2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Safe Policy Improvement Based on ε-Bisimulation

Yuan Zhuang

[1] Nanjing university

Abstract. This paper studies the Safety Policy Improvement (SPI) problem in Batch Reinforcement Learning (Batch RL), which aims to train a policy from a fixed dataset without environment interaction, while ensuring its performance is no worse than the baseline policy used for data collection. Most existing SPI methods impose constraints on training, but these constraints often make the training overly conservative, especially in complex environment where satisfying the constraints requires large amounts of data. Meanwhile, ε-bisimulation, a general state abstraction technique, has been widely used to enhance sample efficiency in reinforcement learning (RL). However, applying ε-bisimulation transforms the original dataset into one over abstracted observation, which typically violates the assumption of independent and identically distributed (i.i.d.) samples required by existing SPI methods. To address this limitation, this paper proposes a constraint for policy learning that incorporates ε-bisimulation to improve sample efficiency while ensuring that the learned policy satisfies the SPI requirement.

Keywords: Batch Reinforcement Learning, Safe policy Improvement, ε-bisimulation.

1 Introduction

Recently, Reinforcement Learning (RL) has achieved remarkable success in tasks involving long-term planning, global optimization, and sequential decision-making [1]. Representative breakthroughs include the Deep Q-Network (DQN), which enabled end-to-end control in Atari games; AlphaGo, which defeated the world Go champion Lee Sedol [2]; and Reinforcement Learning from Human Feedback (RLHF) [3], a key technique for aligning large language models such as ChatGPT with human intentions. Classical RL relies on an agent interacting with its environment through trial and error, gradually learning an optimal behavior policy based on reward signals from past interactions. However, this paradigm is often impractical in domains where trial-and-error incurs high costs or involves significant risks, such as healthcare, industrial control, and finance. Batch Reinforcement Learning (Batch RL) provides an alternative by learning effective policies directly from fixed data, without requiring further interaction with the environment [4, 5].

Safe Policy Improvement (SPI) is a fundamental topic in Batch RL, focusing on learning a policy from fixed data that performs at least as well as the baseline policy that generated it [6-9]. SPI also holds significant practical value. For example, policies

may need to be deployed simultaneously across many independent devices (e.g., widespread software updates on smartphones), where failures can lead to extremely high repair costs. Moreover, policy evaluation may require a long period of time (e.g., in crop management or clinical trials), during which deploying a bad policy could cause severe consequences. Research on SPI can greatly reduce the risks associated with such scenarios, ensuring the stability and consistency of deployed policies.

Most SPI approaches focus on the model-based RL paradigm to address this fundamental problem in the context of infinite-horizon discounted Markov decision processes (MDPs) [6-9]. These approaches impose constraints that restrict training, allowing policy learning only when the constraints are met. Typically, the constraints depend on the available data, as sufficient samples are needed to accurately capture the agent-environment interaction dynamics. As a result, the learned optimal policy from these samples is more likely to perform well in the true environment, thus outperforming the baseline policy. A fundamental constraint is that training is permitted only if the number of samples for each state-action pair in the MDP exceeds a threshold. Based on this, several algorithms have been proposed to improve policy learning efficiency. However, since the available data is usually limited, existing SPI methods often result in conservative training. As a result, improving sample efficiency has become a critical challenge for enabling the practical application of SPI in the true environment.

At the same time, ε-bisimulation (also known as approximation stochastic bisimulation), a general state abstraction technique, reduces the size of an MDP while ensuring that the loss compared to the original problem remains bounded[10, 11]. In recent years, this technique has been applied within model-based RL paradigms to reduce learning complexity and improve sample efficiency [12-14]. ε-Bisimulation can maps states with similar transition probabilities and rewards into an abstracted state (or abstract observation). This approach is naturally suited to improving sample efficiency in the SPI problem, as it enables the sharing of samples across grouped states, which helps alleviate the issue of insufficient data and better meets the constraints. However, since shared samples are not perfectly equivalent, the policies learned from these samples often differ from those learned without sharing. As a result, the existing constraints must be adjusted to account for these differences.

This paper investigates the SPI problem with ε-bisimulation, which involves an ε-bisimulation function that maps similar states to abstract states. This allows the fixed dataset to be transformed into an abstract one, thereby improving sample efficiency in RL. However, such abstract datasets often violate the independent and identically distributed (i.i.d.) assumption, which most state-of-the-art SPI methods rely on, making it challenging to directly apply existing SPI techniques for policy learning. To address this issue, we propose a policy learning constraint that supports non-i.i.d. abstract datasets. The main contribution of this work is the design of this constraint, which can be used when applying reinforcement learning with ε-bisimulation to improve sample efficiency. Additionally, we prove that the policy learned under this constraint will outperform the baseline policy with high probability.

2 Preliminaries

2.1 MDPs and Reinforcement Learning

We briefly introduce the notations for Markov Decision Processes (MDPs) and Reinforcement Learning (RL). For a comprehensive introduction, we refer readers to the relevant literature [1, 15].

An MDP is defined by $M = (S, A, T, R, s_{init}, \gamma)$, where S is the state space, A is the action space, $R: S \times A \to \mathbb{R}$ is the reward function, where $R(s, a)$ represents the reward the agent receives after taking action a in state s, s_{init} is the initial state, and $\gamma \in [0,1)$ is the discount factor. The true environment is modelled as an unknown finite MDP $M^* = (S, A, R, T^*, \gamma, s_{init})$ with unknown transition probability T^*.

