A Appendix

A.1 Detailed Theoretical Analysis

In this section, we will introduce theoretical properties of the ILQ method in detail. First, we investigate the convergence of value iterations using the ILB operator in tabular MDPs, as confirmed in Theorem 1. Additionally, unlike value regularization methods, we do not intend for ILQ to be a pessimistic algorithm. Instead, it aims to retain reasonable estimates of OOD action-values under appropriate restrictions. Thus, in Theorem 4, we analyze the action-value gap between the fixed point of policy evaluation and the Bellman optimality value.

Now we begin by presenting the analysis of convergence. To facilitate reading, the definition of our ILB operator is restated here.

Definition 1. *The Imagination-Limited Bellman (ILB) operator is defined as*

$$\mathcal{T}_{\mathrm{ILB}}Q(s,a) = \begin{cases} r(s,a)+\gamma\mathbb{E}_{s'\sim P}\big[\max\limits_{\tilde{a}'\sim\pi} Q(s',\tilde{a}')\big], & \text{if } \beta(a|s)>0 \\ \min\left\{y^Q_{\mathrm{img}}, y^Q_{\mathrm{lmt}}\right\}+\delta, & \text{otherwise.} \end{cases} \tag{22}$$

where β is the behavior policy,

$$y^Q_{\mathrm{img}} = \widehat{r}(s,a)+\gamma\mathbb{E}_{\widehat{s}'\sim\widehat{P}(\cdot|s,a)}\left[\max_{\tilde{a}'\sim\pi} Q(\widehat{s}',\tilde{a}')\right], \tag{23}$$

and

$$y^Q_{\mathrm{lmt}} = \max_{\widehat{a}\in\mathrm{Supp}(\beta(\cdot|s))} Q(s,\widehat{a}) \tag{24}$$

are the imagined value and its limitation, respectively. The $\widehat{P}$ is the empirical transition kernel, $\widehat{r}$ is the empirical reward function, δ is a hyperparameter with a small absolute value, and $\mathrm{Supp}(\cdot)$ *means support-constrained on the dataset.*

Theorem 1 (Convergence). *The ILB operator defined in* (2) *is a γ-contraction operator in the $\mathcal{L}_\infty$ norm, and Q-function iteration rule obeying the ILB operator can converge to a unique fixed point.*

Proof. Let Q_1 and Q_2 be two arbitrary Q-functions. To prove the γ-contraction property of the ILB operator, we have to demonstrate that the following inequality holds:

$$\begin{aligned} &\|\mathcal{T}_{\mathrm{ILB}}Q_1-\mathcal{T}_{\mathrm{ILB}}Q_2\|_\infty \\ &=\max_{s,a}|\mathcal{T}_{\mathrm{ILB}}Q_1(s,a)-\mathcal{T}_{\mathrm{ILB}}Q_2(s,a)| \\ &\le\gamma\|Q_1-Q_2\|_\infty. \end{aligned} \tag{25}$$

Thus, we are required to carefully investigate $|\mathcal{T}_{\mathrm{ILB}}Q_1(s,a)-\mathcal{T}_{\mathrm{ILB}}Q_2(s,a)|$. We first consider the case of $a\in\mathrm{Supp}(\beta(\cdot\mid s))$. According to the definition of ILB operator (2), one has

$$\begin{aligned} &|\mathcal{T}_{\mathrm{ILB}}Q_1(s,a)-\mathcal{T}_{\mathrm{ILB}}Q_2(s,a)| \\ &=\left|\left(r(s,a)+\gamma\mathbb{E}_{s'\sim P}\left[\max_{\tilde{a}'\sim\pi}Q_1(s',\tilde{a}')\right]\right)\right. \\ &\quad\left.-\left(r(s,a)+\gamma\mathbb{E}_{s'\sim P}\left[\max_{\tilde{a}'\sim\pi}Q_2(s',\tilde{a}')\right]\right)\right| \\ &=\gamma\left|\mathbb{E}_{s'\sim P}\left[\max_{\tilde{a}'\sim\pi}Q_1(s',\tilde{a}')-\max_{\tilde{a}'\sim\pi}Q_2(s',\tilde{a}')\right]\right| \\ &\le\gamma\mathbb{E}_{s'\sim P}\left|\max_{\tilde{a}'\sim\pi}Q_1(s',\tilde{a}')-\max_{\tilde{a}'\sim\pi}Q_2(s',\tilde{a}')\right| \\ &\le\gamma\|Q_1-Q_2\|_\infty. \end{aligned} \tag{26}$$

Otherwise, when $a\notin\mathrm{Supp}(\beta(\cdot\mid s))$, we have the $\mathcal{T}_{\mathrm{ILB}}Q(s,a)=\min\left\{y^Q_{\mathrm{img}},y^Q_{\mathrm{lmt}}\right\}+\delta$. Therefore,

$$\begin{aligned} &|\mathcal{T}_{\mathrm{ILB}}Q_1(s,a)-\mathcal{T}_{\mathrm{ILB}}Q_2(s,a)| \\ &=\left|\left(\min\left\{y^{Q_1}_{\mathrm{img}},y^{Q_1}_{\mathrm{lmt}}\right\}+\delta\right)-\left(\min\left\{y^{Q_2}_{\mathrm{img}},y^{Q_2}_{\mathrm{lmt}}\right\}+\delta\right)\right| \\ &=\left|\min\left\{y^{Q_1}_{\mathrm{img}},y^{Q_1}_{\mathrm{lmt}}\right\}-\min\left\{y^{Q_2}_{\mathrm{img}},y^{Q_2}_{\mathrm{lmt}}\right\}\right| \end{aligned} \tag{27}$$

There exist four possible cases for the inner part on the RHS of (27) above, including $\left|y^{Q_1}_{\mathrm{lmt}}-y^{Q_2}_{\mathrm{lmt}}\right|$, $\left|y^{Q_1}_{\mathrm{img}}-y^{Q_2}_{\mathrm{img}}\right|$, $\left|y^{Q_1}_{\mathrm{img}}-y^{Q_2}_{\mathrm{lmt}}\right|$ and $\left|y^{Q_1}_{\mathrm{lmt}}-y^{Q_2}_{\mathrm{img}}\right|$. For the simplest case $\left|y^{Q_1}_{\mathrm{lmt}}-y^{Q_2}_{\mathrm{lmt}}\right|$, one can, analogous to the derivation process in (26), easily verify that the γ-contraction inequality holds.

