Imagination-Limited Q-Learning for Offline Reinforcement Learning

Wenhui Liu[1], **Zhijian Wu**[1], **Jingchao Wang**[1], **Dingjiang Huang**[1*], **Shuigeng Zhou**[2]
[1]East China Normal University
[2]Fudan University
{whliu_14, zjwu_97, jcwang}@stu.ecnu.edu.cn, djhuang@dase.ecnu.edu.cn, sgzhou@fudan.edu.cn

Abstract

Offline reinforcement learning seeks to derive improved policies entirely from historical data but often struggles with over-optimistic value estimates for out-of-distribution (OOD) actions. This issue is typically mitigated via policy constraint or conservative value regularization methods. However, these approaches may impose overly constraints or biased value estimates, potentially limiting performance improvements. To balance exploitation and restriction, we propose an Imagination-Limited Q-learning (ILQ) method, which aims to maintain the optimism that OOD actions deserve within appropriate limits. Specifically, we utilize the dynamics model to imagine OOD action-values, and then clip the imagined values with the maximum behavior values. Such design maintains reasonable evaluation of OOD actions to the furthest extent, while avoiding its over-optimism. Theoretically, we prove the convergence of the proposed ILQ under tabular Markov decision processes. Particularly, we demonstrate that the error bound between estimated values and optimality values of OOD state-actions possesses the same magnitude as that of in-distribution ones, thereby indicating that the bias in value estimates is effectively mitigated. Empirically, our method achieves state-of-the-art performance on a wide range of tasks in the D4RL benchmark.

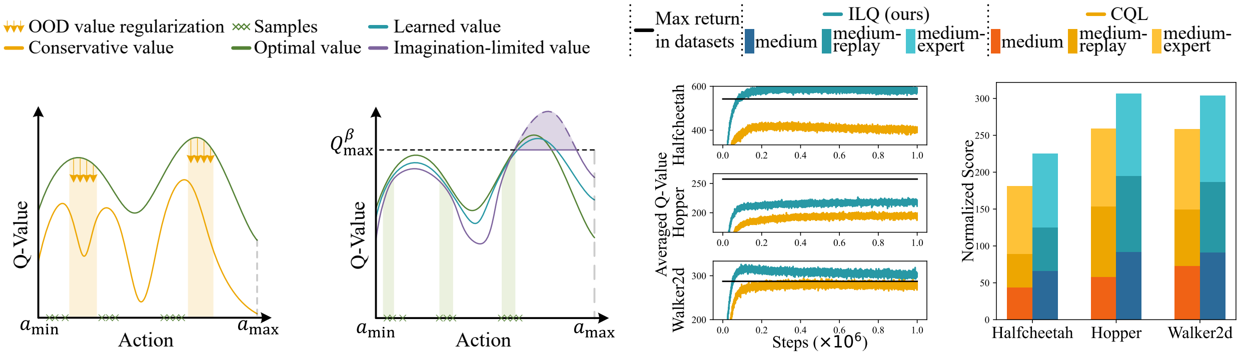
1 Introduction

Offline Reinforcement Learning (RL) [Lange *et al.*, 2012; Fujimoto *et al.*, 2019] is designed to learn optimal policies purely from a static dataset previously collected by an unknown policy (behavior policy). By eliminating the need for online interaction with environments, it offers dual benefits. On the one hand, it can mitigate the expensive costs [Gu *et al.*, 2017] and potential risks [Sallab *et al.*, 2017] associated with trial-and-error learning in real-world applications; on the other hand, it can be leveraged to enhance the generalization ability and scalability of RL models when the logged data is massive and diverse. However, the offline learning paradigm unavoidably incurs distributional shifts [Levine *et al.*, 2020] of state-action visitation between the learned policy and the behavior policy, which makes it difficult to correctly assess out-of-distribution (OOD) action-values. Especially, over-optimistic estimates may even invalidate the learned policy.

To address this challenge, two main classes of technical routes are commonly employed in the model-free approaches [Prudencio *et al.*, 2023]. 1) Policy constraint: It usually explicitly restricts the gap between the learned and behavior policy. The batch-constrained Q-learning (BCQ) [Fujimoto *et al.*, 2019] was devised to restrict the action space via adding perturbations on a state-conditioned behavior model. Kumar et al. [Kumar *et al.*, 2019] proposed BEAR to reduce maximum mean discrepancy between the learned policy and the behavior one. Subsequently, different methods corresponding to other metrics have been proposed, such as KL divergence for BRAC [Wu *et al.*, 2019] and mean squared error for TD3-BC [Fujimoto and Gu, 2021; Srinivasan and Knottenbelt, 2024], which similarly seeks to steer the policy closer to actions in the dataset. However, these learned models are limited to the neighborhood of the behavior policy hindering their performance, especially when the dataset is collected by poor policies. 2) Value regularization: It aims to utilize value regularizations to suppress OOD action-values. The conservative Q-learning (CQL) [Kumar *et al.*, 2020] was designed to penalize the expectation of OOD action-values, thus mitigating optimistic estimates outside the dataset. Kostrikov et al. [Kostrikov *et al.*, 2022] introduced implicit Q-learning (IQL) to estimate value function through expectile regression to implicitly depress OOD action-values. The MCQ [Lyu *et al.*, 2022] was proposed to regularize OOD action-values with the maximum behavior value. And the OAC-BVR [Huang *et al.*, 2024] was developed to regard the difference between the Q-function and the behavior value as a regularization term. While these methods effectively limit OOD optimism, they also introduce uncontrollable bias into Q-value estimates, as illustrated in Fig. 1(a). Especially, the Q-values under CQL exhibit noticeable bias of pessimism over all MuJoCo tasks, as shown in Fig. 1(c).

To mitigate value bias while maintaining appropriate restrictions on over-optimism, we propose an Imagination-Limited Q-learning (ILQ) method. The insight of our method is straightforward: For in-sample state-action pairs, we adopt

[*]Corresponding author



(a) (b) (c) (d)

Figure 1: (a) illustrates the fundamental principle of value regularization methods. While effectively suppressing OOD action-values, it may introduce uncontrolled bias in estimations. In contrast, instead of indiscriminately suppressing OOD action-values, ILQ, depicted in (b), envisions reasonable values (purple line) and then appropriately limits potential over-estimations using the maximum behavior value $Q^{\beta}_{\max}$ (black dashed line), resulting in more appropriate policy evaluation (cyan line). (c) demonstrates that Q-value estimations of CQL across MuJoCo "-v2" tasks are notably compromised, falling well below maximum returns (black line) in datasets. Conversely, ILQ maintains reasonably optimistic Q-value estimations in anticipation of superior policies. Finally, (d) shows that ILQ's ultimate performance is significantly improved, particularly in medium tasks.

the standard Bellman backup based on in-sample transitions to estimate their values. For OOD state-action pairs, instead of blindly applying value regularizations to suppress their action-values, we envision *what the values would be without any restrictions*. The one-step bootstrapped values under an imaged dynamics model would be reasonable estimations, ideally approximating the ground truth when the imaged model closely matches the environment. However, errors in the model fitting are inevitable, and optimistic estimates may still exist. Therefore, we need to further consider *how to appropriately limit the imagination values*. We employ the maximum behavior value as the ceiling of the imagined one. Specifically, if the imagined value is less than the maximum value, it is maintained; otherwise, the maximum behavior value is applied. Figure 1(b) illustrates the intuition behind our method. This design ensures a more reasonable evaluation with appropriate constraints on OOD actions, thereby avoiding unnecessary value suppression and improving the generalization ability of the learned policy.