A policy is defined as $\pi: S \to \Delta(A)$, where $\Delta(A)$ denotes a probability distribution over the action set A. The value function of a policy π in MDP M is defined as $V_M^\pi(s) = E_{\pi,M}[\sum_{t\geq 0} \gamma^t R(s_t, a_t) | s_0 = s, a_t \sim \pi(s_t)]$, representing the expected discounted return when starting from state s and following π. The value of M is denoted as $\rho(\pi, M) = V_M^\pi(s_{init})$. The optimal policy over all policies $\Pi: \{\pi: S \to \Delta(A)\}$ is $\pi^* = \arg\max_{\pi\in\Pi} \rho(\pi, M)$, while the Π'-optimal policy over a subset $\Pi' \in \Pi$ is $\pi^*_{\Pi'} = \arg\max_{\pi\in\Pi'} \rho(\pi, M)$. The value function is upper bounded by $V_{max} \leq \frac{R_{max}}{1-\gamma}$, where R_{max} is the maximum reward.

In this paper, we consider the batch RL setting [4], where the algorithm does its best at learning a policy from a fixed set of experience. Given a dataset of transitions $D = \{(s_j, a_j, r_j, s'_j) | j \in [1, N]\}$, we denote by $N_D(s, a)$ the state-action pair counts, and by $N_D(s, a, s')$ the number of transitions from (s, a) to s'. A vanilla batch RL approach, referred to as Basic RL, adopts a model-based manner [16] by explicitly constructing a Maximum Likelihood Estimation (MLE) MDP $\widehat{M} = (S, A, R, \hat{T}, s_{init}, \gamma)$, where the estimated transition probability is given by:

$$\forall s, s' \in S, a \in A, \hat{T}(s'|s, a) = \frac{N_D(s, a, s')}{N_D(s, a)} \quad (1)$$

Once the model $\widehat{M}$ is constructed, the optimal policy can be derived through dynamic programming on $\widehat{M}$ [17], Q-learning with experience replay until convergence [18], etc.

If the estimated MDP $\widehat{M}$ closely approximates M^*, the optimal policy learned from $\widehat{M}$ may perform optimal in the true environment. However, datasets are often limited, particularly in high-risk fields such as healthcare and finance. With insufficient data, the learned policy may lack robustness in state-action pairs with fewer samples, potentially leading to high-risk decisions.

2.2 Safe Policy Improvement

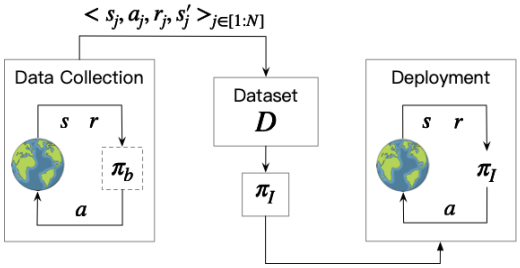

Fig.1. Illustration of the SPI problem in Batch RL

The Safe Policy Improvement (SPI) problem focuses on guaranteeing the performance of policies learned from fixed datasets in the true environment [6-9]. Fig. 1 illustrates the framework of the SPI problem. This problem typically assumes a baseline policy π_b, which generates the fixed dataset $D = \{(s_j, a_j, r_j, s'_j)|j \in [1, N]\}$. The goal of SPI is to learn a policy π_I from D such that, with probability at least $1 - \delta$, its performance deviates from that of the behavior policy π_b by no more than an admissible performance loss ζ:

$$\rho(\pi_I, M^*) \geq \rho(\pi_b, M^*) - \zeta \tag{2}$$

Constraints in Policy Learning. We consider a representative constraint for SPI [6], which specifies δ and ζ in advance and derives a corresponding minimum sample size requirement. Specifically, learning is allowed only if the number of samples $N_D(s, a)$ for every state-action pair $(s, a) \in S \times A$ exceeds a threshold $N_\wedge$, which is derived based on the specified δ and ζ as follows:

$$\forall(s, a) \in S \times A, N_D(s, a) \geq N_\wedge = \frac{8V_{max}^2}{\zeta^2(1-\gamma)^2} \log \frac{2|S||A|2^{|S|}}{\delta} \tag{3}$$

This work marks a milestone in the development of SPI. Earlier methods typically imposed MDP-level constraints, often assuming uniformly distributed environment dynamics, which limited their applicability [19, 20]. In contrast, this work introduces state-action-pair-wise constraints, significantly expanding SPI to a broader class of MDPs. This idea has inspired many follow-up studies. For example, SPIBB builds on this theory to guarantee safe policy improvement even with partial state-action coverage [7, 21, 22]. Leveraging this property, later works have further extended SPI to more challenging settings, such as partially observable MDP [9].

Analysis of the i.i.d. Assumption for Samples. The constraints in this work, as well as those in other SPI methods not discussed here, are all derived under the i.i.d. (independent and identically distributed) sample assumption[6, 8, 9]. To highlight this dependency, we review how the constraints in this work are derived based on the i.i.d. assumption.

First, we introduce the i.i.d. assumption for samples in the fixed dataset. An agent may revisit a state $s \in S$ multiple times and take an action $a \in A$. The i.i.d. assumption

means that, at any given time, the sampling of the next state s' is always based on the transition function $T(\cdot|s,a)$, and the samples are independent of each other. Formally, let $\tau_1, \tau_2, \ldots \tau_{N_D(s,a)}$ denote the $N_D(s,a)$ time steps where the agent takes action a in state s within the dataset D. The sequence of the $N_D(s,a)$ next states s',sampled after taking action a in state s, is denoted by $Y_{s,a} = \left(s'_{(\tau_1+1)}, s'_{(\tau_2+1)}, \ldots, s'_{\left(\tau_{N_D(s,a)}+1\right)}\right)$. Define $Y^i_{s,a}$ to represent the i-th element in the sequence $Y_{s,a}$, where $1 \le i \le N_D(s,a)$. If, for all $(s,a) \in S \times A$, each $Y^i_{s,a}$ in the sequence $Y_{s,a}$ is independently drawn from the transition function $T(\cdot|s,a)$, then dataset D is considered to satisfy the i.i.d. assumption.