For the second case, we have

$$\begin{aligned} &\left|y^{Q_1}_{\mathrm{img}}-y^{Q_2}_{\mathrm{img}}\right| \\ &=\left|\left(\widehat{r}(s,a)+\gamma\mathbb{E}_{\widehat{s}'\sim\widehat{P}(\cdot|s,a)}\left[\max_{\tilde{a}'\sim\pi}Q_1(\widehat{s}',\tilde{a}')\right]\right)\right. \\ &\quad\left.-\left(\widehat{r}(s,a)+\gamma\mathbb{E}_{\widehat{s}'\sim\widehat{P}(\cdot|s,a)}\left[\max_{\tilde{a}'\sim\pi}Q_2(\widehat{s}',\tilde{a}')\right]\right)\right| \\ &=\gamma\left|\sum_{\widehat{s}'}\widehat{P}(\widehat{s}'|s,a)\left[\max_{\tilde{a}'\sim\pi}Q_1(\widehat{s}',\tilde{a}')-\max_{\tilde{a}'\sim\pi}Q_2(\widehat{s}',\tilde{a}')\right]\right| \\ &\le\gamma\sum_{\widehat{s}'}\widehat{P}(\widehat{s}'|s,a)\left|\max_{\tilde{a}'\sim\pi}Q_1(\widehat{s}',\tilde{a}')-\max_{\tilde{a}'\sim\pi}Q_2(\widehat{s}',\tilde{a}')\right| \\ &\le\gamma\sum_{\widehat{s}'}\widehat{P}(\widehat{s}'|s,a)\|Q_1-Q_2\|_\infty \\ &=\gamma\|Q_1-Q_2\|_\infty. \end{aligned}$$

Now we consider two cross term cases. Without loss of generality, we only proof $\left|y^{Q_1}_{\mathrm{img}}-y^{Q_2}_{\mathrm{lmt}}\right|$ holds the contraction inequality. In this situation, we have $y^{Q_1}_{\mathrm{img}}\le y^{Q_1}_{\mathrm{lmt}}$ and $y^{Q_2}_{\mathrm{lmt}}\le y^{Q_2}_{\mathrm{img}}$. Therefore,

$$\left|y^{Q_1}_{\mathrm{img}}-y^{Q_2}_{\mathrm{lmt}}\right| = \begin{cases} y^{Q_1}_{\mathrm{img}}-y^{Q_2}_{\mathrm{lmt}}\le y^{Q_1}_{\mathrm{lmt}}-y^{Q_2}_{\mathrm{lmt}}, & \text{if } y^{Q_1}_{\mathrm{img}}>y^{Q_2}_{\mathrm{lmt}} \\ y^{Q_2}_{\mathrm{lmt}}-y^{Q_1}_{\mathrm{img}}\le y^{Q_2}_{\mathrm{img}}-y^{Q_1}_{\mathrm{img}}, & \text{otherwise.} \end{cases} \tag{28}$$

It can be rewritten as

$$\left|y_{\text{img}}^{Q_1} - y_{\text{lmt}}^{Q_2}\right| \leq \max\left\{\left|y_{\text{lmt}}^{Q_1} - y_{\text{lmt}}^{Q_2}\right|, \left|y_{\text{img}}^{Q_2} - y_{\text{img}}^{Q_1}\right|\right\}. \tag{29}$$

This means the third case can be bounded by either the first case or the second case. Hence, it also satisfies the γ-contraction inequality.

By combining these together, the (25) is obtained, i.e., the ILB operator is a contraction operator over space $\mathcal{S} \times \mathcal{A}$ with $\mathcal{L}_\infty$ norm when $\gamma < 1$. According to the Banach fixed-point theorem (contraction mapping theorem), the ILB operator converges to a unique fixed point. □

This shows that the convergence of the proposed ILB as a policy evaluation operator is guaranteed. In practice, the γ can be utilized to adjust both the convergence speed and the long-term influence of rewards. Nevertheless, it is typically fixed to 0.99 in almost all algorithms.

In the next step, we will analyze the error bound between the converged Q-value and the optimal Q-value. To accomplish this, we will first introduce the support-constrained Bellman optimality operator.

Lemma 1. *The support-constrained Bellman optimality operator*

$$\mathcal{T}_{\text{Supp}}Q(s,a) := r(s,a) + \gamma \mathbb{E}_{s' \sim P}\left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q(s',a')\right]$$

is also a γ-contraction operator and has a fixed point.

Proof. This result can be demonstrated by a derivation similar to (26). □

For clarity and ease of reference, the assumptions are also restated here. We make some commonly used assumptions about the reward function [Huang *et al.*, 2024, Assumption 1].

1. The reward function is bounded, i.e., $|r(s,a)| \leq r_{\max}$. Actually, this is consistent with what is required by its definition $r(s,a) : \mathcal{S} \times \mathcal{A} \to [-r_{\max}, r_{\max}]$.
2. Similar to the Lipschitz condition, i.e., $|r(s,\tilde{a}_1) - r(s,\tilde{a}_2)| \leq \ell \|\tilde{a}_1 - \tilde{a}_2\|_\infty$, $\forall s \in \mathcal{S}$ and $\forall \tilde{a}_1, \tilde{a}_2 \in \mathcal{A}$, where ℓ is a constant. This requires that the reward function satisfies Lipschitz continuity with respect to actions.

And the error bound assumption between the empirical models and the real ones are required, which is also utilized in both [Kumar *et al.*, 2020] and [Huang *et al.*, 2024]. Suppose the $\widehat{r}$ and $\widehat{P}$ are the empirical reward function and empirical transition dynamics, respectively, the following relationships

$$\left\|\widehat{r}(s,a) - r(s,a)\right\|_1 \leq \zeta_r/\sqrt{D}, \tag{30}$$

$$\left\|\widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a)\right\|_1 \leq \zeta_P/\sqrt{D}, \tag{31}$$

hold with high probability $\geq 1-\zeta$, $\zeta \in (0,1)$, where D is the constant related to the dataset size, ζ_r and ζ_P are constants related to ζ.