We validate the effectiveness of ILQ both theoretically and empirically. We prove that the policy evaluation Bellman operator of ILQ is a contraction operator under tabular Markov Decision Processes (MDPs), ensuring convergence. Particularly, we analyze the action-value gap between the fixed point obtained by our method and that obtained by the Bellman optimality equation, demonstrating that the error bound of OOD action-values can reach the same order of magnitude as in-distribution ones. Empirically, our method maintains reasonably optimistic Q-values compared to conservative Q-learning (Fig. 1(c)) and achieves state-of-the-art performance across a wide range of tasks in the standard benchmark.

2 Background

Reinforcement Learning (RL) is commonly modeled as a Markov Decision Process (MDP), characterized by a tuple $(\mathcal{S}, \mathcal{A}, r, P, \rho_0, \gamma)$ [Sutton and Barto, 2018], where $\mathcal{S}$ is state space, $\mathcal{A}$ is action space, $r : \mathcal{S} \times \mathcal{A} \to [-r_{\max}, r_{\max}]$ is reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is transition dynamics, ρ_0 is probability distribution of initial states, and $\gamma \in [0, 1)$ is discount factor. The RL agent takes actions on the environment according to its policy, defined as $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

During the learning process, evaluating the expected return of a state-action pair (s, a) under a policy π, called the action-value (or Q-value) $Q(s, a)$, is essential. Theoretically, it can be obtained by $\mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a]$. In practice, it is usually approximated by minimizing the Bellman residual $\mathbb{E}[(Q(s, a) - (\mathcal{T}Q)(s, a))^2]$, where $\mathcal{T}$ is the Bellman optimality operator defined as

$$(\mathcal{T}Q)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \sim \pi} Q(s', a') \right]. \quad (1)$$

Generally, a delayed approximator Q^- is applied in the above target $\mathcal{T}Q$ for training stability [Mnih *et al.*, 2015].

2.1 Offline Reinforcement Learning

In offline RL, the agent is no longer allowed to interact with the environment and learns policies exclusively from a limited static dataset $\mathcal{D}$, which is typically gathered from an unknown behavior policy β. The dataset $\mathcal{D}$ is usually represented by a set of transition tuples $\{(s, a, r, s')\}$. The goal of the agent remains to maximize the expectation of cumulative rewards. Without online correction, erroneously optimistic estimates of OOD action-values are inevitable [Levine *et al.*, 2020]. These biases can cause the learned policy to favor incorrect OOD actions, potentially leading to it failure.

The representative value regularization method is CQL [Kumar *et al.*, 2020], which adds penalties for OOD action-values to the standard Q-value update objective, as follows:

$$\alpha_{\text{CQL}} \Big(\mathbb{E}_{s \sim \mathcal{D}, a \sim \mu(\cdot|s)}[Q(s, a)] - \mathbb{E}_{s \sim \mathcal{D}, a \sim \beta(\cdot|s)}[Q(s, a)] \Big) + \frac{1}{2} \mathbb{E}\Big[\big(Q(s, a) - \mathcal{T}Q(s, a) \big)^2 \Big],$$

where $\mu(\cdot|s)$ is a distribution to produce OOD actions, α_{CQL} is a hyperparameter to adjust the degree of conservatism.

3 Related Work

3.1 Model-free Offline RL

Recent advancements in offline RL have focused on addressing the challenges posed by OOD actions. Importance sampling methods [Nachum *et al.*, 2019] have been developed to correct evaluation under distributional shifts, but they often suffer from high variance. Policy constraint methods [Kumar *et al.*, 2019; Wu *et al.*, 2021; Fujimoto and Gu, 2021; Li *et al.*, 2023; Chen *et al.*, 2024] are employed to limit the deviation of learned policies; however, their performance tends to degrade when the behavior policy is suboptimal.

Our method aligns more closely with value regularization approaches. For instance, CQL [Kumar *et al.*, 2020] and CSVE [Chen *et al.*, 2023] directly penalize OOD action-values, effectively controlling over-optimism but often resulting in overly pessimistic estimates. Methods like MCQ [Lyu *et al.*, 2022] and OAC-BVR [Huang *et al.*, 2024] attempt to relax restrictions by assigning maximum behavior values or behavior values, respectively, to OOD action values. However, this introduces uncontrolled value bias for OOD actions. Although MCQ additionally employs policy constraint weighting to mitigate this bias in practical implementations, it offers limited theoretical guarantees for OOD action-value estimates. In contrast, our proposed method preserves more honest estimates of OOD action-values within the limitation of maximum behavior values, and offers theoretical guarantees for its value estimates.

3.2 Model-based Offline RL

Model-based methods aim to enhance collected datasets by generating synthetic trajectories using learned dynamics models. Various strategies have been proposed to effectively leverage these synthetic data, including uncertainty quantification [Ovadia *et al.*, 2019; Kidambi *et al.*, 2020; Yu *et al.*, 2020; Diehl *et al.*, 2021], conservative value estimation [Yu *et al.*, 2021], representation balancing [Lee *et al.*, 2021], and adversarial learning [Bhardwaj *et al.*, 2023]. In contrast, our proposed ILQ avoids trajectories generation, relying solely on the dynamics model to produce one-step subsequent states and rewards of in-sample states for estimating imagined OOD action-values. While ILQ utilizes the dynamics model, it remains fundamentally a model-free learning framework and circumvents challenges of error accumulation associated with longer trajectories in model-based methods.

4 Imagination-Limited Q-learning Method

We start by elucidating our novel Imagination-Limited Bellman (ILB) operator in Subsection 4.1. Subsequently, we elaborate on its practical implementation details and theoretical analysis in Subsection 4.2 and 4.3, respectively. And the Imagination-Limited Q-learning (ILQ) algorithm is ultimately summarized in Subsection 4.4.

4.1 Imagination-Limited Bellman Operator

In online RL, researchers typically use the Bellman optimality operator Eq. (1) to evaluate policies. However, in offline settings, the absence of online corrections makes policy evaluation highly susceptible to OOD over-optimism [Levine *et al.*, 2020] under the standard operator. Existing value regularization methods primarily focus on directly restricting the OOD action-values [Kumar *et al.*, 2020; Lyu *et al.*, 2022; Chen *et al.*, 2023; Huang *et al.*, 2024], which introduces uncontrollable bias in value estimates and lacks theoretical guarantees for OOD actions.

We argue that establishing reasonable estimates for out-of-distribution state-actions should take precedence, followed by the imposition of suitable restrictions, rather than directly employing value regularization. To achieve this goal, we introduce a novel Imagination-Limited Bellman operator, defined as follows.

