Journal of Scientific Computing (2025) 103:51
https://doi.org/10.1007/s10915-025-02873-6

An Analysis of Offline Model-Based Learning with Action Noise

Haoya Li[1] · **Tanmay Gangwani**[2] · **Lexing Ying**[1]

Received: 26 May 2023 / Revised: 13 January 2025 / Accepted: 10 March 2025

Abstract
This paper studies the problem of action noise in model-based offline reinforcement learning, i.e., the actions recorded in the transition trajectories are polluted with noise. Though this is a relatively new problem in the literature, it has become a relevant issue since offline reinforcement learning has become more and more widely used. This is particularly important in applications where an offline dataset collected without the intention to improve the policy is repurposed for reinforcement learning. This paper presents an error analysis for the value function due to the action noise and provides numerical studies. This work also prepares for further developments of novel algorithms for addressing action noise.

Keywords Reinforcement learning · Offline reinforcement learning · Action noise

1 Introduction

Markov decision process (MDP) is the central mathematical model for reinforcement learning (RL). In recent years, offline MDP learning algorithms [19, 27], for which the data are collected by a behavior policy prior to the training stage, have received significant attention. Many model-based [4, 6, 8, 14, 23, 29, 32, 35, 40, 41] and model-free algorithms [2, 3, 5, 9–11, 16–18, 24, 26, 30, 43] have been proposed. Several convergence results [7, 13, 34, 37] have also been developed. Among them, the model-based offline MDP methods benefit from the conceptual simplicity since learning the MDP model can be treated as a supervised learning problem. The model extracted from the offline dataset enjoys generalization ability if the behavior policy is sufficiently exploratory. In this paper, we consider this model-based setting.

✉ Haoya Li
lihaoya@stanford.edu

Tanmay Gangwani
gangwt@amazon.com

Lexing Ying
lexing@stanford.edu

[1] Stanford University, Stanford, USA

[2] Amazon Search, Seattle, USA

Published online: 01 April 2025

For offline model-based algorithms, the statistical errors of the estimated rewards and transitions due to finite sampling have been well-studied in the literature. In the situation of partially observed states, the theory of partially observed MDP has also been extensively investigated. In this paper, we are interested in the relatively unexplored issue of *action noise*, i.e., the recorded actions in the offline dataset are contaminated by noises. There are two common scenarios where this issue can occur. In the first scenario, the actions are not recorded carefully when the offline trajectories are stored. In the second scenario, the dataset is originally recorded without the intention of optimizing the policy, and hence, the action information is simply ignored. For example, the initial goal can be the prediction of the next state or its distribution while the policy remains fixed as the behavior policy or to classify the trajectories of states to facilitate other downstream tasks. It is only later that one switches the goal to the optimization of the policy and uses the dataset for reinforcement learning. In this scenario, the actions must be inferred from other information, and such an inference procedure often results in estimation errors. Motivated by these scenarios, it is therefore highly relevant to study the impact of action noise on offline MDP algorithms.

1.1 Related Works

Offline RL has grown into a large topic. Unlike online or off-policy RL, no further interaction with the true environment is allowed in the offline RL setting, and knowledge can only be extracted from a fixed dataset. In this regard, offline RL is similar to supervised learning. However, there is an essential difference between them. While it is usually reasonable to assume that the training data and the test data follow the same or a similar distribution for supervised learning tasks, this is not the case for offline RL. When performing reinforcement learning tasks from offline datasets, the learned policy must be different from the behavior policy if one wants to make a nontrivial improvement. Thus, the distribution of the agent's trajectories is different from that in the offline dataset when the learned policy is applied. This phenomenon is usually referred to as distributional shift or covariate shift, which is a major challenge in the field of offline RL. A variety of algorithms have been proposed to address this challenge using either a model-based or model-free strategy. A common ingredient in these algorithms is learning the behavior policy, which is also not accessible. Modern machine learning models such as neural networks and variational auto-encoders (VAE) [15] have been used to this end. This model should not be confused with the model of transition and reward used in model-based methods.