Theorem 2. *Suppose Q_{β^*} is the fixed point of the support-constrained Bellman optimality operator. The following gap can be obtained*

$$|Q_{\beta^*}(s,\pi(s)) - Q_{\beta^*}(s,\beta(s))| \leq \ell\epsilon_\pi + \gamma\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P, \tag{32}$$

where $\epsilon_\pi := \max_s \|\pi(s) - \beta(s)\|_\infty$ *and* $\epsilon_P := \left\|P^\pi - P^\beta\right\|_\infty$.

Proof. Since Q_{β^*} is the fixed point of the $\mathcal{T}_{\text{Supp}}Q$, the following equation holds

$$Q_{\beta^*}(s,a) = \mathcal{T}_{\text{Supp}}Q_{\beta^*}(s,a) \tag{33}$$

for all (s,a) in the state-action space. This yields

$$\begin{aligned}
&|Q_{\beta^*}(s,\pi(s)) - Q_{\beta^*}(s,\beta(s))| \\
&= \Bigg| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))}\left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s',a')\right] \\
&\quad - r(s,\beta(s)) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,\beta(s))}\left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s',a')\right]\Bigg| \\
&\leq |r(s,\pi(s)) - r(s,\beta(s))| \\
&\quad + \gamma\Bigg|\mathbb{E}_{s' \sim P(\cdot|s,\pi(s))}\left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s',a')\right] \\
&\quad - \mathbb{E}_{s' \sim P(\cdot|s,\beta(s))}\left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s',a')\right]\Bigg| \\
&\leq |r(s,\pi(s)) - r(s,\beta(s))| \\
&\quad + \gamma \sum_{s' \in \mathcal{S}}\Bigg\{\left|P(s' \mid s,\pi(s)) - P(s' \mid s,\beta(s))\right| \\
&\quad \cdot \left|\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s',a')\right|\Bigg\} \\
&\leq \ell \max_s \|\pi(s) - \beta(s)\|_\infty + \gamma\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\left\|P^\pi - P^\beta\right\|_\infty \qquad (34) \\
&= \ell\epsilon_\pi + \gamma\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P.
\end{aligned}$$

The first term in the last inequality (34) holds on the basis of the assumption of Lipschitz condition. The second term holds based on the boundedness of rewards. Actually, for any Q-function, we have

$$|Q(s,a)| = \left|\mathbb{E}\sum_{t=1}^{\infty}\gamma^t r_t\right| \leq \mathbb{E}\sum_{t=1}^{\infty}\gamma^t|r_t| \leq \frac{r_{\max}}{1-\gamma}. \tag{35}$$

Clearly, it still holds for Q_{β^*}. We thus obtain the final inequality. □

Theorem 3. *Suppose Q_{β^*} is the fixed point of support-constrained Bellman optimality operator. The gap between the imagination value $y_{\text{img}}^{Q_{\beta^*}}$ and Q_{β^*} has:*

$$\begin{aligned}
&\left|y_{\text{img}}^{Q_{\beta^*}} - Q_{\beta^*}(s,a)\right| \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma\ell\epsilon_\pi + \gamma^2\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P + \gamma\frac{\zeta_P}{\sqrt{D}}\frac{r_{\max}}{1-\gamma}.
\end{aligned} \tag{36}$$

Proof. By applying the definition of y and (33), we obtain

$$
\begin{aligned}
&\left|y_{\text{img}}^{Q_{\beta^*}} - Q_{\beta^*}(s,a)\right| \\
&= \left|y_{\text{img}}^{Q_{\beta^*}} - \mathcal{T}_{\text{Supp}} Q_{\beta^*}(s,a)\right| \\
&= \Bigg| \widehat{r}(s,a) + \gamma \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{\tilde{a}' \sim \pi} Q_{\beta^*}(\widehat{s}', \tilde{a}') \right] \\
&\quad - r(s,a) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right] \Bigg| \\
&\leq |\widehat{r}(s,a) - r(s,a)| + \gamma \Bigg| \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{\tilde{a}' \sim \pi} Q_{\beta^*}(\widehat{s}', \tilde{a}') \right] \\
&\quad - \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right] \Bigg| \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma \Bigg| \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{\tilde{a}' \sim \pi} Q_{\beta^*}(\widehat{s}', \tilde{a}') \right] \\
&\quad - \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right] \Bigg|
\end{aligned}
$$

The last inequality is derived based on the concentration assumption (10) of the empirical reward model. Now, using the triangle inequality, we can infer that

$$
\begin{aligned}
&\left|y_{\text{img}}^{Q_{\beta^*}} - Q_{\beta^*}(s,a)\right| \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma \Bigg| \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{\tilde{a}' \sim \pi} Q_{\beta^*}(\widehat{s}', \tilde{a}') \right] \\
&\quad - \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|\widehat{s}'))} Q_{\beta^*}(\widehat{s}', a') \right] \Bigg| \\
&\quad + \gamma \Bigg| \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|\widehat{s}'))} Q_{\beta^*}(\widehat{s}', a') \right] \\
&\quad - \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right] \Bigg| \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \Bigg| \max_{\tilde{a}' \sim \pi} Q_{\beta^*}(\widehat{s}', \tilde{a}') \\
&\quad - \max_{a' \in \text{Supp}(\beta(\cdot|\widehat{s}'))} Q_{\beta^*}(\widehat{s}', a') \Bigg| \\
&\quad + \gamma \Bigg| \mathbb{E}_{\widehat{s}' \sim \widehat{P}(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|\widehat{s}'))} Q_{\beta^*}(\widehat{s}', a') \right] \\
&\quad - \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right] \Bigg| \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma \sum_{\widehat{s}'} \widehat{P}(\widehat{s}'|s,a) \Bigg(\ell \left\| \pi(\widehat{s}') - \beta(\widehat{s}') \right\|_\infty \\
&\qquad + \gamma \frac{|\mathcal{S}| r_{\max}}{1-\gamma} \left\| P^\pi - P^\beta \right\|_\infty \Bigg) \\
&\quad + \gamma \sum_{s'} \Bigg(\left| \widehat{P}(s'|s,a) - P(s'|s,a) \right| \\
&\qquad \cdot \left| \max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right| \Bigg) \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma \ell \epsilon_\pi + \gamma^2 \frac{|\mathcal{S}| r_{\max}}{1-\gamma} \epsilon_P + \gamma \frac{\zeta_P}{\sqrt{D}} \frac{r_{\max}}{1-\gamma}.
\end{aligned}
$$

The second term and third term of the last inequality are obtained by a derivation process similar to the proof of (34). The last term of the last inequality is built on the error bound assumption of the empirical dynamics model and (35). □

Based on these theorems, we now estimate the action-value gap between the fixed point of the ILB operator and Q_{β^*}.

