\right]\right| = \gamma\mathbb{E}\left[\min\left\{Q_t^A(s_{t+1},b_K^*),Q_t^B(s_{t+1},b^*)\right\} - \min\left\{Q_t^B(s_{t+1},a_K^*),Q_t^A(s_{t+1},a^*)\right\} \mid P_t\right] \\
\leq&\gamma\mathbb{E}\left[\min\left\{Q_t^A(s_{t+1},a^*),Q_t^B(s_{t+1},b^*)\right\} - \min\left\{Q_t^B(s_{t+1},a^*),Q_t^A(s_{t+1},a^*)\right\} \mid P_t\right] \\
\leq&\gamma\mathbb{E}\left[Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right] - \gamma\mathbb{E}\left[\min\left\{Q_t^B(s_{t+1},a^*),Q_t^A(s_{t+1},a^*)\right\} \mid P_t\right] \\
=&\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) \geq Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^A(s_{t+1},a^*) \mid Q_t^B\left(s_{t+1},a^*\right) \geq Q_t^A\left(s_{t+1},a^*\right),P_t\right] \\
&+\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^A(s_{t+1},a^*) \mid Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right),P_t\right] \\
&-\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) \geq Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^A(s_{t+1},a^*) \mid Q_t^B\left(s_{t+1},a^*\right) \geq Q_t^A\left(s_{t+1},a^*\right),P_t\right] \\
&-\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^B(s_{t+1},a^*) \mid Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right),P_t\right] \\
=&\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\mathbb{E}\left[\underbrace{Q_t^A(s_{t+1},a^*) - Q_t^B(s_{t+1},a^*)}_{\leq\|\Delta_t^{BA}\|} \mid Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right),P_t\right] \\
\leq&\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) < Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\|\Delta_t^{BA}\| \leq \gamma\|\Delta_t^{BA}\|,
\end{aligned}
\tag{13}
$$

where the first inequality is based on the monotonicity in Theorem 1. From Theorem 1, we have that the expected value of $\min\left\{Q_t^A(s_{t+1},b_K^*),Q_t^B(s_{t+1},b^*)\right\} - \min\left\{Q_t^B(s_{t+1},a_K^*),Q_t^A(s_{t+1},a^*)\right\}$ is no more than the one of $\min\left\{Q_t^A(s_{t+1},b_1^*),Q_t^B(s_{t+1},b^*)\right\} - \min\left\{Q_t^B(s_{t+1},a_N^*),Q_t^A(s_{t+1},a^*)\right\}$. Since $a^* = b_1^* = a_N^*$, we can have the first inequality above. The second inequality is based on that since $\min\left\{Q_t^A(s_{t+1},a^*),Q_t^B(s_{t+1},b^*)\right\}$ is no more than $Q_t^A(s_{t+1},a^*)$, the expected value of the former is also no more than the one of latter.

Now assume $\mathbb{E}\left[\min\left\{Q_t^A(s_{t+1},b_K^*),Q_t^B(s_{t+1},b^*)\right\}\right] < \mathbb{E}\left[\min\left\{Q_t^B(s_{t+1},a_K^*),Q_t^A(s_{t+1},a^*)\right\} \mid P_t\right]$ and therefore