Next, we discuss the relationship between the number of the state-action pairs $N_D(s,a)$ in D and the accuracy of the estimated $\widehat{M}$, which is crucial for deriving learning constraints. Intuitively, having more $N_D(s,a)$ samples help $\widehat{M}$ better approximate the true environment M^*. If $\widehat{M}$ is sufficiently accurate, the optimal policy learned from $\widehat{M}$ is likely to perform optimally in $\widehat{M}$, and thus outperform the behavior policy. The accuracy of the estimated MDP $\widehat{M}$ is typically measured by the difference between its transition function $\hat{T}$ and the true transition function T^*. A common metric is the L_1 distance, which sums the absolute differences in transition probabilities over all state-action pairs.

Lemma 1 (L_1 inequality) [23]. According to Hoeffding's inequality, if dataset D satisfy the i.i.d. assumption, then, for all $\epsilon > 0$,

$$\forall s, s' \in S, a \in A, \Pr\left(\left\|T^*(s'|s,a) - \hat{T}(s'|s,a)\right\|_1 \ge \epsilon\right) \le \left(2^{|S|} - 2\right) exp^{-\frac{1}{2}N_D(s,a)\epsilon^2} \quad (5)$$

Here, $\|\cdot\|_1$ denotes the L_1 distance, i.e., the sum of the absolute differences of the corresponding components of two vectors.

Finally, we show that the L_1 inequality plays a crucial role in deriving the learning constraint, which in turn highlights the necessity of the i.i.d. assumption. Let uncertainty set $\Xi_e^{\widehat{M}}$ is the set of MDPs with transition function $T(\cdot|s,a)$, such that L_1 distance between $T(\cdot|s,a)$ and $\hat{T}(\cdot|s,a)$ is smaller than $e(s,a)$ for every state-action pair, that is:

$$\Xi_e^{\widehat{M}} = \{M = (S, A, R, T, \gamma, s_{init}) | \forall (s,a) \in S \times A, s' \in S \; s.t. \; \left\|T(s'|s,a) - \hat{T}(s'|s,a)\right\|_1 \le e(s,a)\} \quad (6)$$

By setting $e(s,a) = \epsilon = \sqrt{\frac{2}{N(s,a)} \log\left(\frac{2|S||A|2^{|S|}}{\delta}\right)}$, and applying the L_1 inequality, we have $\Pr\left(M^* \notin \Xi_e^{\widehat{M}}\right) \le \delta$, which means that $\Xi_e^{\widehat{M}}$ includes the true environment M^* with a probability of $1 - \delta$ (the proof can be found in Proposition 9 of [6]). Therefore, if the optimal policy π_I, learned based on $\widehat{M}$, is optimal for all MDPs in $\Xi_e^{\widehat{M}}$, then π_I will outperform the baseline policy π_b with a probability of $1 - \delta$. Therefore, the SPI problem is formulated as follows:

$$\pi_I \in \arg\max_{\pi} \rho\left(\pi, \widehat{M}\right), \text{s.t.} \, \forall M \in \Xi_e^{\widehat{M}}, \rho(\pi, M) \ge \rho(\pi_b, M) - \zeta \quad (7)$$

According to Theorem 8 in [6], the solution can be derived by enforcing the learning constraint specified in Equation (3). Specifically, if the sampling of all state-action pairs in the dataset D is at least the threshold $N_\wedge$ described in equation (3), then the formula can be solved.

In summary, the constraints presented in this paper, as well as those proposed by other SPI methods, rely on the i.i.d. assumption of D to derive learning constraints, ensuring that the strategy trained under these constraints is likely to outperform the baseline strategy with an admissible performance loss.

2.3 RL based on ε-Bisimulation

ε-Bisimulation (also known as approximation stochastic bisimulation) is a state abstraction technique that maps the original states space S in an MDP into smaller abstract state $\bar{S}$ in an abstract MDP, which reduces the problem complexity while maintaining a bounded loss with respect to the original problem [11, 24]. Recently, driven by the rapid advancements in reinforcement learning, ε-bisimulation has also been incorporated into the RL paradigm to improve sample efficiency [12, 14, 25, 26]. In their setting, there exits an ε-bisimulation function $\phi: S \to \bar{S}$. And the agent acts in an MDP that returns states s, but instead of observing the true state s, the agent observes abstract states $\phi(s)$.

In this section, we introduce the notion of ε-bisimulation and discuss how to learn policies in RL based on ε-Bisimulation.

Definition 1 (ε-bisimulation function, ϕ) [11, 24]. Given two states $s_1, s_2 \in S$, if for any action $a \in A$, the difference in their transition probabilities to any abstract state $\bar{s}' \in \bar{S}$ is bounded by η, then s_1 and s_2 can be mapped into the same abstract state under the function ϕ, i.e.:

$$\phi(s_1) = \phi(s_2) \Longrightarrow \forall \bar{s}' \in \bar{S}, a \in A: |T(\bar{s}'|s_1, a) - T(\bar{s}'|s_2, a)| \leq \eta \tag{8}$$

Under the ε-bisimulation function $\phi: S \to \bar{S}$, the original dataset $D = \{(s_j, a_j, r_j, s_j')|j \in [1, N]\}$ collected from the underlying MDP can be transformed into an abstracted dataset $\mathcal{D} = \{(\bar{s}_j, a_j, r_j, \bar{s}_j')|j \in [1, N]\}$, where $\bar{s}_j = \phi(s_j)$ and $\bar{s}_j' = \phi(s_j, a_j, r_j, s_j')$. This dataset captures the transitions between abstract states, thereby enabling policy learning in the abstracted state space.