Theorem 4 (**Action-value gap**). *Suppose Q_{ILB} and Q_{β^*} denote the fixed point of the ILB operator and support-constrained Bellman optimality operator, separately. The action-value gap can be bounded as*

$$
\begin{aligned}
&\left\| Q_{\text{ILB}}(s,a) - Q_{\beta^*}(s,a) \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \frac{\zeta_r}{\sqrt{D}} + \frac{\ell}{1-\gamma} \epsilon_\pi \\
&\quad + \frac{\gamma |\mathcal{S}| r_{\max}}{(1-\gamma)^2} \epsilon_P + \frac{\gamma r_{\max}}{(1-\gamma)^2} \frac{\zeta_P}{\sqrt{D}} + \frac{1}{1-\gamma} |\delta|,
\end{aligned} \tag{37}
$$

where ζ_r, ζ_P are defined in (10) *and* (11)*, ϵ_r, ϵ_P are defined in Theorem 2, δ is defined in the ILB operator.*

Proof. If $a \in \text{Supp}(\beta(\cdot|s))$, we have

$$
\mathcal{T}_{\text{ILB}} Q_{\beta^*}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P} \left[\max_{\tilde{a}' \sim \pi} Q_{\beta^*}(s', \tilde{a}') \right] \tag{38}
$$

Hence,

$$
\begin{aligned}
&|\mathcal{T}_{\text{ILB}} Q_{\beta^*}(s,a) - \mathcal{T}_{\text{Supp}} Q_{\beta^*}(s,a)| \\
&= \Bigg| \gamma \mathbb{E}_{s' \sim P} \left[\max_{\tilde{a}' \sim \pi} Q_{\beta^*}(s', \tilde{a}') \right] \\
&\quad - \gamma \mathbb{E}_{s' \sim P} \left[\max_{a' \in \text{Supp}(\beta(\cdot|s'))} Q_{\beta^*}(s', a') \right] \Bigg| \\
&\leq \gamma \mathbb{E}_{s' \sim P} \left\| Q_{\beta^*}(s', \pi(s')) - Q_{\beta^*}(s', \beta(s')) \right\|_\infty \\
&\leq \gamma \ell \epsilon_\pi + \gamma^2 \frac{|\mathcal{S}| r_{\max}}{1-\gamma} \epsilon_P.
\end{aligned} \tag{39}
$$

The last inequality is obtained by utilizing Theorem 2. Now we can estimate the error bound in the support region of β.

$$
\begin{aligned}
&|Q_{\text{ILB}}(s,a) - Q_{\beta^*}(s,a)| \\
&= |\mathcal{T}_{\text{ILB}} Q_{\text{ILB}}(s,a) - \mathcal{T}_{\text{Supp}} Q_{\beta^*}(s,a)| \\
&= |\mathcal{T}_{\text{ILB}} Q_{\text{ILB}}(s,a) - \mathcal{T}_{\text{ILB}} Q_{\beta^*}(s,a) \\
&\quad + \mathcal{T}_{\text{ILB}} Q_{\beta^*}(s,a) - \mathcal{T}_{\text{Supp}} Q_{\beta^*}(s,a)| \\
&\leq |\mathcal{T}_{\text{ILB}} Q_{\text{ILB}}(s,a) - \mathcal{T}_{\text{ILB}} Q_{\beta^*}(s,a)| \\
&\quad + |\mathcal{T}_{\text{ILB}} Q_{\beta^*}(s,a) - \mathcal{T}_{\text{Supp}} Q_{\beta^*}(s,a)| \\
&\leq \gamma \left| Q_{\text{ILB}}(s,a) - Q_{\beta^*}(s,a) \right| + \gamma \ell \epsilon_\pi + \gamma^2 \frac{|\mathcal{S}| r_{\max}}{1-\gamma} \epsilon_P.
\end{aligned} \tag{40}
$$

We use the contraction property of ILB operator to get the first term in the last inequality, and derive the second term by applying equation (39) directly. By transposing terms, we can get

$$|Q_{\mathrm{ILB}}(s,a) - Q_{\beta^*}(s,a)| \leq \frac{\gamma}{1-\gamma}\ell\epsilon_\pi + \gamma^2\frac{|\mathcal{S}|r_{\max}}{(1-\gamma)^2}\epsilon_P. \tag{41}$$

At last, we consider the error bound in the case of $a \notin \mathrm{Supp}(\beta(\cdot|s))$. Similarly, from the fixed point property and γ-contraction inequality, it follows that

$$\begin{aligned}
&|Q_{\mathrm{ILB}}(s,a) - Q_{\beta^*}(s,a)| \\
&\leq |\mathcal{T}_{\mathrm{ILB}}Q_{\mathrm{ILB}}(s,a) - \mathcal{T}_{\mathrm{ILB}}Q_{\beta^*}(s,a)| \\
&\quad + |\mathcal{T}_{\mathrm{ILB}}Q_{\beta^*}(s,a) - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)| \\
&\leq \gamma\,|Q_{\mathrm{ILB}}(s,a) - Q_{\beta^*}(s,a)| \\
&\quad + |\mathcal{T}_{\mathrm{ILB}}Q_{\beta^*}(s,a) - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)|\,.
\end{aligned} \tag{42}$$