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

$$\begin{aligned}&\mathcal{T}_{\mathrm{ILB}}Q(s,a)\\ &\quad=\begin{cases} r(s,a)+\gamma\mathbb{E}_{s'\sim P}\big[\max\limits_{\tilde{a}'\sim\pi}Q(s',\tilde{a}')\big], & \textit{if}\ \beta(a|s)>0\\ \min\left\{y^{Q}_{\mathrm{img}},y^{Q}_{\mathrm{lmt}}\right\}+\delta, & \textit{otherwise.}\end{cases}\end{aligned} \tag{2}$$

where β is the behavior policy,

$$y^{Q}_{\mathrm{img}}=\widehat{r}(s,a)+\gamma\mathbb{E}_{\widehat{s}'\sim\widehat{P}(\cdot|s,a)}\left[\max_{\tilde{a}'\sim\pi}Q(\widehat{s}',\tilde{a}')\right], \tag{3}$$

and

$$y^{Q}_{\mathrm{lmt}}=\max_{\widehat{a}\in\mathrm{Supp}(\beta(\cdot|s))}Q(s,\widehat{a}) \tag{4}$$

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.*

Here is the insight behind the proposed ILB operator. For an in-sample state-action pair (s,a), i.e., $\beta(a\mid s)>0$, we have its corresponding transition (s,a,r,s') in the dataset, allowing us to apply the standard Bellman operator without any obstacles. However, for an out-of-sample state-action pair (s,a^{oos}), the standard Bellman backup cannot be applied solely due to the absence of its successor state and reward. To address this, one could utilize empirical dynamics model to predict the next state $\widehat{s}'$ and reward $\widehat{r}$ and obtain the imagination value y^{Q}_{img} as Eq. (3), which provides a relatively accurate approximation for an OOD state-action. Nevertheless, it may still result in optimistic estimates because of fitting errors. To tackle this issue, we use the maximum in-distribution action-value Eq. (4) as the upper limit for the imagined value. This design offers dual benefits: First, it maximally maintains the imagined value, reducing estimation bias on OOD actions; second, the maximum behavior value ensures that there is always an in-distribution action-value greater than or equal to the OOD ones, encouraging the policy to more likely favor in-distribution actions during the actor improvement process.

We analyze the ILB operator's properties and demonstrate that the ILB operator exhibits the same γ-contraction property as the standard Bellman operator, ensuring convergence in policy evaluation. The proof is provided in the Appendix.

Theorem 1 (**Convergence**)**.** *The ILB operator defined in Eq.* (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.*

4.2 Practical Implementation of Imagination and Limitation Value

Imagination Value

We fit the environment dynamics to derive empirical transfer kernel $\widehat{P}$ and reward function $\widehat{r}$ in the simplest manner using

$$\max_{\widehat{T}_\psi} \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\log \widehat{T}_\psi(r, s' \mid s, a)\right], \tag{5}$$

where $\widehat{T}_\psi$ stands for both $\widehat{P}$ and $\widehat{r}$ for brevity, and is represented by a multivariate Gaussian distribution with parameters ψ practically. We then obtain the imagined value y^{Q}_{img}.

Limitation Value

In fact, the behavior policy $\beta(\cdot \mid s)$ in Eq. (4) is unknown and needs to be empirically modeled. In light of the expressiveness of diffusion models [Ho *et al.*, 2020], we fit the behavior policy using a conditional diffusion model. Specifically, it is constructed via a reverse diffusion chain, formulated as

$$\text{Diff}_\omega(a \mid s) := \mathcal{N}(a^K; 0, I)\prod_{k=1}^{K} p_\omega(a^{k-1} \mid a^k, s) \tag{6}$$

where superscript k denotes the diffusion timestep, $a := a^0$ is the final sampled action, a^k, $k = 1, \cdots, K-1$, are latent variables, $a^K \sim \mathcal{N}(0, I)$ is Gaussian noise. Typically, $p_\omega(a^{k-1} \mid a^k, s)$ is modeled as a Gaussian distribution $\mathcal{N}\left(a^{k-1}; \mu_\omega(a^k, s, k), \Sigma_\omega(a^k, s, k)\right)$ with the covariance matrix $\Sigma_\omega(a^k, s, k) = \beta_k I$ and the mean defined as

$$\mu_\omega(a^k, s, k) = \frac{1}{\sqrt{\alpha_k}}\left(a^k - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\xi_\omega(a^k, s, k)\right), \tag{7}$$

where β_k is the variance schedule, $\alpha_k := 1 - \beta_k$, $\bar{\alpha}_k := \prod_{i=1}^{k}\alpha_i$, and $\xi_\omega(\cdot)$ is the noise prediction network with parameters ω. The conditional diffusion model is optimized by maximizing the evidence lower bound, which can be simplified [Ho *et al.*, 2020] to minimize the following objective

$$\min_{\omega} \mathbb{E}_{\substack{k\sim\mathcal{U},\xi\sim\mathcal{N}(0,I)\\(s,a)\sim\mathcal{D}}}\left\|\xi - \xi_\omega(\sqrt{\bar{\alpha}_k}a + \sqrt{1-\bar{\alpha}_k}\xi, s, k)\right\|^2, \tag{8}$$

where $\mathcal{U}$ is an uniform distribution over $\{1, \cdots, K\}$. Similar diffusion behavior modeling methods are also applied in other works [Wang *et al.*, 2023; Hansen-Estruch *et al.*, 2023].

Accordingly, we adopt the limitation value as shown in the following equation:

$$y^{Q}_{\text{lmt}} = \max_{\substack{\widehat{a}_m\sim\text{Diff}_\omega(\cdot\mid s)\\ m=1,\cdots,M}} Q\left(s, \widehat{a}_m\right), \tag{9}$$

where M is the number of sampled actions. Due to possible errors in the fitting process, this may result in a deviation between the estimated value Eq. (9) and the true value Eq. (4). In order to offset this gap, we introduced the hyperparameter δ in definition Eq. (2), typically set to a small absolute value.

4.3 Theoretical Analysis

We now theoretically discuss the action-value gap between the fixed point in Theorem 1 and the Bellman optimality value. Before proceeding further, 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.

In addition, 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 hold with high probability $\geq 1 - \zeta$, $\zeta \in (0, 1)$,

$$\left\|\widehat{r}(s, a) - r(s, a)\right\|_1 \leq \zeta_r/\sqrt{D}, \tag{10}$$

$$\left\|\widehat{P}(\cdot \mid s, a) - P(\cdot \mid s, a)\right\|_1 \leq \zeta_P/\sqrt{D}, \tag{11}$$

where D is the constant related to the dataset size, ζ_r and ζ_P are constants related to ζ.

We begin by analyzing the Bellman optimality value gap between the learned policy and behavior policy.

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

$$|Q_{\beta^*}(s, \pi(s)) - Q_{\beta^*}(s, \beta(s))| \leq \ell\epsilon_\pi + \gamma\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P, \tag{12}$$

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

Accordingly, we could prove the error bound between the imagination value and Bellman optimality value.