$$
\begin{aligned}
&\left|\mathbb{E}\left[F_t^{BA}\left(s_t,a_t\right) \mid P_t\right]\right| = \gamma\mathbb{E}\left[\min\left\{Q_t^B(s_{t+1},a_K^*),Q_t^A(s_{t+1},a^*)\right\} - \min\left\{Q_t^A(s_{t+1},b_K^*),Q_t^B(s_{t+1},b^*)\right\} \mid P_t\right] \\
\leq&\gamma\mathbb{E}\left[\min\left\{Q_t^B(s_{t+1},b^*),Q_t^A(s_{t+1},a^*)\right\} - \min\left\{Q_t^A(s_{t+1},b^*),Q_t^B(s_{t+1},b^*)\right\} \mid P_t\right] \\
\leq&\gamma\mathbb{E}\left[Q_t^B\left(s_{t+1},b^*\right) \mid P_t\right] - \gamma\mathbb{E}\left[\min\left\{Q_t^A(s_{t+1},b^*),Q_t^B(s_{t+1},b^*)\right\} \mid P_t\right] \\
=&\gamma P\left(Q_t^B\left(s_{t+1},b^*\right) \geq Q_t^A\left(s_{t+1},b^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^B(s_{t+1},b^*) \mid Q_t^B\left(s_{t+1},b^*\right) \geq Q_t^A\left(s_{t+1},b^*\right),P_t\right] \\
&+\gamma P\left(Q_t^B\left(s_{t+1},b^*\right) < Q_t^A\left(s_{t+1},b^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^B(s_{t+1},b^*) \mid Q_t^B\left(s_{t+1},b^*\right) < Q_t^A\left(s_{t+1},b^*\right),P_t\right] \\
&-\gamma P\left(Q_t^B\left(s_{t+1},b^*\right) \geq Q_t^A\left(s_{t+1},b^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^A(s_{t+1},b^*) \mid Q_t^B\left(s_{t+1},b^*\right) \geq Q_t^A\left(s_{t+1},b^*\right),P_t\right] \\
&-\gamma P\left(Q_t^B\left(s_{t+1},b^*\right) < Q_t^A\left(s_{t+1},b^*\right) \mid P_t\right)\mathbb{E}\left[Q_t^B(s_{t+1},b^*) \mid Q_t^B\left(s_{t+1},b^*\right) < Q_t^A\left(s_{t+1},b^*\right),P_t\right] \\
=&\gamma P\left(Q_t^B\left(s_{t+1},b^*\right) \geq Q_t^A\left(s_{t+1},b^*\right) \mid P_t\right)\mathbb{E}\left[\underbrace{Q_t^B(s_{t+1},b^*) - Q_t^A(s_{t+1},b^*)}_{\leq\|\Delta_t^{BA}\|} \mid Q_t^B\left(s_{t+1},b^*\right) \geq Q_t^A\left(s_{t+1},b^*\right),P_t\right] \\
\leq&\gamma P\left(Q_t^B\left(s_{t+1},a^*\right) \geq Q_t^A\left(s_{t+1},a^*\right) \mid P_t\right)\|\Delta_t^{BA}\| \leq \gamma\|\Delta_t^{BA}\|,
\end{aligned}
\tag{14}
$$

where the first inequality is based on the monotonicity in Theorem 1. From Theorem 1, we have the expected value of $\min\left\{Q_t^B(s_{t+1},a_K^*),Q_t^A(s_{t+1},a^*)\right\} - \min\left\{Q_t^A(s_{t+1},b_K^*),Q_t^B(s_{t+1},b^*)\right\}$ is no more than the one of $\min\left\{Q_t^B(s_{t+1},a_1^*),Q_t^A(s_{t+1},a^*)\right\} - \min\left\{Q_t^A(s_{t+1},b_N^*),Q_t^B(s_{t+1},b^*)\right\}$. Since $b^* = a_1^* = b_N^*$, we can have the first inequality above. The second inequality is based on that since $\min\left\{Q_t^B(s_{t+1},b^*),Q_t^A(s_{t+1},a^*)\right\}$ is no more than $Q_t^B(s_{t+1},b^*)$, the expected value of the former is also no more than the one of latter.

Clearly, one of the assumptions must hold at each time step and in both cases we obtain the desired result that $\left|\mathbb{E}\left[F_t^{BA} \mid P_t\right]\right| \leq \gamma\|\Delta_t^{BA}\|$. Applying the lemma yields convergence of Δ_t^{BA} to zero, which in turn ensures that the original process also converges in the limit. □

B.2 Convergence Analysis on Simultaneous Updating

In Algorithm 2 of the paper, we update our two Q-functions with the same target value in each time step. In this section, we further prove that our action candidate based clipped Double Q-learning can also converge to the optimal Q-function $Q^*\left(s,a\right)$ under such updating method.