For the second term on the RHS above, we now have

$$\begin{aligned}
&|\mathcal{T}_{\mathrm{ILB}}Q_{\beta^*}(s,a) - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)| \\
&= \left|\min\left\{y_{\mathrm{img}}^{Q_{\beta^*}}, y_{\mathrm{lmt}}^{Q_{\beta^*}}\right\} + \delta - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)\right| \\
&= \max\Big\{\left|y_{\mathrm{img}}^{Q_{\beta^*}} + \delta - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)\right|, \\
&\quad \left|y_{\mathrm{lmt}}^{Q_{\beta^*}} + \delta - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)\right|\Big\}.
\end{aligned} \tag{43}$$

The first case can be estimated by the Theorem 3, and we can see that

$$\begin{aligned}
&\left|y_{\mathrm{img}}^{Q_{\beta^*}} + \delta - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)\right| \\
&\leq \frac{\zeta_r}{\sqrt{D}} + \gamma\ell\epsilon_\pi + \gamma^2\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P + \gamma\frac{\zeta_P}{\sqrt{D}}\frac{r_{\max}}{1-\gamma} + |\delta|.
\end{aligned} \tag{44}$$

For the second case, by the Theorem 2, we have

$$\begin{aligned}
&\left|y_{\mathrm{lmt}}^{Q_{\beta^*}} + \delta - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)\right| \\
&= \left|\max_{\widehat{a}\in\mathrm{Supp}(\beta(\cdot|s))} Q_{\beta^*}(s,\widehat{a}) + \delta - Q_{\beta^*}(s,a)\right| \\
&\leq \left|\max_{\widehat{a}\in\mathrm{Supp}(\beta(\cdot|s))} Q_{\beta^*}(s,\widehat{a}) - Q_{\beta^*}(s,a)\right| + |\delta| \\
&\leq \ell\epsilon_\pi + \gamma\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P + |\delta|.
\end{aligned} \tag{45}$$

Combining (42) to (45) together, we have

$$\begin{aligned}
&|Q_{\mathrm{ILB}}(s,a) - Q_{\beta^*}(s,a)| \\
&\leq \frac{1}{1-\gamma}\,|\mathcal{T}_{\mathrm{ILB}}Q_{\beta^*}(s,a) - \mathcal{T}_{\mathrm{Supp}}Q_{\beta^*}(s,a)| \\
&\leq \max\Big\{\frac{1}{1-\gamma}\frac{\zeta_r}{\sqrt{D}} + \frac{\gamma\ell}{1-\gamma}\epsilon_\pi \\
&\quad + \frac{\gamma^2|\mathcal{S}|r_{\max}}{(1-\gamma)^2}\epsilon_P + \frac{\gamma r_{\max}}{(1-\gamma)^2}\frac{\zeta_P}{\sqrt{D}} + \frac{1}{1-\gamma}|\delta|, \\
&\quad \frac{\ell}{1-\gamma}\epsilon_\pi + \frac{\gamma|\mathcal{S}|r_{\max}}{(1-\gamma)^2}\epsilon_P + \frac{1}{1-\gamma}|\delta|\Big\} \\
&\leq \frac{1}{1-\gamma}\frac{\zeta_r}{\sqrt{D}} + \frac{\ell}{1-\gamma}\epsilon_\pi \\
&\quad + \frac{\gamma|\mathcal{S}|r_{\max}}{(1-\gamma)^2}\epsilon_P + \frac{\gamma r_{\max}}{(1-\gamma)^2}\frac{\zeta_P}{\sqrt{D}} + \frac{1}{1-\gamma}|\delta|.
\end{aligned} \tag{46}$$

Taking together (41) and (46), the theorem is proved. □

According to (41) and (46), we conclude that the error bounds for in-sample and out-of-sample actions are of the same magnitude $\mathcal{O}(r_{\max}/(1-\gamma)^2)$. This result aligns with the conclusion of CQL [Kumar *et al.*, 2020] within the support region. Notably, the theoretical optimal value of delta is 0, based on the assumption of no error in the maximum behavior value. In practice, the optimal value may fluctuate around 0. Nevertheless, $\delta = 0$ consistently provides good performance across all tasks in experiments.

A.2 Experimental Settings

Evaluation Metric

The standard performance indicator is the normalized score, defined as

$$\text{normalized score} = 100 \times \frac{\text{learned score} - \text{random score}}{\text{expert score} - \text{random score}},$$

where the learned score is obtained by the test method, the expert score and random score are two constants taken from the D4RL [Fu *et al.*, 2020] benchmark.

Competitors

In the MuJoCo tasks, we compare our method with prior state-of-the-art methods, including BCQ [Fujimoto *et al.*, 2019], CQL [Kumar *et al.*, 2020], UWAC [Wu *et al.*, 2021], One-step [Brandfonbrener *et al.*, 2021], TD3+BC [Fujimoto and Gu, 2021], IQL [Kostrikov *et al.*, 2022], MCQ [Lyu *et al.*, 2022], CSVE [Chen *et al.*, 2023], OAP [Yang *et al.*, 2023], DTQL [Chen *et al.*, 2024], OAC-BVR [Huang *et al.*, 2024], and TD3-BST [Srinivasan and Knottenbelt, 2024]. BC stands for behavior cloning, with results sourced from OAC-BVR. The CQL results are from IQL, while the performances of BCQ and UWAC are derived from the reproduction experiments of MCQ, as their original experiments were conducted on "-v0" datasets. Results for other algorithms are taken from their respective original papers.

We also conduct evaluation on Maze2d "-v1" tasks to further examine the effectiveness of ILQ. Here, we compare our method with ROMI-BCQ [Wang *et al.*, 2021], BEAR [Kumar *et al.*, 2019], CQL [Kumar *et al.*, 2020], IQL [Kostrikov

et al., 2022], MCQ [Lyu *et al.*, 2022], Diffuser [Janner *et al.*, 2022], and PlanCP [Sun *et al.*, 2023]. As mentioned above, the results of BCQ and CQL are reported from IQL and MCQ, respectively. The performances of other methods are obtained from their original papers.