An abstract policy $\bar{\pi}: \bar{S} \to \Delta(A)$ can be optimized using batch RL algorithm in a model-based manner. Specifically, we consider the estimated abstract MDP $\hat{\bar{M}} = \left(\bar{S}, A, \hat{\bar{T}}, R, \bar{s}_0, \gamma\right)$, where $\hat{\bar{T}}$ denotes the transition dynamics over the abstract states, which can be derived from the abstracted dataset $\mathcal{D}$ as follows:

$$\forall \bar{s}, \bar{s}' \in S, a \in A, \hat{\bar{T}}(\bar{s}'|\bar{s}, a) = \frac{N_{\mathcal{D}}(\bar{s}, a, \bar{s}')}{N_{\mathcal{D}}(\bar{s}, a)} \tag{11}$$

Where $N_{\mathcal{D}}(\bar{s}, a)$ denotes the number of samples in $\mathcal{D}$ where action a is taken in $\bar{s}$, and $N_{\mathcal{D}}(\bar{s}, a, \bar{s}')$ denotes the number of samples transitioning to $\bar{s}'$. The value of abstract MDP, $\rho\left(\bar{\pi}, \hat{\bar{M}}\right)$, naturally mirrors that of the original MDP. The optimal abstract policy is defined as $\bar{\pi}^* \in \arg\max_{\pi \in \Pi} \rho\left(\bar{\pi}, \hat{\bar{M}}\right)$, and can be directly learned by vanilla batch RL algorithms[17, 18, 27].

While ε-bisimulation facilitates improved sample efficiency, it inevitably incurs a performance gap between the policy learned from the abstract dataset and that learned from the original dataset, thereby presenting challenges in ensuring the performance of the abstract policy.

3 SPI based on ε-Bisimulation

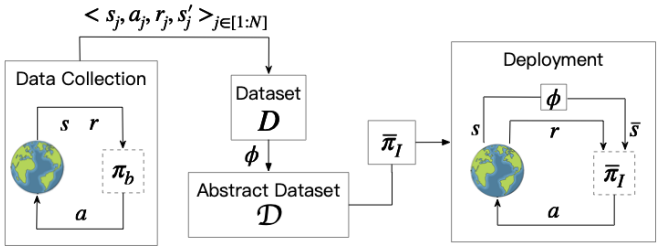

Fig.2. Illustration of the SPI with ε-bisimulation

To enhance the sample efficiency of Safe Policy Improvement (SPI), we study the SPI problem under ε-bisimulation. As illustrated in Fig. 2, the framework assumes an ε-bisimulation function $\phi: S \to \bar{S}$, which maps the original dataset $D = \{(s_j, a_j, r_j, s'_j)|j \in [1, N]\}$ into an abstract dataset $\mathcal{D} = \{(\bar{s}_j, a_j, r_j, \bar{s}'_j)|j \in [1, N]\}$. The objective is to design constraints that enable learning an abstract policy $\bar{\pi}_I$ from $\mathcal{D}$, such that it outperforms a baseline policy π_b in the true environment M^*. Formally, given δ and ζ, the goal is to learn $\bar{\pi}_I$ satisfying:

$$\rho(\bar{\pi}_I, M^*) \geq \rho(\pi_b, M^*) - \zeta$$

Where $\rho(\cdot, M^*)$ denotes the expected return (also called performance) in M^*.

3.1 Theoretical Challenges

The key challenge lies in designing appropriate constraints for learning the abstract policy. However, existing SPI methods cannot directly provide such constraints, as they rely on the i.i.d. assumption, which does not hold for the abstract dataset $\mathcal{D}$. To illustrate why this assumption may be violated, we present a simple example shown in Fig. 4.

As shown in Fig. 4, the environment is a simple MDP, where the circles denote the state set $S = \{s_0, s_1, s_2, s_3\}$, and the agent can take the same action A={a} in all states (for simplicity, this unique action is omitted in the figure）. The arrows with associated probabilities indicate state transitions by taking the unique action. The colored circles represent the abstract state set $\bar{S} = \{\bar{s}_0, \bar{s}_1, \bar{s}_2\}$. According to the ε-bisimulation function ϕ, states s_0 and $s_2$are clustered into the same abstract state $\bar{s}_0$.

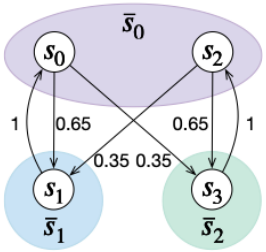

Fig.3. A Simple MDP [12]

Example 3.1. Take $\bar{s}_0$ in Fig. 3 as an example. The i.i.d. assumption means that whenever the agent visits $\bar{s}_0$ and takes the same action, the next abstract state ($\bar{s}_1$ or $\bar{s}_2$) is drawn from the same distribution, and each sample is independent.

First, we show that transitions from $\bar{s}_0$ under the same action do not follow the same distribution. Let $\phi^{-1}(\bar{s}_0) = \{s_0, s_2\}$ denotes the set of all states that are mapped to $\bar{s}_0$ by ϕ. In different time steps, the agent may be in either s_0 or s_2, both corresponding to the same abstract state $\bar{s}_0$. However, after taking the same action, their transition probabilities to other abstract states are different: for s_0, the transition distribution is $\{\bar{s}_1: 0.65, \bar{s}_2: 0.35\}$, while for s_2, it becomes $\{\bar{s}_1: 0.4, \bar{s}_2: 0.6\}$. Clearly, these distributions are not identical, violating the i.i.d. assumption.

Next, we show a counterexample to illustrate that the transitions originating from $\bar{s}_0$ are not independent. Define $Y^0_{\bar{s}_0} = \bar{s}_1$ and $Y^1_{\bar{s}_0} = \bar{s}_1$ as two transitions collected from $\bar{s}_0$ by taking same action, both of which lead to the abstract state $\bar{s}_1$. If these two transitions are independent, their joint probability should satisfy the following condition:

$$\Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1, Y^1_{\bar{s}_0} = \bar{s}_1\right) = \Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1\right) \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1\right) \tag{12}$$

Example 3.2. Suppose that when the agent visits $\bar{s}_0$ for the first time, the underlying environment state is s_0. The probability of transitioning to $\bar{s}_1$ is then $\Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1\right) = \Pr(\bar{s}_1|s_0) = 0.5$. For the second visit to $\bar{s}_0$, we do not fix the underlying state. Instead, we consider all possible states that may lead to this visit.