Specificially, with the collected experience $\langle s_t,a_t,r_t,s_{t+1}\rangle$, we set the target value y_t as below:

$$
y_t = r_t + \gamma\min\left\{Q^B\left(s_{t+1},a_K^*\right),Q^A\left(s_{t+1},a^*\right)\right\}, \tag{15}
$$

where $a_K^* = \arg\max_{a \in \mathcal{M}_K} Q^A(s_{t+1}, a)$ with $\mathcal{M}_K = \left\{a_i \mid Q^B(s_{t+1}, a_i) \in \text{top } K \text{ values in } Q^B(s_{t+1}, \cdot)\right\}$ and $a^* = \arg\max_a Q^A(s_{t+1}, a)$. Then, both Q-functions are updated as below:

$$\begin{aligned} Q_{t+1}^A(s_t, a_t) &\leftarrow Q_t^A(s_t, a_t) + \alpha_t(s_t, a_t)\left(y_t - Q_t^A(s_t, a_t)\right) \\ Q_{t+1}^B(s_t, a_t) &\leftarrow Q_t^B(s_t, a_t) + \alpha_t(s_t, a_t)\left(y_t - Q_t^B(s_t, a_t)\right) \end{aligned} \tag{16}$$

Theorem 3. *Given the following conditions:*
1) Each state action pair is sampled an infinite number of times.
2) The MDP is finite, that is $|S \times A| < \infty$.
3) $\gamma \in [0, 1)$.
4) Q values are stored in a lookup table.
5) Both Q^A *and* Q^B *receive an infinite number of updates.*
6) The learning rates satisfy $\alpha_t(s, a) \in [0, 1], \sum_t \alpha_t(s, a) = \infty, \sum_t (\alpha_t(s, a))^2 < \infty$ *with probability 1 and* $\alpha_t(s, a) = 0, \forall (s, a) \neq (s_t, a_t)$.
7) $\text{Var}[r(s, a)] < \infty, \forall s, a$.
Then, our proposed action candidate based clipped Double Q-learning under simultaneous updating will converge to the optimal value function Q^* *with probability 1.*

Proof. We apply Lemma 1 with $P_t = \left\{Q_0^A, Q_0^B, s_0, a_0, \alpha_0, r_1, s_1, \ldots, s_t, a_t\right\}$, $X = S \times A$, $\Delta_t = Q_t^A - Q^*$, $\zeta_t = \alpha_t$ and $F_t(s_t, a_t) = r_t + \gamma \min\left\{Q_t^B(s_{t+1}, a_K^*), Q_t^A(s_{t+1}, a^*)\right\} - Q^*(s_t, a_t)$, where $a^* = \arg\max_a Q_t^A(s_{t+1}, a)$ and $a_K^* = \arg\max_{\mathrm{a} \in \boldsymbol{\mathcal{M}}_{\mathrm{K}}} \mathrm{Q}^{\mathrm{A}}(\mathrm{s}_{\mathrm{t}+1}, \mathrm{a})$. The conditions 1 and 4 of Lemma 1 can hold by the conditions 2 and 7 of Theorem 3, respectively. Condition 2 in Lemma 1 holds by the condition 6 in Theorem 3 along with our selection of $\zeta_t = \alpha_t$. Further, we have

$$\begin{aligned} \Delta_{t+1}(s_t, a_t) &= (1 - \alpha_t(s_t, a_t))\left(Q_t^A(s_t, a_t) - Q^*(s_t, a_t)\right) + \alpha_t(s_t, a_t)(y_t - Q^*(s_t, a_t)) \\ &= (1 - \alpha_t(s_t, a_t))\Delta_t(s_t, a_t) + \alpha_t(s_t, a_t) F_t(s_t, a_t), \end{aligned} \tag{17}$$

where we have defined $F_t(s_t, a_t)$ as:

$$\begin{aligned} F_t(s_t, a_t) &= y_t - Q_t^*(s_t, a_t) = y_t - Q_t^*(s_t, a_t) + \gamma Q_t^A(s_{t+1}, a^*) - \gamma Q_t^A(s_{t+1}, a^*) \\ &= F_t^Q(s_t, a_t) + c_t, \end{aligned} \tag{18}$$

where $F_t^Q = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$ denotes the value of F_t under standard Q-learing and

$$c_t = \gamma \min\left\{Q^B(s_{t+1}, a_K^*), Q^A(s_{t+1}, a^*)\right\} - \gamma Q_t^A(s_{t+1}, a^*). \tag{19}$$