To further evaluate the proposed ILQ, we conduct additional comparisons on Adroit tasks. The results for TD3+BC [Fujimoto and Gu, 2021] are taken from MCQ [Lyu *et al.*, 2022], as the original paper does not include experiments on Adroit domain. The performance for BCQ [Fujimoto *et al.*, 2019], CQL [Kumar *et al.*, 2020], and IQL [Kostrikov *et al.*, 2022] are sourced from DTQL [Chen *et al.*, 2024], while the results for other methods are derived from their respective original reports.

Parameter Settings

The basic hyperparameters of ILQ are described in Table 3. In addition, hyperparameters of the behavior policy model follow the settings in Diffusion-QL [Wang *et al.*, 2023]. Thus, a 3-layer MLPs with 256 hidden units, 5 diffusion time steps, and corresponding variance schedule [Song *et al.*, 2021] are implemented for the diffusion model. For the dynamics model, we follow the implementation of MOPO [Yu *et al.*, 2020] with 4-layers MLPs with 200 hidden units. We only utilize its reward penalty coefficient 2 for hopper-m, walker2d-mr, and 1 for hopper-r, hopper-mr and walker2d-m tasks. Both of behavior policy and dynamics model are optimized by Adam [Kingma and Ba, 2014] with learning rate 3×10^{-4} and 1×10^{-3}, respectively. In addition, we use a cosine learning schedule for adroit tasks. The main hyperparameters η and δ associated with MuJoCo "-v2" are listed in Table 4, and the main hyperparameters associated with Maze2D "-v1" are listed in Table 5 and Adroit "-v0" are listed in Table 6. All experiments were conducted on the device with $4\times$ Tesla V100 GPUs. Our code required for conducting all experiments will be made publicly available upon acceptance.

Table 3: Basic Hyperparameters of ILQ

Hyperparameters	Value
Actor Architechture	input-256-256-256-output
Critic Architechture	input-256-256-256-1
Optimizer	Adam [Kingma and Ba, 2014]
Batch size	256
(Critic, Actor) Learning rate	$(3 \times 10^{-4}, 1 \times 10^{-4})$ for hopper-r, hopper-mr, walker2d-mr, adroit tasks $(5 \times 10^{-4}, 3 \times 10^{-4})$ for others
Entropy	True for all except adroit tasks
Training steps	10^6
Behavior training steps	3×10^5
Dynamics training epochs	40
Discount factor γ	0.99
Target update rate τ	0.005
Sampling Number M	10

Table 4: Main Hyperparameters on MuJoCo Datasets

Task	Trade-off Factor η	Offset δ
halfcheetah-r	0.95	2
hopper-r	0.9	1
walker2d-r	0.7	1
halfcheetah-m	0.95	1
hopper-m	0.95	-2
walker2d-m	0.9	0.5
halfcheetah-mr	0.95	2
hopper-mr	0.8	-0.5
walker2d-mr	0.9	1
halfcheetah-me	0.6	1
hopper-me	0.4	-0.5
walker2d-me	0.8	1

Table 5: Main Hyperparameters on Maze2D Datasets

Task	Trade-off Factor η	Offset δ
maze2d-u	0.95	-0.5
maze2d-ud	0.95	0
maze2d-m	0.95	0
maze2d-md	0.95	0
maze2d-l	0.95	0
maze2d-ld	0.95	0

A.3 More Experimental Results

Score Curve Results

The score curves for MuJoCo tasks are typically of primary interest, which are illustrated in Fig. 6

A.4 More Sensitive Analyses

Sensitive Analysis of Sampling Number M

We did not finetune the sampling number M, which was set to 10 in all experiments. According to the practical implementation of ILB operator, M implicitly influences the estimation of the maximum behavior value. To assess its impact on performance, we conduct extra sensitivity analyses on the halfcheetah-m, hopper-mr, and walker2d-me tasks. The experimental results indicate that performance on these tasks remains stable when M is set to 5, 10, and 15, respectively, as illustrated in the bar charts in Fig. 7.

Sensitive Analysis of Dynamics Model Accuracy

Our analysis includes: (1) Training epochs: The Gaussian-fitted dynamics model shows stable performance across different epoch settings (Table 7), indicating robustness to this hyperparameter. (2) Training data quantity: We also conducted experiments using 20% reduced training data for the dynamics model. While Table 8 shows a slight performance decrease on most of tasks when using only 80% data, the method maintains reasonable effectiveness.

A.5 More Ablation Studies

We conduct additional ablation studies to assess the effectiveness of both the imagination and limitation components. The results without the imagination component are shown