$$\begin{aligned}
&\Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1\right) = \sum_{\bar{s} \in \bar{S}} \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1 \middle| Y^0_{\bar{s}_0} = \bar{s}\right) \Pr\left(Y^0_{\bar{s}_0} = \bar{s}\right) \\
&= \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1 \middle| Y^0_{\bar{s}_0} = \bar{s}_0\right) \Pr\left(Y^0_{\bar{s}_0} = \bar{s}_0\right) + \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1 \middle| Y^0_{\bar{s}_0} = \bar{s}_1\right) \Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1\right) \\
&+ \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1 \middle| Y^0_{\bar{s}_0} = \bar{s}_2\right) \Pr\left(Y^0_{\bar{s}_0} = \bar{s}_2\right) \\
&= \Pr\left(Y^1_{\bar{s}_0} = s_1 \middle| Y^0_{\bar{s}_0} = s_1\right) \Pr\left(Y^0_{\bar{s}_0} = s_1\right) + \Pr\left(Y^1_{\bar{s}_0} = s_1 \middle| Y^0_{\bar{s}_0} = s_3\right) \Pr\left(Y^0_{\bar{s}_0} = s_3\right) \\
&= 0.65 \times 0.65 + 0.35 \times 0.35 \\
&= 0.545
\end{aligned}$$

Additionally，the joint probability is given by:

$$\begin{aligned}
&\Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1, Y^1_{\bar{s}_0} = \bar{s}_1\right) = \Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1\right) \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1 | Y^0_{\bar{s}_0} = \bar{s}_1\right) \\
&= \Pr(\bar{s}_1|s_0)\left(\Pr(\bar{s}_1|s_0) \Pr(s_0|\bar{s}_1)\right) \\
&= 0.65 \times (0.65 \times 1) \\
&= 0.4225
\end{aligned}$$

Here, $\Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1 | Y^0_{\bar{s}_0} = \bar{s}_1\right) = \Pr(\bar{s}_1|s_0) \Pr(s_0|\bar{s}_1)$, since the abstract state $\bar{s}_1$ transitions to s_0 with probability 1. Clearly, the samples in this example do not satisfy the independence condition (Equation 13), as $\Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1\right) \Pr\left(Y^1_{\bar{s}_0} = \bar{s}_1\right) = 0.52 \times 0.6 \neq \Pr\left(Y^0_{\bar{s}_0} = \bar{s}_1, Y^1_{\bar{s}_0} = \bar{s}_1\right)$, meaning the sampling of $Y^0_{\bar{s}_0}$ and $Y^1_{\bar{s}_0}$ are not independent.

In summary, in the SPI problem based on ε-Bisimulation, the abstract dataset $\mathcal{D}$ does not satisfy the i.i.d. assumption. Existing methods, relying on this assumption to design policy constraints, cannot be directly applied to ensure the performance of the learned policy from $\mathcal{D}$.

3.2 Constraints for Non-i.i.d. Datasets

For the SPI problem based on ε-bisimulation, where the abstract datasets $\mathcal{D}$ does not satisfy the i.i.d. assumption, we propose constraints to ensure the generated policy meets the problem's performance requirements.

We investigate how to design learning constraints for $\mathcal{D}$ to ensure that the generated abstract policy $\bar{\pi}$ outperforms the baseline π_b in the true environment with high probability. This problem be converted into ensuring that $\bar{\pi}$ outperforms an abstract baseline

$\bar{\pi}_b: \bar{S} \to A$, which can be compute by $\phi: \bar{S} \to S$ and $\pi_b: S \to A$ (See Appendix A for detailed definition.). This is because, given a true MDP $M^* = (S, A, T^*, R, s_0, \gamma)$ and the corresponding MDP $\bar{M}_\omega^* = (\bar{S}, A, \bar{T}_\omega^*, R, \bar{s}_0, \gamma)$, the performance difference between the policy in the abstract model and its performance in the true model is bounded. In the case of strict abstraction, the two performances are nearly identical（Detailed proof can be found in Lemma 6 of [12]）.

RL with a model-based manner first constructs an estimated abstract MDP $\hat{\bar{M}} = \left(\bar{S}, A, \hat{\bar{T}}, R, \bar{s}_0, \gamma\right)$ and then learns the optimal policy in $\hat{\bar{M}}$. The error between $\hat{\bar{M}}$ and the true abstract MDP $\bar{M}_\omega^*$ cannot be measured using the L_1 inequality, as it depends on the i.i.d. assumption. Fortunately, for a dataset $\mathcal{D}$ that does not follow the i.i.d. assumption, the difference between $\hat{\bar{M}}$ and $\bar{M}_\omega^*$ can be quantified using Azuma-Hoeffding Inequality.

Lemma 2 (Abstract L_1 inequality)[12]. Given an abstract dataset $\mathcal{D} = \left\{\left(\bar{s}_j, a_j, r_j, \bar{s}_j'\right) | j \in [1, N]\right\}$, the deviation between the transition function $\hat{\bar{T}}$ in $\hat{\bar{M}}$ the true transition function $\bar{T}_\omega^*$ in $\bar{M}_\omega^*$is bounded by the following inequality:

$$\forall \bar{s}, a \in \bar{S} \times A, \bar{s}' \in \bar{S}, \Pr\left(\left\|\hat{\bar{T}}(\bar{s}'|\bar{s}, a) - \bar{T}_\omega(\bar{s}'|\bar{s}, a)\right\|_1 \geq \epsilon\right) \leq 2^{|\bar{S}|} e^{-\frac{1}{8} N_{\mathcal{D}}(\bar{s}, a)\epsilon^2} \quad (13)$$

This inequality characterizes how the number of samples $N_{\mathcal{D}}(\bar{s}, a)$ collected for each abstract state-action pair in $\mathcal{D}$ affects the accuracy of the estimated transition function $\hat{\bar{T}}$ in terms of its L_1 distance to the true transition $\bar{T}_\omega^*$.