As $\mathbb{E}\left[F_t^Q | P_t\right] \leq \gamma \|\Delta_t\|$ is a well-known result, condition 3 of Lemma 1 holds if it can be shown that c_t converges to 0 with probability 1. Further, due to $\min\left\{Q^B(s_{t+1}, a_K^*), Q^A(s_{t+1}, a^*)\right\}$ is no more than $Q^B(s_{t+1}, a_K^*)$ and $Q^B(s_{t+1}, a_K^*)$ is also no more than $Q^B(s_{t+1}, a_1^*)$ (based on the Property 1), we can have $\min\left\{Q^B(s_{t+1}, a_K^*), Q^A(s_{t+1}, a^*)\right\} \leq Q^B(s_{t+1}, a_1^*)$. Since $a_1^* = b^*$, we can have $Q^B(s_{t+1}, a_1^*) = Q^B(s_{t+1}, b^*)$. Finally, due to $Q_t^A(s_{t+1}, a^*) \geq Q_t^A(s_{t+1}, b^*)$, we can know that:

$$c_t = \gamma \min\left\{Q^B(s_{t+1}, a_K^*), Q^A(s_{t+1}, a^*)\right\} - \gamma Q_t^A(s_{t+1}, a^*) \leq \gamma Q^B(s_{t+1}, b^*) - \gamma Q^A(s_{t+1}, b^*). \tag{20}$$

Thus, c_t converges to 0 if Δ_t^{BA} converges to 0 where $\Delta_t^{BA}(s_t, a_t) = Q_t^B(s_t, a_t) - Q_t^A(s_t, a_t)$. The update of Δ_t^{BA} at time t is the sum of updates of Q^A and Q^B:

$$\begin{aligned} \Delta_{t+1}^{BA}(s_t, a_t) &= \Delta_t^{BA}(s_t, a_t) + \alpha_t(s_t, a_t)\left(y_t - Q_t^B(s_t, a_t) - \left(y_t - Q_t^A(s_t, a_t)\right)\right) \\ &= \Delta_t^{BA}(s_t, a_t) + \alpha_t(s_t, a_t)\left(Q_t^A(s_t, a_t) - Q_t^B(s_t, a_t)\right) \\ &= (1 - \alpha_t(s_t, a_t))\Delta_t^{BA}(s_t, a_t). \end{aligned} \tag{21}$$

Clearly, Δ_t^{BA} converges to 0, which then shows we have satisfied condition 3 of Lemma , which implies that $Q^A(s_t, a_t)$ converges to $Q^*(s_t, a_t)$. Similarly, wo get the convergence of $Q^B(s_t, a_t)$ to the optimal value function by choosing $\Delta_t = Q_t^B - Q^*$ and repeating the same arguments, thus proving Theorem 3. □

C. Additional Experimental Results

In this section, we provide some additional experimental results on Grid World, MinAtar and MuJoCo tasks.

Grid World In Grid World environment, for action candidate based clipped Double Q-learning (AC-CDQ), we further evaluate its performance on the grid world game with size 3×3 and 4×4. As shown in Fig 1, benefiting from the precise estimation about the optimal action value (closest to the dash line), AC-CDQ ($K = 2$) presents the superior performance.

MinAtar In MinAtar games, for action candidate based clipped Double DQN (AC-CDDQN), we additionally list the learning curves about the averaged reward and the estimated maximum action value with different numbers of the action candidates ($K = \{2, 3, 4\}$). As shown in Fig 2, except for the case that $K = 4$ in Breakout game, our method can consistently perform better than clipped Double DQN. Moreover, as shown in the two plots on the right, our deep version can effectively balance the overestimated DQN and the underestimated clipped Double DQN. Further, it also empirically follows the monotonicity in Theorem 1, that is as the number K of action candidates decreases, the underestimation bias in clipped Double DQN reduces monotonically.