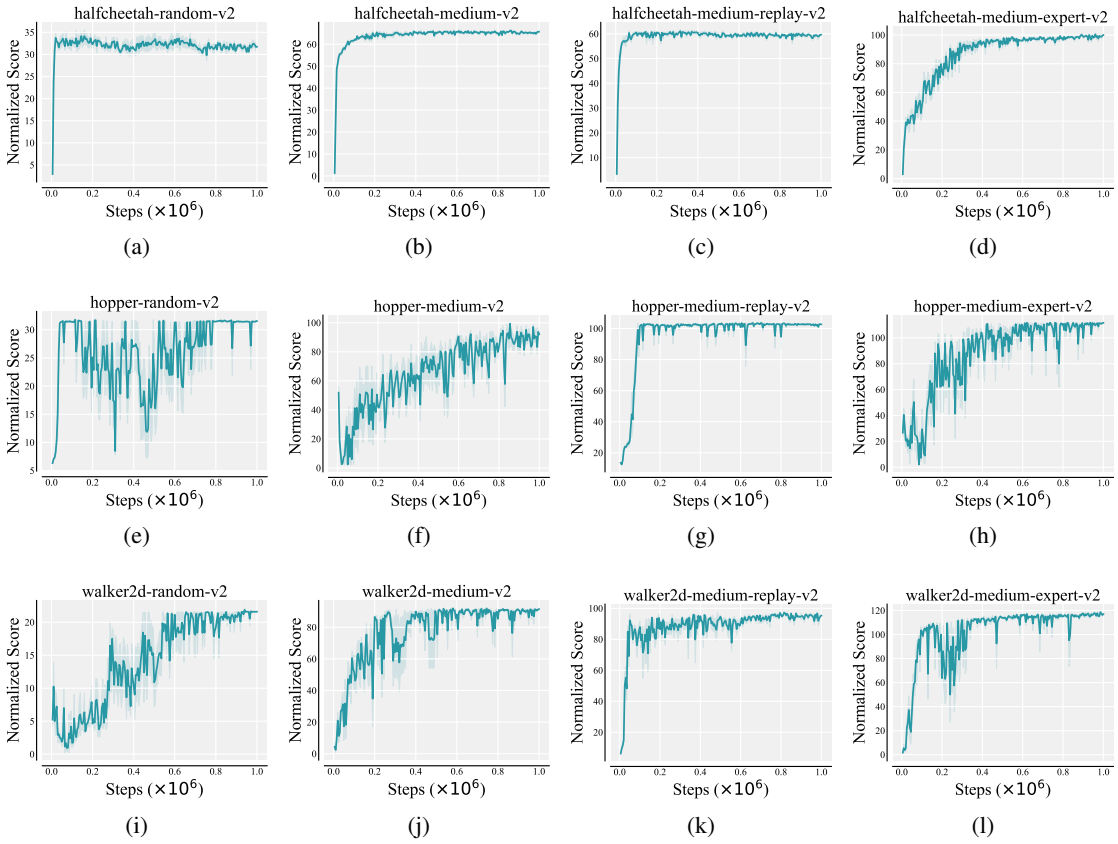


Figure 6: Normailzed score curves of ILQ on MuJoCo "-v2". The results are averaged over 5 different random seeds. Shaded areas indicate standard deviation.

Table 6: Main Hyperparameters on Adroit Datasets

Task	Trade-off Factor η	Offset δ
pen-human	0.8	-1
pen-cloned	0.8	0

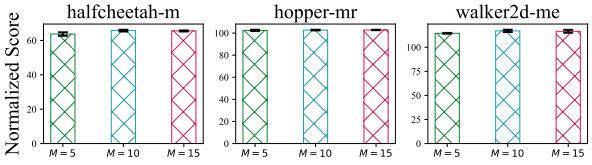


Figure 7: Performances of ILQ under different sampling number M.

Table 7: Changing training epochs of dynamics model. 'ha'=halfcheetah, 'ho'=hopper, 'wa'=walker2d, 'm'=medium, 'mr'=medium-replay. Epochs=40 is the default setting.

Epochs	ha-m	ho-m	wa-m	ha-mr	ho-mr	wa-mr
35	$64.4_{\pm 0.6}$	$93.5_{\pm 5.9}$	$92.0_{\pm 1.1}$	$58.2_{\pm 1.1}$	$102.9_{\pm 0.5}$	$91.5_{\pm 6.0}$
40	$65.7_{\pm 0.5}$	$92.1_{\pm 5.8}$	$91.5_{\pm 0.7}$	$59.6_{\pm 1.0}$	$102.7_{\pm 0.3}$	$95.3_{\pm 1.8}$
45	$65.0_{\pm 0.4}$	$93.9_{\pm 6.0}$	$90.0_{\pm 0.5}$	$58.1_{\pm 0.3}$	$102.4_{\pm 0.5}$	$93.3_{\pm 0.7}$
50	$64.1_{\pm 0.2}$	$90.9_{\pm 9.1}$	$89.4_{\pm 0.8}$	$59.2_{\pm 0.6}$	$103.1_{\pm 0.7}$	$97.9_{\pm 0.7}$

Table 8: Reducing training data for dynamics model.

Data ratio	ha-m	ho-m	wa-m	ha-mr	ho-mr	wa-mr
All	$65.7_{\pm 0.5}$	$92.1_{\pm 5.8}$	$91.5_{\pm 0.7}$	$59.6_{\pm 1.0}$	$102.7_{\pm 0.3}$	$95.3_{\pm 1.8}$
80%	$64.7_{\pm 0.0}$	$98.4_{\pm 4.9}$	$89.8_{\pm 3.3}$	$58.9_{\pm 1.6}$	$102.6_{\pm 0.2}$	$88.0_{\pm 5.5}$

in Fig. 8. While competitive performance is achieved on some tasks, such as in Fig. 8(b) and (d), performance significantly drops on other tasks. As illustrated in Fig. 8(a), (c), (e), and (f), performance deteriorates sharply, with the normalized score curves exhibiting highly oscillatory behavior. As discussed previously, this is due to the use of the maximum behavior value as the target for OOD actions, which introduces uncontrollable bias and ultimately hampers policy improvement. These results highlight the critical role of the imagination component in providing calibrated target values within proper constraints.

In further studies on the limitation component, slightly better performance is observed when the imagination component is exclusively used, as shown in Fig. 9(b). This suggests that, in certain cases, the imagination component can offer reliable guidance. However, as seen in Fig. 9(a), (e), and (f), performance drops significantly, and in some cases, policies completely collapse, as evidenced in the hopper-medium-expert

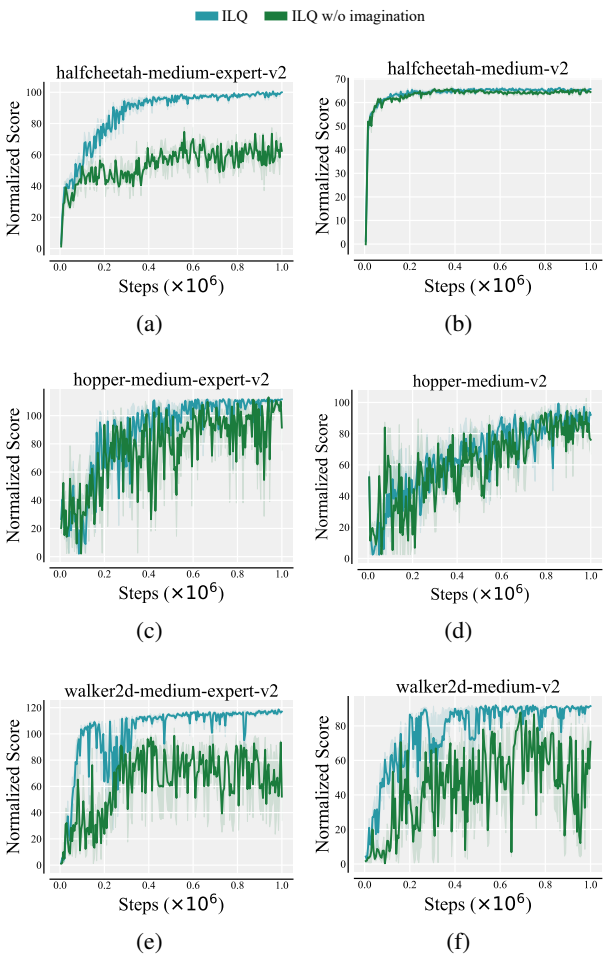


Figure 8: Performance comparison of the ILQ algorithm with and without the imagined value y^Q_{img} in the target value.