Constraints for Training with the Abstract Dataset $\mathcal{D}$. Given parameters δ and ζ', the abstract dataset $\mathcal{D} = \left\{\left(\bar{s}_j, a_j, r_j, \bar{s}_j'\right) | j \in [1, N]\right\}$ is required to contain sufficient samples for every abstract state-action pair $(\bar{s}, a)$.

Lemma 3(Constrains in $\mathcal{D}$). The number of samples $N_{\mathcal{D}}(\bar{s}, a)$ must satisfy:

$$N_{\mathcal{D}}(\bar{s}, a) \geq N_\wedge = \frac{32\gamma^2 Rmax^2}{\zeta'^2 (1-\gamma)^4} \ln \frac{|\bar{S}||A|2^{|\bar{S}|}}{\delta} \quad (14)$$

Under this condition, with probability at least $1 - \delta$, the optimal policy $\bar{\pi}_I$ learned from the estimated abstract MDP $\hat{\bar{M}}$ constructed from $\mathcal{D}$ is an approximate improvement over the baseline policy $\bar{\pi}_b$, i.e., $\rho(\bar{\pi}_I, \bar{M}_\omega^*) \geq \rho(\bar{\pi}_b, \bar{M}_\omega^*) - \zeta'$.

The proof is provided in Appendix B.

Then, we derive the performance guarantee in the true environment M^*.

Theorem 1. Based on Lemma 3 and the performance gap of the abstract policy between the abstract MDP and the true MDP M^*, we can further derive that:

$$\rho(\bar{\pi}_I, M^*) \geq \rho(\pi_b, M^*) - \zeta \quad (15)$$

Here, $\zeta = \zeta' + \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$.

The proof is provided in Appendix C.

Based on Theorem 1, the user can set an acceptable performance error ζ and derive the sampling threshold $N_\wedge$ for every abstract state-action pair. Specifically, ζ' is calculated as $\zeta' = \zeta - \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$, and $N_\wedge$ is derived from equation (14). When the sample size

for every abstract state-action pair $N_{\mathcal{D}}(\bar{s}, a) \geq N_{\wedge}$ in the abstract dataset $\mathcal{D}$, the policy learned from $\mathcal{D}$ satisfies the SPI problem requirements in the true environment.

4 Related Works

Batch reinforcement learning (Batch RL)[4], also known as offline reinforcement learning [5], focuses on how to learn a policy from pre-collected fixed dataset when the agent cannot directly interact with the environment. This paper addresses the safety policy improvement (SPI) problem in batch RL [6-9], which involves learning a policy from a fixed dataset and guaranteeing that the learned policy outperforms the baseline policy used to generate those samples. In reinforcement learning, the concept of "safety" can have multiple meanings [28], including parameter uncertainty [29], model uncertainty [30], external interruptibility [31, 32], and safety concerns in exploration in risky environments [33, 34]. The SPI problem primarily concerns safety related to parameter uncertainty.

Early approaches to the SPI problem mainly used the model-free RL paradigm, where policies are learned from a dataset without constructing an environment model [19, 20]. These methods work well only when nondeterministic parameters, like transition probabilities, follow a uniform distribution. Otherwise, they return the baseline policy and cannot generate the target policy. The paper [6] adopts a model-based approach, minimizing robust baseline regret, which transforms the SPI problem into a state-action pairs version, allowing it to handle nondeterministic parameters that don't follow a uniform distribution. It proves that the SPI problem is NP-hard and introduces constraints to approximate target policy learning, though scalability remains limited. Paper [7] proposes the baseline-guided SPI method (SPIBB), which builds on [6] by adding constraints for different state-action pairs, ensuring performance even when some pairs don't satisfy the constraints. Other works either improve effectiveness [21, 22] or extend SPI to more complex settings, such as partially observable MDPs [9]. A common feature of these works is their reliance on the i.i.d. assumption of fixed dataset.

5 Conclusion and Future

We propose a constraint for SPI based on ε-bisimulation. Given a fixed dataset, a safety requirement parameter, and a ε-bisimulation function, the proposed constraint restricts policy trains only when the specified constraint is satisfied. This approach facilitates adopt ε-bisimulation and provides with high probability guarantees that the learned policy outperforms the baseline policy that generated the dataset.

There are two promising directions for future research. The first is to adaptively learn the ε-bisimulation function from different datasets, and further derive corresponding training constraints under such learned ε-bisimulation function. The second is to extend our framework to more complex models, such as partially observable Markov decision processes (POMDPs), to improve sample efficiency.

References

1. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G.: Human-level control through deep reinforcement learning. nature 518, 529-533 (2015)
3. Retzlaff, C.O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M.E., Holzinger, A.: Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. Journal of Artificial Intelligence Research 79, 359-415 (2024)
4. Lange, S., Gabel, T., Riedmiller, M.: Batch reinforcement learning. Reinforcement learning: State-of-the-art, pp. 45-73. Springer (2012)
5. Jia, Z., Rakhlin, A., Sekhari, A., Wei, C.-Y.: Offline Reinforcement Learning: Role of State Aggregation and Trajectory Data. arXiv preprint arXiv:2403.17091 (2024)
6. Ghavamzadeh, M., Petrik, M., Chow, Y.: Safe policy improvement by minimizing robust baseline regret. Advances in Neural Information Processing Systems 29, (2016)
7. Laroche, R., Trichelair, P., Des Combes, R.T.: Safe policy improvement with baseline bootstrapping. In: International conference on machine learning, pp. 3652-3661. PMLR, (Year)
8. Scholl, P., Dietrich, F., Otte, C., Udluft, S.: Safe policy improvement approaches and their limitations. In: International Conference on Agents and Artificial Intelligence, pp. 74-98. Springer, (Year)
9. Simão, T.D., Suilen, M., Jansen, N.: Safe Policy Improvement for POMDPs via Finite-State Controllers. arXiv preprint arXiv:2301.04939 (2023)
10. Abel, D., Hershkowitz, D., Littman, M.: Near optimal behavior via approximate state abstraction. In: International Conference on Machine Learning, pp. 2915-2923. PMLR, (Year)
11. Li, L., Walsh, T.J., Littman, M.L.: Towards a unified theory of state abstraction for MDPs. In: AI&M. (Year)
12. Starre, R.A., Loog, M., Congeduti, E., Oliehoek, F.A.: An Analysis of Model-Based Reinforcement Learning From Abstracted Observations. Transactions on Machine Learning Research (2023)
13. Paduraru, C., Kaplow, R., Precup, D., Pineau, J.: Model-based reinforcement learning with state aggregation. In: 8th European Workshop on Reinforcement Learning. (Year)
14. Starre, R.A., Loog, M., Oliehoek, F.A.: Model-Based Reinforcement Learning with State Abstraction: A Survey. In: BNAIC/BeNeLearn 2022. (Year)
15. Bellman, R.: A Markovian decision process. Journal of mathematics and mechanics 679-684 (1957)
16. Moerland, T.M., Broekens, J., Plaat, A., Jonker, C.M.: Model-based reinforcement learning: A survey. Foundations and Trends® in Machine Learning 16, 1-118 (2023)
17. Chatterjee, K., Henzinger, T.A.: Value iteration. 25 Years of Model Checking: History, Achievements, Perspectives, pp. 107-138. Springer (2008)
18. Watkins, C.J., Dayan, P.: Q-learning. Machine learning 8, 279-292 (1992)
19. Thomas, P., Theocharous, G., Ghavamzadeh, M.: High-confidence off-policy evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence. (Year)