MuJoCo Tasks In MuJoCo tasks, for action candidate based TD3 (AC-TD3), we provide the additional learning curves on Walker2D-v2, Pusher-v2, Hopper-v2 and Reacher-v2 in Fig 3. Moreover, we also test the performance variance under different number of the action candidates ($K = \{32, 64, 128\}$) in Walker2D-v2, Pusher-v2, Swimmer-v2 and Ant-v2 in Fig 4. The plots show that AC-TD3 can consistently obtain the robust and superior performance with different action candidate sizes.

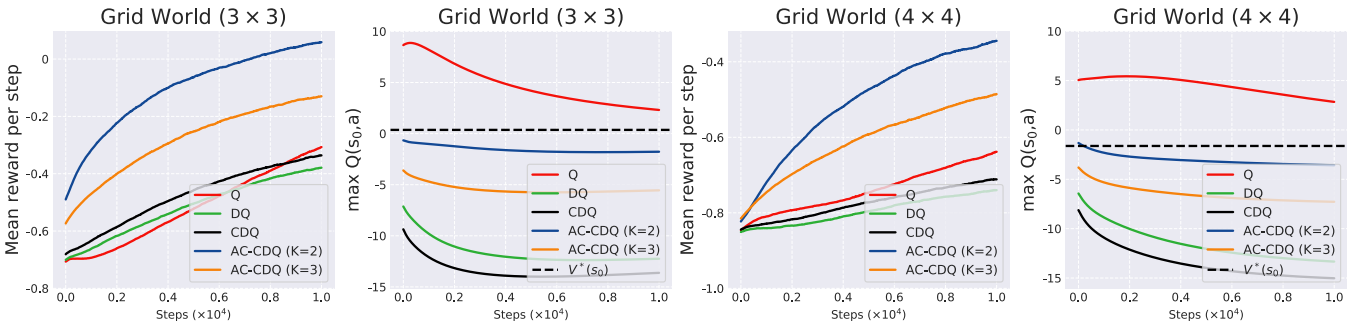

Figure 1: Learning curves about the mean reward per step and the estimated maximum action value from the state s_0 (the black dash line demotes the optimal state value $V^*(s_0)$). The results are averaged over 10000 experiments and each experiment contains 10000 steps. We set the number of the action candidates to 2 and 3, respectively. Q: Q-learning, DQ: Double Q-learning, CDQ: clipped Double Q-learning.

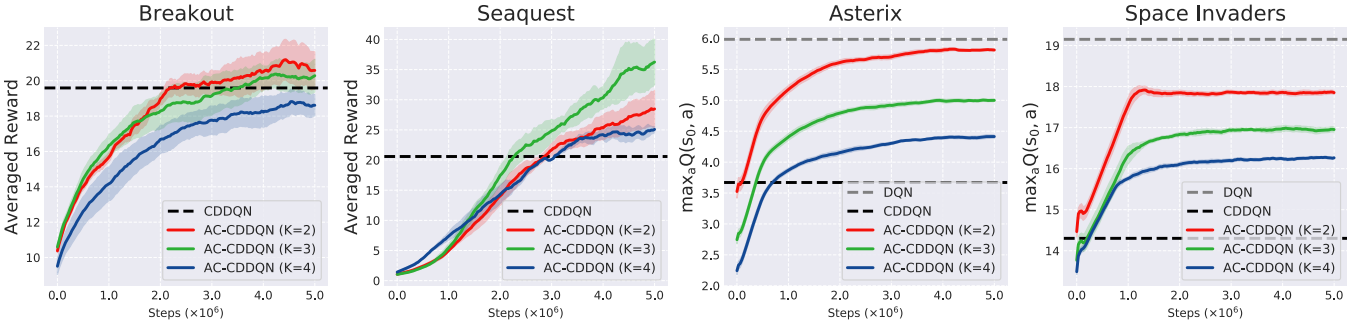

Figure 2: Learning curves about the averaged reward (two plots on the left) and estimated maximum action value (two plots on the right) for AC-CDDQN with different numbers of the action candidates (K=2, 3, 4). CDDQN: clipped Double DQN.