An early example of a model-free method is [10], where the algorithm is built upon Q-learning and the learned policy is forced to be a neural network perturbation of the behavior policy. Several subsequent methods adopt the idea of policy constraint, which means that the learned policy is restricted in a neighborhood of the behavior policy under certain metrics. For example, the advantage-weighted regression (AWR) method [26] implicitly adopts the KL-divergence constraint and a behavior cloning metric is used in [9]. Other metrics, such as the maximum mean discrepancy (MMD) used in [17], essentially require that the learned policy and the behavior policy must have equal support, which is shown to allow more flexibility. Policy constraints can also be enforced through learning the representation of actions. For example, the authors of [43] utilize the latent action of the CVAE trained to mimic the behavior policy, and the authors of [5] further improve this idea by updating the latent representation with the advantage function as weights.

In general, the idea of pessimism or conservatism is widely used to avoid being over-optimistic or exploiting the error caused by distributional shift, which often appears as a

conservative adjustment to the model, value function, or Q function according to the uncertainty in the information of the offline dataset. For example, a random convex combination of Q functions, instead of a single Q function, is used to perform policy updates in [2]. In [11], the standard deviation of an ensemble of independently trained Q functions is subtracted from the Q function. In [18, 30], a mini-max formulation is used for the Q function to penalize large Q values without sufficient data coverage. In [24], a conservative Bellman operator is used where out-of-distribution (OOD) actions are treated differently. Another way to follow the conservatism principle is to update the policy only once, which was proposed by Brandfonbrener et al. in [3] and dubbed the one-step method. No off-policy evaluation (OPE) is needed for one-step methods, which alleviate the iterative exploitation of errors. In light of this, Kostrikov et al. proposed implicit Q-learning (IQL) [16] where the expectile regression is used to modify the Bellman maximal operator, and an approximated optimal Q function is learned. The policy is then extracted from the approximate optimal Q function with an AWR routine.

In spite of the huge success brought by model-free offline RL algorithms, one often needs to tune the hyperparameters carefully to achieve ideal performance. The model-based methods, on the other hand, take advantage of the more established and reliable supervised learning methods to extract a model from the offline data and then obtain an improved policy by planning with the learned model. Commonly used models include the discrete model, kernel nonlinear regulator (KNR), Gaussian process (GP), and neural network (NN).

Early examples of model-based RL studies include [14, 41]. In [41], the reward is penalized by an estimation of uncertainty, quantified by the Frobenius norm of the covariance matrix of the Gaussian transition model. In [14], the transition dynamic is halted, and the reward is set as a negative constant if the uncertainty is larger than a threshold, where the uncertainty is measured by the maximal distance between independently trained models. Eysenbach et al. [8] augment the reward with a log-likelihood ratio term computed from a classifier that distinguishes the real transitions from fake ones. Besides modifying the reward, conservatism can also be integrated into model-based methods through the transition model. One common strategy is to learn a policy that optimizes the worst case for an ensemble of transition models, which is adopted by a line of work including [6, 29, 34]. Another way to use the model is to generate imaginary trajectories and feed them to a model-free method. For example, the COMBO method in [40] sends the imaginary trajectories generated from a learned transition model to the CQL method [18]. An additional time-reversed transition model is learned in [35] to generate reverse imaginations, and a double-check procedure of the imagination is added in [23] to further enhance conservatism.

Another line of work addressing uncertainty in MDPs is the field of robust MDPs, which investigates the complexity and tractability of MDP problems with adversarially chosen dynamics. An earlier example of robust MDP research is [36], which explores various assumptions about the structure of the uncertainty set, such as the (s, a)-rectangular assumption and the s-rectangular assumption. It was also shown that, without specific structural constraints on the uncertainty set, the corresponding problem becomes NP-hard. More recently, [38] provided non-asymptotic performance analysis under the (s, a)-rectangular and s-rectangular assumptions with different f-divergence type uncertainty sets. In the RL context, this field is often referred to as robust RL. For instance, [33] combines robust MDP concepts with deep reinforcement learning and conducts experiments on MuJoCo problems. The authors of [33] also consider a setting where actions are subject to adversarial perturbations, which bears some similarity to the setting studied in our paper. However, the algorithm proposed in [33] relies on a known model or an online setting. This is fundamentally different from

the problem addressed in our work, where perturbations (not necessarily adversarial) are introduced in the offline dataset. One can also refer to [25] for a more detailed review.

1.2 Comparison and Contributions

At a high level, the suboptimality of an RL policy learned from an offline dataset can be owed to two main sources, the first being the distributional shift and the second being the finite sample size (or statistical error). The statistical error should not be confused with the distributional shift since it exists even with a generative model, i.e., even when new data points can be collected from interacting directly with the environment, though the idea of conservatism can also be applied to the setting with a generative model. For example, [21] proposes to perturb the empirical reward by a uniform noise or intentionally select suboptimal actions with a random suboptimality in order to avoid the excessive exploitation of the empirical model.