20. Kakade, S., Langford, J.: Approximately optimal approximate reinforcement learning. In: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 267-274. (Year)
21. Simão, T.D., Spaan, M.T.: Safe policy improvement with baseline bootstrapping in factored environments. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4967-4974. (Year)
22. Nadjahi, K., Laroche, R., Tachet des Combes, R.: Safe policy improvement with soft baseline bootstrapping. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III, pp. 53-68. Springer, (Year)
23. Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., Weinberger, M.J.: Inequalities for the L1 deviation of the empirical distribution. Hewlett-Packard Labs, Tech. Rep 125 (2003)
24. Auer, P., Jaksch, T., Ortner, R.: Near-optimal regret bounds for reinforcement learning. Advances in neural information processing systems 21, (2008)
25. Ortner, R., Maillard, O.-A., Ryabko, D.: Selecting near-optimal approximate state representations in reinforcement learning. In: International Conference on Algorithmic Learning Theory, pp. 140-154. Springer, (Year)
26. Abel, D., Arumugam, D., Lehnert, L., Littman, M.: State abstractions for lifelong reinforcement learning. In: International Conference on Machine Learning, pp. 10-19. PMLR, (Year)
27. Ernst, D., Geurts, P., Wehenkel, L.: Tree-based batch mode reinforcement learning. Journal of Machine Learning Research 6, (2005)
28. Garcıa, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research 16, 1437-1480 (2015)
29. Thomas, P., Theocharous, G., Ghavamzadeh, M.: High confidence policy improvement. In: International Conference on Machine Learning, pp. 2380-2388. PMLR, (Year)
30. Altman, E.: Constrained Markov decision processes. Routledge (2021)
31. El Mhamdi, E.M., Guerraoui, R., Hendrikx, H., Maurer, A.: Dynamic safe interruptibility for decentralized multi-agent reinforcement learning. Advances in Neural Information Processing Systems 30, (2017)
32. Orseau, L., Armstrong, M.: Safely interruptible agents. In: Conference on Uncertainty in Artificial Intelligence. Association for Uncertainty in Artificial Intelligence, (Year)
33. Schulman, J.: Trust Region Policy Optimization. arXiv preprint arXiv:1502.05477 (2015)
34. Fatemi, M., Sharma, S., Van Seijen, H., Kahou, S.E.: Dead-ends and secure exploration in reinforcement learning. In: International Conference on Machine Learning, pp. 1873-1881. PMLR, (Year)

Appendix

A. Definition of the Abstract Baseline Policy

We define an abstract baseline policy $\bar{\pi}_b$ that preserves the action choices of the original baseline policy π_b under the ε-bisimulation function $\phi: S \to \bar{S}$. Specifically, for any abstract state $\bar{s} \in \bar{S}$, $\bar{\pi}_b(\bar{s})$ selects an action consistent with π_b over the set of original states mapped to $\bar{s}$, i.e.,

$$\forall \bar{s} \in \bar{S}, \bar{\pi}_b(\bar{s}) = \begin{cases} a_i, if\ s_i \in \phi^{-1}(\bar{s})\ and\ \pi_b(s_i) = a_i \\ a_j, if\ s_j \in \phi^{-1}(\bar{s})\ and\ \pi_b(s_j) = a_j \\ \cdots \end{cases}$$

By construction, $\bar{\pi}_b$ can be viewed as the abstraction of π_b in the abstract MDP. Consequently, $\bar{\pi}_b$ and π_b are performance-equivalent in their corresponding models.

B. Proof of Lemma 3.

In this section, we prove Lemma 3. First, we start with uncertainty set $\Xi_e^{\hat{\bar{M}}}$:

$$\Xi_e^{\hat{\bar{M}}} = \left\{ \bar{M} = (\bar{S}, A, R, \bar{T}, \gamma, \bar{s}_0) \middle| \forall (\bar{s}, a) \in \bar{S} \times A, \bar{s}' \in \bar{S}\ s.t. \left\| \bar{T}(\bar{s}'|\bar{s}, a) - \hat{\bar{T}}(\bar{s}'|\bar{s}, a) \right\|_1 \le e(\bar{s}, a) \right\} \quad (20)$$

which is the set of MDPs with transition function $T(\cdot|s,a)$, such that L_1 distance between $T(\cdot|s,a)$ and $\hat{T}(\cdot|s,a)$ is smaller than $e(s,a)$ for every state-action pair.