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

$$\begin{aligned}&\left|y^{Q_{\beta^*}}_{\text{img}} - Q_{\beta^*}(s, a)\right|\\ &\leq \frac{\zeta_r}{\sqrt{D}} + \gamma\ell\epsilon_\pi + \gamma^2\frac{|\mathcal{S}|r_{\max}}{1-\gamma}\epsilon_P + \gamma\frac{\zeta_P}{\sqrt{D}}\frac{r_{\max}}{1-\gamma}.\end{aligned} \tag{13}$$

Based on above theorems, we can 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*

$$\begin{aligned}&\|Q_{\text{ILB}}(s, a) - Q_{\beta^*}(s, a)\|_\infty\\ &\leq \frac{1}{1-\gamma}\frac{\zeta_r}{\sqrt{D}} + \frac{\ell}{1-\gamma}\epsilon_\pi\\ &\quad + \frac{\gamma|\mathcal{S}|r_{\max}}{(1-\gamma)^2}\epsilon_P + \frac{\gamma r_{\max}}{(1-\gamma)^2}\frac{\zeta_P}{\sqrt{D}} + \frac{1}{1-\gamma}|\delta|,\end{aligned} \tag{14}$$

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

According to Theorem 4, we conclude that 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. All proofs are detailed in the Appendix.

4.4 The ILQ Algorithm

In deep RL, the Q-function is commonly approximated by a neural network with parameters θ, while the corresponding target network has parameters θ^{-}. As described in the background, it can be optimized by minimizing the temporal difference (TD) loss $\mathbb{E}_{(s,a,r,s')}[(Q_\theta(s,a) - \mathcal{T}Q_{\theta^-}(s,a))^2]$. Intuitively, under the ILB operator we developed, the corresponding loss function can be constructed as

$$\mathbb{E}_{(s,a,r,s')}\left[\left(Q_\theta(s,a) - \mathcal{T}_{\text{ILB}}Q_{\theta^-}(s,a)\right)^2\right]. \tag{15}$$

Since in-sample and out-of-sample state-action pairs have different TD targets, as defined by our ILB operator, a common way is to split the above loss function into a weighted sum of in-sample and out-of-sample components as follows

$$\begin{aligned}\mathcal{L}_Q(\theta) = {} & \eta\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \mathcal{T}_{\text{ILB}}Q_{\theta^-}(s,a)\right)^2\right] \\ & + (1-\eta)\mathbb{E}_{\substack{s\sim\mathcal{D}\\ a^{\text{oos}}\sim u(\cdot|s)}}\left[\left(Q_\theta(s,a^{\text{oos}}) - \mathcal{T}_{\text{ILB}}Q_{\theta^-}(s,a^{\text{oos}})\right)^2\right],\end{aligned} \tag{16}$$

where η is a trade-off factor, (s, a^{oos}) refers to the out-of-sample state-action pair. The $u(\cdot \mid s)$ is any distribution that can produce out-of-sample actions. In practice, we directly treat policy π as the sampling distribution u.

We also incorporate the double network [Hasselt, 2010] to reduce overestimation, which is widely utilized in both online [Mnih *et al.*, 2015] and offline RL [Fujimoto and Gu, 2021]. Therefore, combining the definition of ILB operator Eq. (2), we have the target, for in-sample transition (s, a, r, s'), as:

$$\mathcal{T}_{\text{ILB}}Q_{\theta^-}(s,a) = r(s,a) + \gamma \min_{j=1,2} \mathbb{E}_{\tilde{a}'\sim\pi_\phi(\cdot|s')}\left[Q_{\theta_j^-}(s',\tilde{a}')\right], \tag{17}$$

where π_ϕ is the learned policy represented by a neural network with parameters ϕ and $Q_{\theta_j^-}$ is the j-th target network. Similarly, for out-of-sample (s, a^{oos}), the target is formulated as:

$$\mathcal{T}_{\text{ILB}}Q_{\theta^-}(s,a^{\text{oos}}) = \min\left\{y^Q_{\text{img}}, y^Q_{\text{lmt}}\right\} + \delta, \tag{18}$$

where

$$y^Q_{\text{img}} = \widehat{r}(s,a^{\text{oos}}) + \gamma \min_{j=1,2} \mathbb{E}_{\substack{\widehat{s}'\sim\widehat{P}(\cdot|s,a^{\text{oos}})\\ \tilde{a}'\sim\pi_\phi(\cdot|\widehat{s}')}}\left[Q_{\theta_j^-}(\widehat{s}',\tilde{a}')\right], \tag{19}$$

$$y^Q_{\text{lmt}} = \min_{j=1,2} \max_{\substack{\widehat{a}_m\sim\text{Diff}_\omega(\cdot|s)\\ m=1,\cdots,M}} Q_{\theta_j^-}(s,\widehat{a}_m). \tag{20}$$

Here Eq. (20) is obtained by coupling Eq. (9).

During policy improvement, we adopt the same objective as in vanilla SAC [Haarnoja *et al.*, 2018] to optimize the actor network without any complex design, as follows:

$$\max_\phi \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi(\cdot|s)}\left[\min_{j=1,2} Q_{\theta_j}(s,a) - \alpha\log\pi_\phi(a \mid s)\right], \tag{21}$$

where α is a multiplier for the entropy.

Combining above steps, our method is derived, with its pseudo-code presented in Algorithm 1.

Algorithm 1 Imagination-Limited Q-Learning (ILQ)

```
Require: The offline dataset $\mathcal{D}$, number of iterations $N$, discount factor $\gamma$, target network update rate $\tau$, trade-off factor $\eta$, and offset parameter $\delta$.
Initialize critic networks $Q_{\theta_1}$, $Q_{\theta_2}$, actor network $\pi_\phi$, target networks $Q_{\theta_1^-}$, $Q_{\theta_2^-}$ with $\theta_i^- \leftarrow \theta_i, i = \{1,2\}$.
// Pre-train the dynamics model and behavior policy
Train the dynamics model $\widehat{T}_\psi(s', r \mid s, a)$ via (5).
Train the diffusion model $\text{Diff}_\omega(\cdot \mid s)$ for modeling the behavior policy by optimizing (8).
// Policy training
for step $n = 1$ to $N$ do
    Sample a mini-batch of transitions $\mathcal{B} = \{(s,a,r,s')\}$ from dataset $\mathcal{D}$.
    Compute target value for $(s,a)$ in $\mathcal{B}$ as (17).
    Sample OOD actions conditioned on sates in $\mathcal{B}$ via $\pi_\phi$, and calculate target value according to (18) via the pre-trained behavior policy and dynamics model.
    Update parameters $\theta_i, i = 1,2$ for each critic network via minimizing (16).
    Update actor $\phi$ via (21).
    Update the target networks
        $\theta_i^- \leftarrow \tau\theta_i + (1-\tau)\theta_i^-, i = 1,2.$
end for
```

5 Experiments

In this section, we empirically validate the effectiveness of our method ILQ. 1) We demonstrate the superiority of ILQ over existing methods by comparing performance across a series of tasks. 2) We conduct sensitivity analyses on the hyperparameters involved in ILQ, confirming the stability of the proposed method. 3) We then perform ablation experiments on both imagination and limitation components to verify their impacts. 4) We also delve into the Q-value estimations to further validate the effectiveness of the two components designed; details are in the Appendix due to space constraints.