In this paper, we propose to analyze yet another source of error that is ubiquitous in real-life offline RL applications, namely the misclassification of the action (or the action noise). Unlike distributional shift and statistical error, the pollution in the action information has been severely underexplored. We summarize our contributions below.

- We propose a methodology for studying a ubiquitous problem in offline RL where the action information in the offline dataset is polluted. This has not been fully investigated by the RL community.
- We provide error analyses of both the infinite-horizon and finite-horizon MDP models for both the infinite-sample limit and the finite-sample cases.
- We include a verification of the theoretical result with numerical examples.

1.3 Contents

The paper is organized as follows. Section 2 covers the basic material of MDP in both discounted infinite-horizon settings and undiscounted finite-horizon settings. Section 3 derives the error estimates solely due to action noise, assuming that the rewards and transition probabilities can be accessed without error. The work in Sect. 4 considers a more realistic scenario where the rewards and transition probabilities need to be estimated from a finite number of samples. Here, an additional statistical error is introduced due to the estimation. Numerical results are provided in both sections to verify the error bounds. Finally, Sect. 5 discusses some future directions, including the extension to continuous actions and the derivation of novel pessimism algorithms based on action noise.

2 Preliminaries

We consider both the discounted infinite-horizon MDP and the undiscounted finite-horizon MDP. This section covers some preliminaries of MDPs for completeness. We refer the readers to the existing literature [1, 28, 31, 39] for more details.

2.1 Infinite-Horizon

A discounted infinite-horizon MDP model is a tuple (S, A, P, r, γ), where S is the state space, A is the action space, P is a transition probability tensor such that the entry $P_{sas'}$ is the probability of transitioning from the state s to the state s' when action a is taken, and r is a reward matrix where r_{sa} is the reward obtained when action a is taken at state s. We assume that the reward is bounded and without loss of generality $r_{sa} \in [0, 1]$. Finally, $0 < \gamma < 1$ is a discount factor.

A policy π is a collection of probability distributions over A indexed by the states in S, where π_{sa} is the probability of choosing action a at state s with $\sum_{a \in A} \pi_{sa} = 1$. For a given policy π, let $(s_0, a_0, s_1, a_1, \ldots)$ be a stochastic trajectory generated by following π. The value function v^π associated with π is the expected discounted accumulative reward

$$v_s^\pi := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{s_k a_k} \mid s_0 = s \right].$$

Associated with π, we define the transition matrix P^π and the reward vector r^π by

$$P_{ss'}^\pi := \sum_a \pi_{sa} P_{sas'}, \quad r_s^\pi := \sum_a \pi_{sa} r_{sa}.$$

The vector v^π then satisfies the Bellman equation

$$v^\pi = r^\pi + \gamma P^\pi v^\pi.$$

Since $0 < \gamma < 1$ and P^π is a transition probability kernel, $(I - \gamma P^\pi)$ is invertible and

$$v^\pi = (I - \gamma P^\pi)^{-1} r^\pi. \tag{1}$$

The goal of MDP is to find the policy that maximizes $e^\top v^\pi$ for some positive weight vector e. In fact, the optimal policy is independent of e (see [22, 28]). Using (1), the problem becomes finding the optimal policy $\pi^\star$ such that

$$\pi^\star := \operatorname*{argmax}_\pi \, e^\top (I - \gamma P^\pi)^{-1} r^\pi.$$

According to the theory of MDP, there is a deterministic optimal policy $\pi^\star$ such that

$$v_s^{\pi^\star} = \max_a [r_{sa} + \gamma P_{sa\cdot} v^{\pi^\star}].$$

Another related function is the state-action value function, defined for a fixed policy π as

$$q_{sa}^\pi := \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r_{s_k a_k} \mid s_0 = s, a_0 = a \right].$$

It is related to the value function v^π via

$$q_{sa}^\pi = r_{sa} + \gamma P_{sa\cdot} v^\pi.$$

At the optimal policy $\pi^\star$,

$$q_{sa}^{\pi^\star} = r_{sa} + \gamma P_{sa\cdot} v^{\pi^\star} = r_{sa} + \gamma P_{sa\cdot} \left[\max_{a' \in A} q_{\cdot a'}^{\pi^\star} \right].$$

Since the rewards are assumed to be bounded in $[0, 1]$, $0 \le v_s^\pi, q_{sa}^\pi \le \frac{1}{1-\gamma}$.