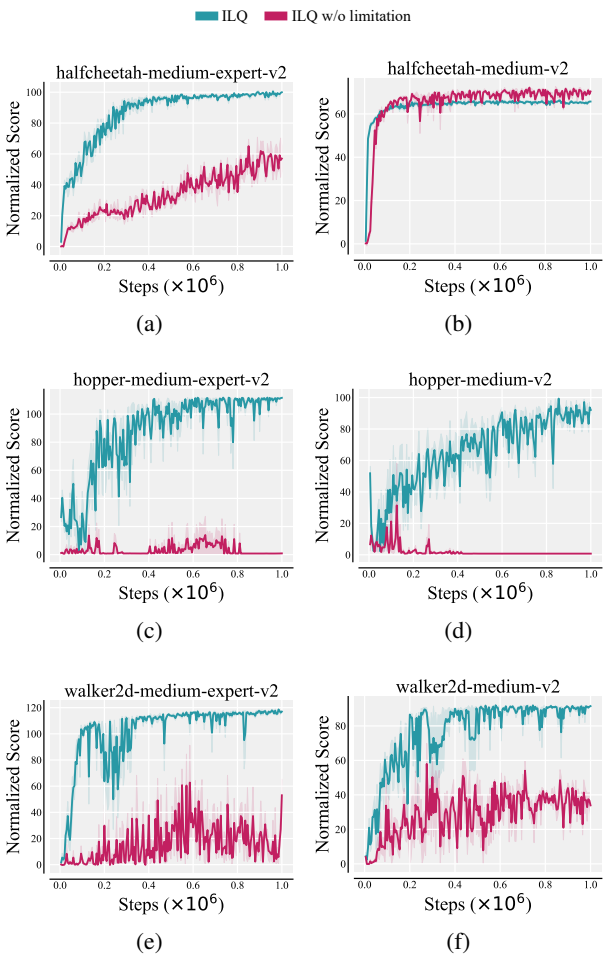


Figure 9: Performance comparison of the ILQ algorithm with and without the limitation value y^Q_{lmt} in the target value.

and hopper-medium tasks (Fig. 9(c) and 9(d)). This underscores the importance of the limitation component in preventing overly optimistic estimates. These comprehensive studies demonstrate that both components are essential for accurate OOD action-value estimation.

A.6 Further Verification in Q-value

ILQ estimates OOD Q-values by preserving the imagined values as much as possible while adhering to the maximum behavior value constraint. This approach ensures appropriately optimistic estimates, as shown in the section of Introduction, thus avoiding the deliberate pessimism of value regularization methods.

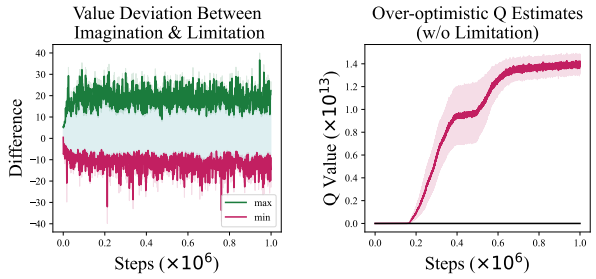
To further understand the interaction between the imagined value and the maximum behavioral value, we examined the difference between them, i.e., $y^Q_{\text{img}} - y^Q_{\text{lmt}}$. Figure 10(a) illustrates how the range (cyan area) of this difference evolves during training. For the upper boundary curve (green), where the imagined value exceeds the limiting value, we retain the limiting value. Conversely, for the lower boundary curve (red), where the limiting value is higher than the imagined value, we retain the imagined value. This suggests a mutually constraining relationship between the two components. As seen in Fig. 4, the absence of the imagined value leads to a significant performance decrease. Meanwhile, as shown in Fig. 5, unconstrained imagining can result in false optimistic estimates, potentially causing the policy to collapse. Notably, in the scenario depicted in Fig. 5(b), this false optimistic estimation even grows exponentially, reaching a Q-value of 10^{13}, as shown in Fig. 10(b).

A.7 Limitations of Theoretical Results

Our theoretical analyses - consistent with most theoretical works in both online and offline RL - assumes tabular MDPs, as formal guarantees under neural network function approximation remain challenging. We will note it as a direction for future research.

A.8 Computational Cost

Regarding computational efficiency, we provide detailed comparisons of computation costs across different methods (measured on a Tesla V100 server) in Table 9, which shows ILQ achieves competitive efficiency relative to baselines.



(a) (b)

Figure 10: (a) illustrates the evolving range of the difference between the imagined value and the limiting value during training. (b) shows the exponential growth of false optimistic value estimations in the w/o limitation scenario on the hopper-medium-v2 task.

Table 9: Training time per 100 steps on hopper-medium task.

	BEAR	CQL	IQL	MCQ	DTQL	ILQ
Time(s)	3.06	2.14	1.01	3.05	2.24	2.21

A.9 Discussion of Lipschitz Continuity Assumption

The Lipschitz condition on reward functions is commonly adopted in offline RL theoretical analyses [Huang *et al.*, 2024], though it represents a strong practical assumption. Mathematically, any continuously differentiable function on a compact set satisfies the Lipschitz condition. In practice, we can verify this by checking: (1) whether the real-world reward function is sufficiently smooth (continuously differentiable), and (2) whether the action space is bounded (compact). For our experimental environments: (1) In MuJoCo and Adroit tasks, the reward functions are continuous and actions are bounded within $[-1, 1]^{|\mathcal{A}|}$, so the condition typically holds. (2) For Maze tasks with sparse rewards, the assumption theoretically fails.