5.1 Experimental Settings

We evaluate ILQ on the D4RL [Fu *et al.*, 2020] benchmark. The commonly used domain is Gym MuJoCo "-v2", including halfcheetah, hopper, and walker2d tasks at four levels: random (r), medium (m), medium-replay (mr), and medium-expert (me). We also assess ILQ on Maze2D "-v1" domain, which offers three layouts with two reward types, i.e., umaze (u), umaze-dense (ud), medium (m), medium-dense (md), large (l), and large-dense (ld). In addition, comparisons on several adroit "-v0" tasks are conducted. Due to page limitations, descriptions of compared algorithms and extra experimental results have been relocated to the Appendix. Details of hyperparameters settings and implementation specifics are also provided in the Appendix to ensure reproducibility.

5.2 Performance Comparison

In comparing ILQ with state-of-the-art methods on MuJoCo tasks, as shown in Table 1, ILQ performs significantly better than policy constraint methods (BCQ [Fujimoto *et al.*,

Table 1: Comparison of normalized average scores for ILQ and existing state-of-the-art methods on MuJoCo tasks over the final 10 evaluations. Experiments are conducted using 5 different random seeds. The highest score is **bolded**.

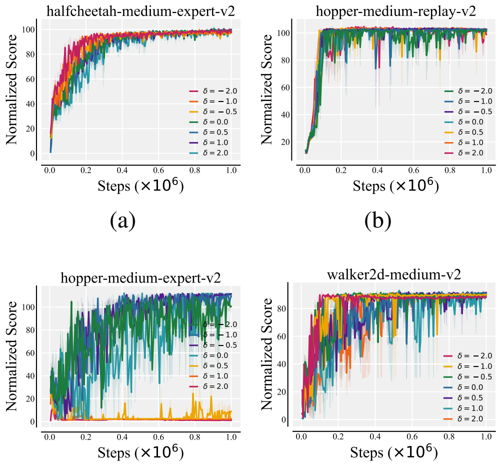
Task Name	BC	BCQ	CQL	UWAC	One-step	TD3+BC	IQL	MCQ	CSVE	OAP	DTQL	OAC-BVR	TD3-BST	ILQ(Ours)
halfcheetah-r	2.2	2.2	17.5	2.3	3.7	11.0	13.1	28.5	26.7	24.0	-	31.1	-	**31.7** ± **0.7**
hopper-r	3.7	7.8	7.9	2.7	5.6	8.5	7.9	**31.8**	27.0	8.8	-	7.4	-	31.6 ± 0.2
walker2d-r	0.2	4.9	5.1	2.0	5.2	1.6	5.4	17.0	6.1	5.1	-	9.8	-	**21.6** ± **0.1**
halfcheetah-m	42.4	46.6	44.0	42.2	48.6	48.3	47.4	64.3	48.6	56.4	57.9	52.2	62.1	**65.7** ± **0.5**
hopper-m	53.4	59.4	58.5	50.9	56.7	59.3	66.3	78.4	99.4	82.0	99.6	95.0	**102.9**	92.1 ± 5.8
walker2d-m	66.9	71.8	72.5	75.4	80.3	83.7	78.3	91.0	82.5	85.6	89.4	86.0	90.7	**91.5** ± **0.7**
halfcheetah-mr	34.9	42.2	45.5	35.9	38.6	44.6	44.2	56.8	54.8	53.4	50.9	48.3	53.0	**59.6** ± **1.0**
hopper-mr	28.1	60.9	95.0	25.3	94.1	60.9	94.7	101.6	91.7	98.5	100.0	95.3	101.2	**102.7** ± **0.3**
walker2d-mr	19.2	57.0	77.2	23.6	49.3	81.8	73.9	91.3	78.5	84.3	88.5	77.3	90.4	**95.3** ± **1.8**
halfcheetah-me	60.4	95.4	91.6	42.7	91.7	90.7	86.7	87.5	93.1	83.4	92.7	93.1	**100.7**	**100.0** ± **0.4**
hopper-me	51.5	106.9	105.4	44.9	83.1	98.0	91.5	111.2	95.2	85.9	109.3	96.5	110.3	**111.6** ± **0.6**
walker2d-me	96.7	107.7	108.8	96.5	112.9	110.1	109.6	114.2	109.0	111.1	110.0	112.0	109.4	**117.0** ± **1.2**
MuJoCo total	459.6	662.8	729.0	444.4	669.8	698.5	719.0	873.6	812.6	778.5	-	804.0	-	**920.4**

Table 2: Comparison of normalized scores on Maze2D and Aroit datasets. The scores are also averaged over the final 10 evaluations across 5 different random seeds.

Task Name	ROMI-BCQ	BEAR	CQL	IQL	MCQ	Diffuser	PlanCP	ILQ(Ours)
maze2d-u	**139.5**	65.7	18.9	47.4	81.5	113.9	116.4	91.9 ± 26.0
maze2d-ud	98.3	32.6	14.4	48.9	107.8	-	-	**116.2** ± **15.4**
maze2d-m	82.4	25.0	14.6	34.9	106.8	121.5	128.5	**163.6** ± **31.4**
maze2d-md	102.6	19.1	30.5	47.1	112.7	-	-	**137.8** ± **9.2**
maze2d-l	83.1	81.0	16.0	58.6	111.2	123.0	130.9	**198.5** ± **23.8**
maze2d-ld	124.0	133.8	46.9	75.4	118.5	-	-	**152.8** ± **10.4**
Maze2D total	629.9	357.2	141.3	312.3	638.5	-	-	**860.8**
Task Name	BCQ	TD3+BC	CQL	IQL	MCQ	DQL	DTQL	ILQ(Ours)
pen-human	68.9	64.8	35.2	71.5	68.5	72.8	64.1	**77.3** ± **7.9**
pen-cloned	44.4	49	27.2	37.3	49.4	57.3	81.3	**85.6** ± **10.2**
Adroit total	113.3	113.8	62.4	108.8	117.9	130.1	145.4	**162.9**

2019], UWAC [Wu *et al.*, 2021], One-step [Brandfonbrener *et al.*, 2021], TD3+BC [Fujimoto and Gu, 2021], and OAP [Yang *et al.*, 2023]) in a wide range of random- and medium-level tasks. This is because policy constraint methods restrict the learned policy to be within a neighborhood of the behavior policy. Although TD3-BST [Srinivasan and Knottenbelt, 2024] introduces fine-grained weighting on the constraints of TD3+BC to enhance performance, it remains inferior to ILQ overall. The value regularization methods (CQL [Kumar *et al.*, 2020], MCQ [Lyu *et al.*, 2022], CSVE [Chen *et al.*, 2023], OAC-BVR [Huang *et al.*, 2024]) show higher scores in average. ILQ continues to show performance beyond them in $^{11}/_{12}$ tasks. DTQL [Chen *et al.*, 2024] integrates implicit value regularization and policy constraints to enhance performance, but it still falls short compared to ILQ, especially in halfcheetah tasks.