2.2 Finite-Horizon

An undiscounted finite-horizon MDP is similarly described by a tuple $(S, A, \{P_h\}, \{r_h\}, H)$, where H is the horizon, i.e., the number of steps of each episode. P_h and r_h are, respectively, the transition probability tensor and the reward tensor at time step h. For a fixed triple (h, s, a), $\sum_{s'} P_{hsas'} = 1$. Again, we assume that the reward r_{hsa} is bounded without loss of generality in $[0, 1]$.

A policy π is a collection of probabilities over A for each s and h: π_{hsa} is the probability of choosing action a at state s at time h with $\sum_a \pi_{hsa} = 1$. For a given policy π, let the sequence $(s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}, s_H)$ be a stochastic trajectory generated by following the policy π. The value function v^π of the policy π is the expected accumulative reward

$$v_{hs}^\pi := \mathbb{E}_\pi \left[r_{Hs_H} + \sum_{t=h}^{H-1} r_{ts_t a_t} \mid s_h = s \right], \quad h = 0, 1, \ldots, H-1,$$

with the terminal condition $v_{Hs} = r_{Hs}$. By defining the transition matrix P^π and reward vector r^π

$$P_{hss'}^\pi := \sum_a \pi_{hsa} P_{hsas'}, \quad r_{hs}^\pi := \sum_a \pi_{hsa} r_{hsa},$$

we notice that the value function satisfies the finite-horizon Bellman equation:

$$v_h^\pi = r_h^\pi + P_h^\pi v_{h+1}^\pi, \quad h = 0, 1, \ldots, H-1.$$

The goal is to seek an optimal deterministic policy that is guaranteed to exist by MDP theory. At the optimal policy $\pi^\star$,

$$v_{hs}^{\pi^\star} = \max_a [r_{hsa} + P_{hsa} \cdot v_{h+1}^{\pi^\star}], \quad h = 0, 1, \ldots, H-1.$$

The associated state-value function is defined as

$$q_{hsa}^\pi := \mathbb{E}_\pi \left[\sum_{t=h}^{H-1} r_{ts_t a_t} \mid s_h = s, a_h = a \right].$$

At the optimal policy $\pi^\star$, it satisfies the following Bellman optimality condition

$$q_{hsa}^{\pi^\star} = r_{hsa} + P_{hsa} \cdot v_{h+1}^{\pi^\star} = r_{hsa} + \mathbb{E}_{s' \sim P_{hsa}\cdot} \left[\max_{a' \in \mathcal{A}} q_{h+1,s'a'}^{\pi^\star} \right].$$

The reason for choosing the discounted infinite-horizon MDP and the undiscounted finite-horizon MDP is as follows. In the finite-horizon case, a discounted model can be converted to an undiscounted model by letting $\tilde{r}_{hsa} = \gamma^h r_{hsa}$ and $\tilde{r}_{Hs} = \gamma^H r_{hs}$. Thus, we can focus our study on the undiscounted setting without loss of generality. In the infinite-horizon case, the discount factor is needed for both theoretical and computational reasons: otherwise, the accumulative reward would be unbounded even if the reward function r is bounded, and the Bellman equation would be singular because $(I - P^\pi)$ is not invertible. From a practical perspective, the discount factor is also natural because $1 - \gamma$ corresponds to the risk-free interest rate in economics. In other fields, the discount factor is also useful for representing a trade-off between short-term and long-term rewards. One can adjust the discount factor closer to 1 if one wants to learn a policy that values the short-term and long-term rewards almost equally while moving it away from 1 if one attaches more importance to short-term

rewards. For a more comprehensive presentation of MDP models, we refer interested readers to [28].

3 Infinite-Sample Analysis of Action Noise

In this section, we give an infinite-sample analysis of error due to action noise for both finite-horizon MDPs and discounted infinite-horizon MDPs. Here *infinite-sample* means that the reward r and the transition P can be obtained with zero error. Therefore, the only sources of error are the action noise and the distributional shift.

3.1 Infinite-Horizon

We model the action noise by a tensor $\eta = (\eta_{saj})_{s\in S, a, j\in A}$, where η_{saj} is the rate of classifying action a as action j at state s. In particular, $\sum_j \eta_{saj} = 1$.