Let $e(\bar{s}, a) = \sqrt{\frac{8}{N_{\mathcal{D}}(\bar{s},a)} \ln \frac{|\bar{S}||A|2^{|\bar{S}|}}{\delta}}$. By the abstract L_1 inequality (Lemma 2), we have:

$$\Pr\left(\bar{M}_\omega^* \notin \Xi_e^{\hat{\bar{M}}} \right) \le \delta \quad (21)$$

This means that with at least $1-\delta$ probability, $\Xi_e^{\hat{\bar{M}}}$ contains the true abstract MDP $\bar{M}_\omega^*$. The SPI problem can be reformulated as:

$$\bar{\pi}_I \in \arg\max_{\bar{\pi}} \rho\left(\bar{\pi}, \hat{\bar{M}}\right), \text{s.t.} \forall \bar{M} \in \Xi_e^{\bar{M}_Y}, \rho(\bar{\pi}, \bar{M}) \ge \rho(\bar{\pi}_b, \bar{M}) - \zeta' \quad (22)$$

This seeks the approximate optimal policy $\bar{\pi}_I$ based on $\hat{\bar{M}}$, ensuring that for any MDP $\bar{M}$ in $\Xi_e^{\bar{M}_Y}$, it is a approximate improvement over $\bar{\pi}_b$. Theorem 8 in [6] can be used to solve this, yielding:

$$\rho(\bar{\pi}_I, \bar{M}_\omega^*) \ge \rho(\bar{\pi}^*, \bar{M}_\omega^*) - \frac{2\gamma Rmax}{(1-\gamma)^2}\|e\|_\infty \ge \rho(\bar{\pi}_b, \bar{M}_\omega^*) - \frac{2\gamma Rmax}{(1-\gamma)^2}\|e\|_\infty \quad (23)$$

Where $\|\cdot\|_\infty$ is the infinity norm, representing the maximum absolute value in the vector. Let $N_\wedge$ denote the minimum sample size for the abstract state-action pair, then $\|e\|_\infty = \sqrt{\frac{8}{N_\wedge} \ln \frac{|\bar{S}||A|2^{|\bar{S}|}}{\delta}}$. Let $\zeta' = \frac{2\gamma Rmax}{(1-\gamma)^2}\|e\|_\infty$, then:

$$N_\wedge = \frac{32\gamma^2 Rmax^2}{\zeta'^2(1-\gamma)^4} \ln \frac{|\bar{S}||A|2^{|\bar{S}|}}{\delta} \quad (24)$$

Thus, when the sample size for all abstract state-action pairs $N_{\mathcal{D}}(\bar{s}, a)$ satisfies:

$$N(\bar{s}, a) \ge N_\wedge = \frac{32\gamma^2 Rmax^2}{\zeta'^2(1-\gamma)^4} \ln \frac{|\bar{S}||A|2^{|\bar{S}|}}{\delta} \quad (25)$$

the approximate optimal abstract policy $\bar{\pi}_I$ is an approximate improvement over $\bar{M}_\omega^*$,i.e.,

$$\rho(\bar{\pi}_I, \bar{M}_\omega^*) \geq \rho(\bar{\pi}_b, \bar{M}_\omega^*) - \zeta' \tag{26}$$

This concludes the proof.

C. Proof of Theorem 1.

In this section, we prove Theorem 1. We begin with the performance gap between an abstract policy $\bar{\pi}$ in the true abstract MDP $\bar{M}_\omega$ and its corresponding true MDP M. According to Lemma 6 in [18], the performance gap satisfies the following inequality constraint:

$$|\rho(\bar{\pi}, \bar{M}_\omega) - \rho(\bar{\pi}, M)| \leq \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2} \tag{27}$$

Let π in equation (27) be $\bar{\pi}_I$, then equation (27) can be rewritten as:

$$|\rho(\bar{\pi}_I, M) - \rho(\bar{\pi}_I, \bar{M}_\omega)| \leq \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2} \tag{28}$$

Moreover, $\bar{\pi}_b$, derived from the baseline policy π_b, has the same performance in both the abstract and actual MDPs as π_b. Therefore,

$$|\rho(\bar{\pi}_b, M) - \rho(\bar{\pi}_b, \bar{M}_\omega)| = 0 \tag{29}$$

Then, by adding both sides of equation (28) and equation (29), we obtain:

$$|\rho(\bar{\pi}_I, M) - \rho(\bar{\pi}_I, \bar{M}_\omega)| + |\rho(\bar{\pi}_b, M) - \rho(\bar{\pi}_b, \bar{M}_\omega)| \leq \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$$

$$\left|\rho(\bar{\pi}_I, M) - \rho(\bar{\pi}_b, M) - \left(\rho(\bar{\pi}_I, \bar{M}_\omega) - \rho(\bar{\pi}_b, \bar{M}_\omega)\right)\right| \leq \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$$

$$\rho(\bar{\pi}_I, M) - \rho(\bar{\pi}_b, M) - \left(\rho(\bar{\pi}_I, \bar{M}_\omega) - \rho(\bar{\pi}_b, \bar{M}_\omega)\right) \geq -\frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$$

$$\rho(\bar{\pi}_I, M) - \rho(\bar{\pi}_b, M) \geq \rho(\bar{\pi}_I, \bar{M}_\omega) - \rho(\bar{\pi}_b, \bar{M}_\omega) - \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$$

We have $\rho(\bar{\pi}_I, \bar{M}_\omega) - \rho(\bar{\pi}_b, \bar{M}_\omega) \geq -\zeta'$(Lemma 3), so

$$\rho(\bar{\pi}_I, M) \geq \rho(\bar{\pi}_b, M) - \zeta' - \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2} \tag{30}$$

Let $\zeta = \zeta' + \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$, the above expression can be written as

$$\rho(\bar{\pi}_I, M) \geq \rho(\bar{\pi}_b, M) - \zeta \tag{30}$$

Where $\zeta = \zeta' + \frac{\gamma\eta|\bar{S}|R_{max}}{(1-\gamma)^2}$.

This concludes the proof.