According to Table 2, neither policy constraint approaches nor value regularization approaches performed well enough on Maze2D tasks. Although ROMI-BCQ [Wang *et al.*, 2021] achieves advanced performance on maze2d-u by utilizing a reverse dynamics model, it performs mediocrely on other tasks. Diffuser [Janner *et al.*, 2022] and PlanCP [Sun *et al.*, 2023] leverage diffusion models to improve their planning capabilities in maze2d-u, while still lag behind ILQ on maze2d-m and maze2d-l, further demonstrating stitching abilities of ILQ. The experimental results on Adroit tasks demonstrate that our method continues to outperform others, highlighting its applicability across different tasks.



(a) (b)

(c) (d)

Figure 2: Performances of ILQ under different values of offset parameter δ.

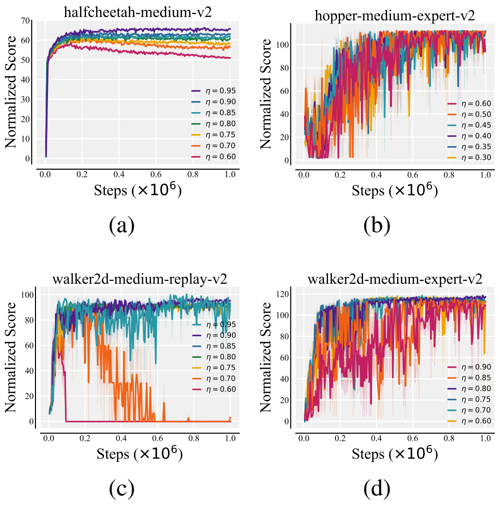
5.3 Parameter Study

Offset parameter δ

To assess impacts of parameter δ, we conduct sensitivity analyses by varying δ across $\{-2, -1, -0.5, 0, 0.5, 1, 2\}$. We evaluate the performance on halfcheetah-me, hopper-mr, hopper-me, and walker2d-m, where each experiment was run over 3 random seeds. Figure 2 shows that the score curve remains stable as δ changes. Overall, performance is slightly lower when δ is negative compared to when it is positive. Notably, a negative δ always ensures a safe learned policy. However, when δ is positive, meaning the limit exceeds the maximum behavior value a little bit, there may exists policy failure, as illustrated in Fig. 2(c). These indeed validate our argument that overly restricting OOD values could inhibit potential performance gains, and taking the maximum behavior value ($\delta = 0$) as a ceiling of the imagined value is consistently a solid choice.

Trade-off Factor η

To evaluate the impact of the trade-off factor η, we conduct sensitivity analyses by varying η around its optimal value.



(a) (b)

(c) (d)

Figure 3: Performances of ILQ under different values of trade-off factor η.

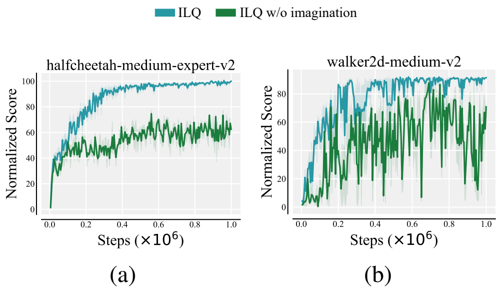
According to results in Fig. 3(a) and 3(c), ILQ achieves robust performance for all variations of η around 0.9, making it a typically safe choice for medium and medium-replay tasks. This suggests that ILQ should place greater trust in in-sample value estimates when evaluating policies. In medium tasks, the Q-value of OOD actions generated by the learned policy rarely reach the maximum behavior value. Consequently, assigning too high a weight to these OOD action-values, i.e., a smaller η, may lead to misplaced trust in OOD actions and ultimately cause policy failure, as illustrated in Fig. 3(c). In medium-expert tasks, the in-sample data comprises a mixture of medium and expert levels, and the value of the OOD actions generated by the learned policy can approach the maximum behavior value. Therefore, the value estimate needs to be more balanced between in-sample and out-of-sample. In this case, the optimal value of η is usually around 0.6 for all medium-expert tasks. The overall results in Fig. 3 show that ILQ maintains stable performance when η varies around its optimal parameter.

5.4 Ablation Study

To understand the contribution of each component in our OOD target action-value Eq. (18), we conduct an ablation study. This study evaluates the impact of removing either the imagination component or the limitation value from the target value. All experiments are run over 3 random seeds. More results can be found in the Appendix.

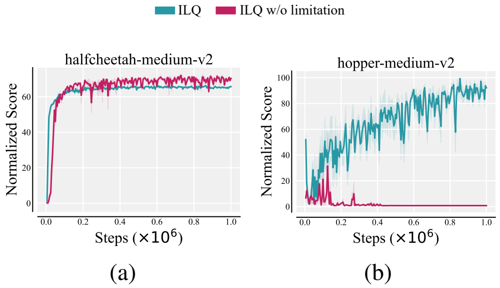
Without Imagination

In this part, we assess the performance of ILQ without the imagination component y^Q_{img}, indicating solely the maximum behavior value is considered as the regularization target. The results are illustrated in Fig. 4. As shown in Fig. 4(a) and (b), the performance degrades significantly and the curve of normalized score exhibits a very oscillatory behavior. This is because directly using the maximum behavior value as the target values for OOD actions introduces uncontrollable bias and ultimately impairs policy improvement. This demonstrates the necessity of the imagination component for providing calibrated target values under the limitation.



(a) (b)

Figure 4: Performance comparison of the ILQ algorithm with and without the imagined value y^Q_{img} in the target value.



(a) (b)

Figure 5: Performance comparison of the ILQ algorithm with and without the limitation value y^Q_{lmt} in the target value.

Without Limitation

Here we exclude the limitation component y^Q_{lmt} from the target value, relying solely on y^Q_{img} for learning. This approach is intended to evaluate the importance of the limitation component. In one specific case, Fig. 5(a), performance increased when relying solely on the imagination value, indicating that the imagination component can sometimes provide highly reliable guidance. However, in most tasks, performance dropped significantly, with some policies in hopper-m task Fig. 5(b) collapsing completely. This underscores the importance of the limitation component in preventing the incorrectly optimistic estimates. These studies suggest both of the two components play critical role on OOD estimates.

6 Conclusion

In conclusion, the Imagination-Limited Q-learning (ILQ) method effectively mitigates bias of value estimations by maintaining reasonable evaluations of OOD action-values within appropriate limits. Specifically, it utilizes a dynamics model to help generate imagined values and capping these with the maximum behavior values for OOD actions, while standard target values for in-distribution ones. Theoretical analysis confirms the convergence of ILQ and demonstrates that the error bound between estimated and optimal values for OOD actions is comparable to that for in-distribution actions, thereby enhancing performance improvements. Empirical results show that ILQ achieves state-of-the-art performances on a wide range of tasks in the D4RL benchmark. We hope this work can provide new insights into the value estimates for offline RL.