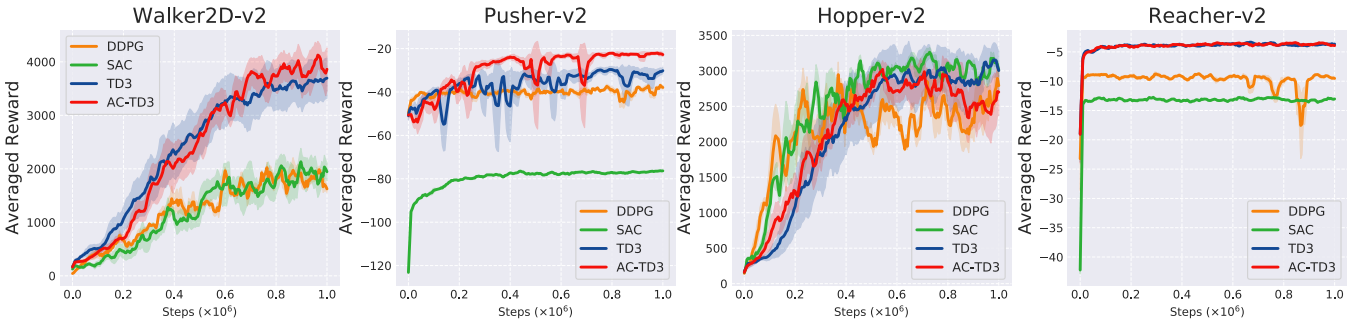

Figure 3: Learning curves for the OpenAI Gym continuous control tasks. The shaded region represents half a standard deviation of the average evaluation over 10 trials.

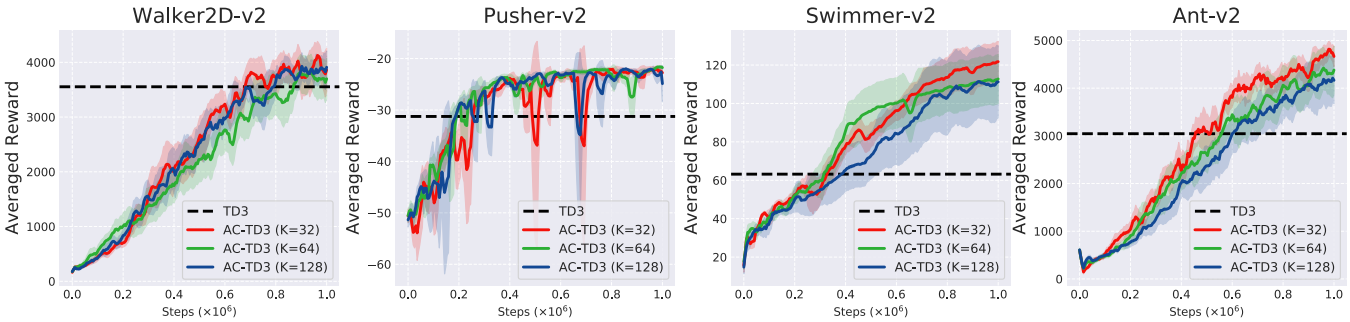

Figure 4: Learning curves for AC-TD3 with different numbers of the action candidates ($K = \{32, 64, 128\}$). The shaded region represents half a standard deviation of the average evaluation over 10 trials.

D. Hyper-parameters Setting

Action Candidate Based Clipped Double DQN In this method, the number of frames is $5 \cdot 10^6$; the discount factor is 0.99; reward scaling is 1.0; the batch size is 32; the buffer size is $1 \cdot 10^6$; the frequency of updating the target network is 1000; the optimizer is RMSprop with learning $2.5 \cdot 10^{-4}$, squared gradient momentum 0.95 and minimum squared gradient 0.01; the iteration per time step is 1.

Action Candidate Based TD3 In this method, the number of frames is $1 \cdot 10^6$; the discount factor is 0.99; reward scaling is 1.0; the batch size is 256; the buffer size is $1 \cdot 10^6$; the frequency of updating the target network is 2; the optimizers for actor and critic are Adams with learning $3 \cdot 10^{-4}$; the iteration per time step is 1. All experiments are conducted on a server with NVIDIA TITAN V.