Acknowledgments

This work was partially supported by the National Science Foundation of China (62072185 and U1711262).

References

[Bhardwaj *et al.*, 2023] Mohak Bhardwaj, Tengyang Xie, Byron Boots, Nan Jiang, and Ching-An Cheng. Adversarial model for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 1245–1269, 2023.

[Brandfonbrener *et al.*, 2021] David Brandfonbrener, Will Whitney, Rajesh Ranganath, and Joan Bruna. Offline rl without off-policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, pages 4933–4946, 2021.

[Chen *et al.*, 2023] Liting Chen, Jie Yan, Zhengdao Shao, Lu Wang, Qingwei Lin, Saravanakumar Rajmohan, Thomas Moscibroda, and Dongmei Zhang. Conservative state value estimation for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 35064–35083, 2023.

[Chen *et al.*, 2024] Tianyu Chen Chen, Zhendong Wang, and Mingyuan Zhou. Diffusion policies creating a trust region for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 1–22, 2024.

[Diehl *et al.*, 2021] Christopher Diehl, Timo Sievernich, Martin Krüger, Frank Hoffmann, and Torsten Bertram. UMBRELLA: Uncertainty-aware model-based offline reinforcement learning leveraging planning. *arXiv preprint arXiv:2111.11097*, 2021.

[Fu *et al.*, 2020] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[Fujimoto and Gu, 2021] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20132–20145, 2021.

[Fujimoto *et al.*, 2019] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

[Gu *et al.*, 2017] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE International Conference on Robotics and Automation*, pages 3389–3396, 2017.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[Hansen-Estruch *et al.*, 2023] Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

[Hasselt, 2010] Hado Hasselt. Double q-learning. In *Advances in Neural Information Processing Systems*, volume 23, pages 1–9, 2010.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[Huang *et al.*, 2024] Longyang Huang, Botao Dong, Wei Xie, and Weidong Zhang. Offline reinforcement learning with behavior value regularization. *IEEE Transactions on Cybernetics*, 54(6):3692–3704, 2024.

[Janner *et al.*, 2022] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pages 9902–9915, 2022.

[Kidambi *et al.*, 2020] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21810–21823, 2020.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kostrikov *et al.*, 2022] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, pages 1–11, 2022.

[Kumar *et al.*, 2019] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, volume 32, pages 1–11, 2019.

[Kumar *et al.*, 2020] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.

[Lange *et al.*, 2012] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pages 45–73. Springer, 2012.

[Lee *et al.*, 2021] Byung-Jun Lee, Jongmin Lee, and Kee-Eung Kim. Representation balancing offline model-based reinforcement learning. In *International Conference on Learning Representations*, pages 1–22, 2021.

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[Li *et al.*, 2023] Jianxiong Li, Xianyuan Zhan, Haoran Xu, Xiangyu Zhu, Jingjing Liu, and Ya-Qin Zhang. When data geometry meets deep function: Generalizing offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, pages 1–35, 2023.

[Lyu *et al.*, 2022] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1711–1724, 2022.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Nachum *et al.*, 2019] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32, pages 1–11, 2019.

[Ovadia *et al.*, 2019] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 1–12, 2019.

[Prudencio *et al.*, 2023] Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Sallab *et al.*, 2017] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:1704.02532*, 2017.

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, pages 1–36, 2021.

[Srinivasan and Knottenbelt, 2024] Padmanaba Srinivasan and William Knottenbelt. Offline reinforcement learning with behavioral supervisor tuning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1–9, 2024.

[Sun *et al.*, 2023] Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Nobel, Mykel J Kochenderfer, and Mac Schwager. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. In *Advances in Neural Information Processing Systems*, volume 36, pages 80324–80337, 2023.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Wang *et al.*, 2021] Jianhao Wang, Wenzhe Li, Haozhe Jiang, Guangxiang Zhu, Siyuan Li, and Chongjie Zhang. Offline reinforcement learning with reverse model-based imagination. In *Advances in Neural Information Processing Systems*, volume 34, pages 29420–29432, 2021.

[Wang *et al.*, 2023] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, pages 1–17, 2023.

[Wu *et al.*, 2019] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[Wu *et al.*, 2021] Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 11319–11328, 2021.

[Yang *et al.*, 2023] Qisen Yang, Shenzhi Wang, Matthieu Gaetan Lin, Shiji Song, and Gao Huang. Boosting offline reinforcement learning with action preference query. In *International Conference on Machine Learning*, pages 39509–39523, 2023.

[Yu *et al.*, 2020] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 14129–14142, 2020.

[Yu *et al.*, 2021] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 28954–28967, 2021.

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*

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

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

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

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

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

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,

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

It can be rewritten as

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

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., $|r(s,a)|\le r_{\max}$. 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., $|r(s,\tilde{a}_1)-r(s,\tilde{a}_2)|\le\ell\|\tilde{a}_1-\tilde{a}_2\|_\infty$, $\forall s\in\mathcal{S}$ 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

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

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

hold with high probability $\ge 1-\zeta$, $\zeta\in(0,1)$, 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*

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

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

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

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

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

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:*

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

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

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

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

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

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*

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

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,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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× 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

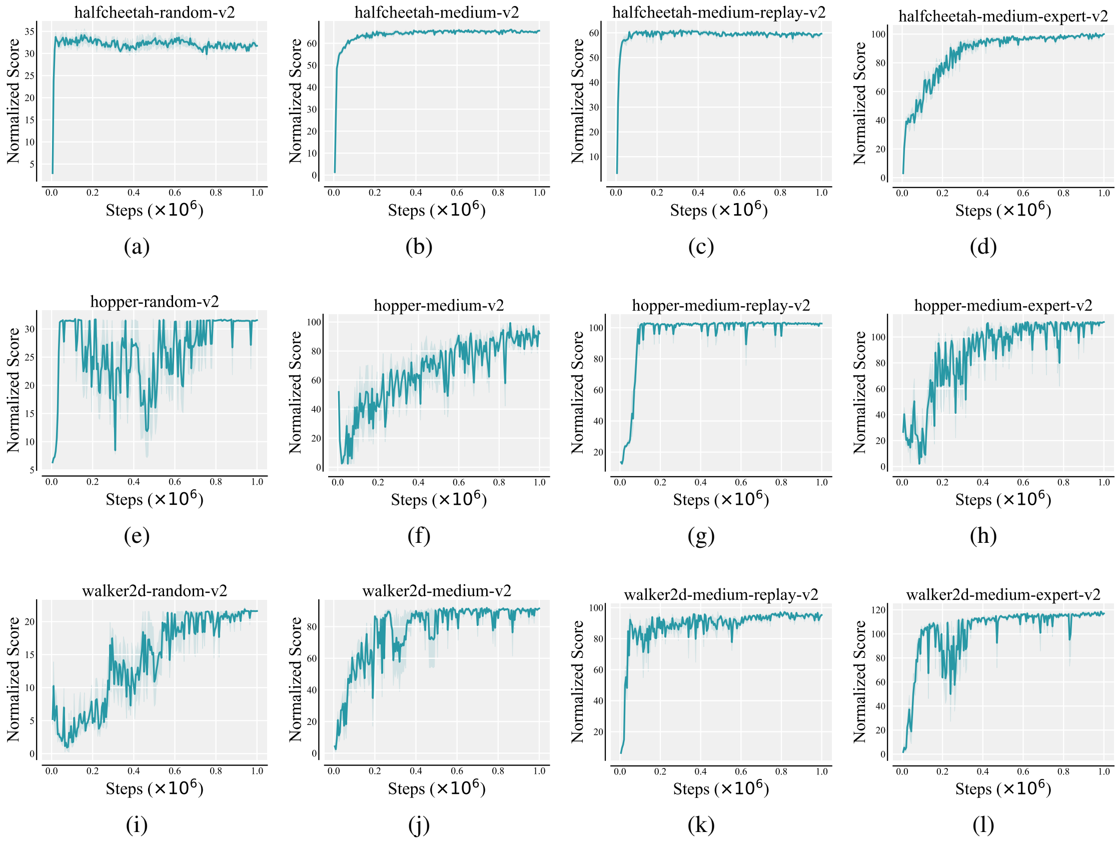
We did not finetune the sampling number M, which was set to 10 in all experiments. According to the practical implementation of ILB operator, M implicitly influences the estimation of the maximum behavior value. To assess its impact on performance, we conduct extra sensitivity analyses on the halfcheetah-m, 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



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

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

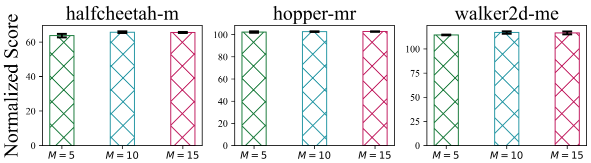


Figure 7: Performances of ILQ under different sampling number M.

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.

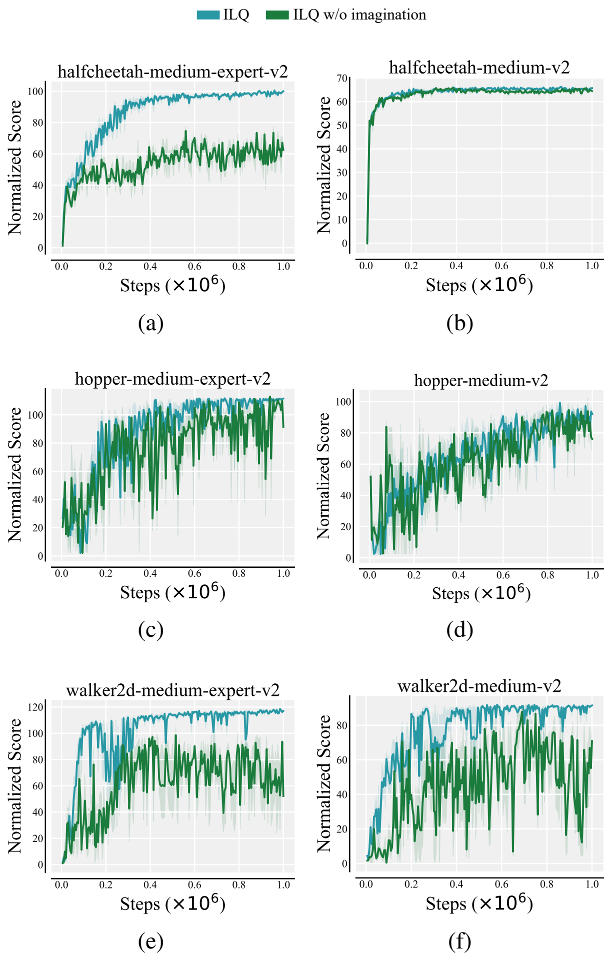
Table 7: Changing training epochs of dynamics model. ‘ha’=halfcheetah, ‘ho’=hopper, ‘wa’=walker2d, ‘m’=medium, ‘mr’=medium-replay. Epochs=40 is the default setting.

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 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

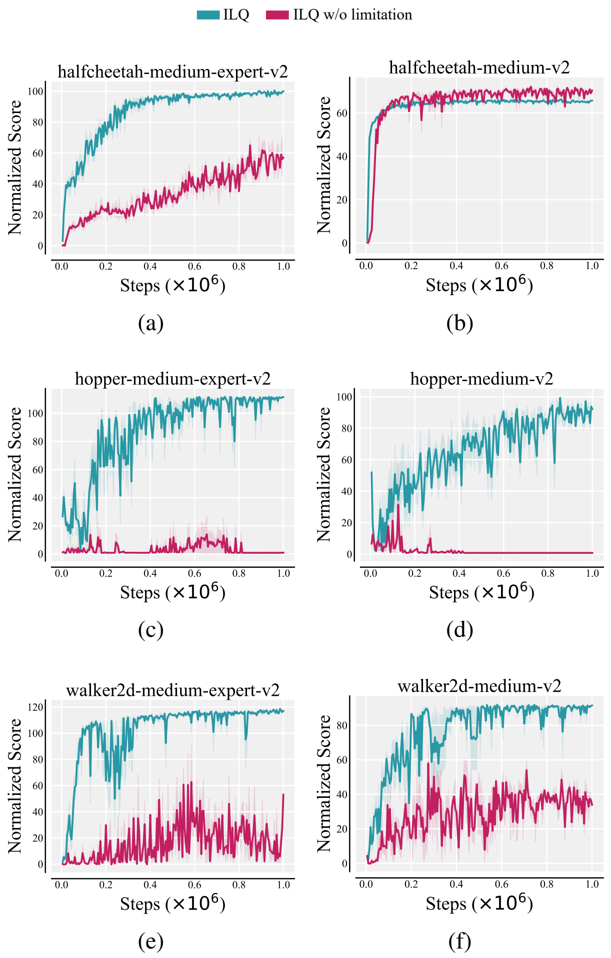


(a) (b)

(c) (d)

(e) (f)

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



(a) (b)

(c) (d)

(e) (f)

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.

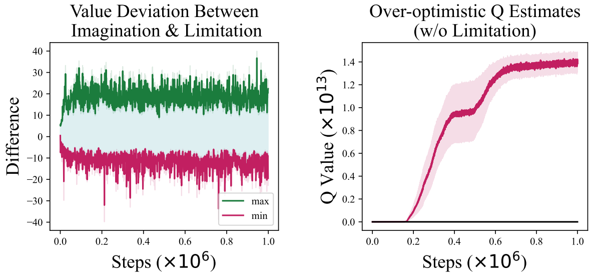
To further understand the interaction between the imagined value and the maximum behavioral value, we examined the difference between them, i.e., $y^{Q}_{\mathrm{img}} - y^{Q}_{\mathrm{lmt}}$. Figure 10(a) illustrates how the range (cyan area) of this difference evolves during training. For the upper 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.