Let $\bar{\pi}$ be the *behavior policy* from which the offline data was collected. Due to the action noise, instead of the true behavior policy $\bar{\pi}_{sa}$, we are faced with the *polluted* behavior policy

$$\bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}$$

at state s and action a. As a consequence, the reward tensor $\breve{r}$ and the probability tensor $\breve{P}$ polluted by the action noise are given by

$$\breve{r}_{sa} := \frac{\bar{\pi}_{sa}\eta_{saa} r_{sa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} r_{sj}}{\bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}}, \tag{2}$$

$$\breve{P}_{sa\cdot} := \frac{\bar{\pi}_{sa}\eta_{saa} P_{sa\cdot} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} P_{sj\cdot}}{\bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}}, \tag{3}$$

respectively. The following lemma characterizes the error between (P, r) and $(\breve{P}, \breve{r})$.

Lemma 1 *Define* $c_\pi := \min_{sa} \bar{\pi}_{sa} > 0$ *and* $c_\eta := \max_{sa} \sum_{j\neq a} \eta_{saj} < 1$. *Then*

$$\max_{s,a} |r_{sa} - \breve{r}_{sa}| \leq \frac{c_\eta}{c_\pi(1 - c_\eta)}, \tag{4}$$

$$\max_{s,a} \| P_{sa\cdot} - \breve{P}_{sa\cdot} \|_1 \leq \frac{2c_\eta}{c_\pi(1 - c_\eta)}. \tag{5}$$

Proof Since $\max_{j\neq a} \eta_{sja} \leq \sum_{a\neq j} \eta_{sja} \leq c_\eta$, we have

$$\sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} \leq \left(\sum_{j\neq a} \bar{\pi}_{sj}\right)\left(\max_{j\neq a} \eta_{sja}\right) \leq 1 \cdot c_\eta = c_\eta. \tag{6}$$

On the other hand, we have

$$\begin{aligned} \bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} &= \bar{\pi}_{sa}(1 - \sum_{j\neq a} \eta_{saj}) \\ + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} &\geq (\min_{sa} \bar{\pi}_{sa})(1 - \max_{sa} \sum_{j\neq a} \eta_{saj}) = c_\pi(1 - c_\eta). \end{aligned}$$

Combing with (6) and (2), we have

$$
\begin{aligned}
|r_{sa} - \breve{r}_{sa}| &= \left| r_{sa} - \frac{\bar{\pi}_{sa}\eta_{saa} r_{sa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} r_{sj}}{\bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}} \right| \\
&= \left| \frac{\sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}(r_{sa} - r_{sj})}{\bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}} \right| \\
&\leq \frac{1}{c_\pi(1-c_\eta)} \left(\sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} \right) \max_{s,a,j} |r_{sa} - r_{sj}| \leq \frac{c_\eta}{c_\pi(1-c_\eta)}.
\end{aligned}
$$

Similarly, from (6) and (3),

$$
\begin{aligned}
\| P_{sa\cdot} - \breve{P}_{sa\cdot} \|_1 &= \left\| \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} \frac{P_{sa\cdot} - P_{sj\cdot}}{\bar{\pi}_{sa}\eta_{saa} + \sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja}} \right\|_1 \\
&\leq \frac{1}{c_\pi(1-c_\eta)} \left(\sum_{j\neq a} \bar{\pi}_{sj}\eta_{sja} \right) \max_{s,a,j} \| P_{sa\cdot} - P_{sj\cdot} \|_1 \leq \frac{2c_\eta}{c_\pi(1-c_\eta)}.
\end{aligned}
$$

□

Given $\breve{r}$ and $\breve{P}$, the MDP model polluted by the action noise is $(S, A, \breve{P}, \breve{r}, \gamma)$. Let us define

- $\breve{\pi}^\star$ to be the optimal policy for $(S, A, \breve{P}, \breve{r}, \gamma)$,
- $\breve{v}^{\breve{\pi}^\star}$ to be the value function at $\breve{\pi}^\star$ for $(S, A, \breve{P}, \breve{r}, \gamma)$, i.e.,

$$
\breve{v}_s^{\breve{\pi}^\star} := \max_a \left[\breve{r}_{sa} + \gamma \mathbb{E}_{s'\sim \breve{P}_{sa\cdot}} \breve{v}_{s'}^{\breve{\pi}^\star} \right]. \tag{7}
$$

Based on Lemma 1, one has the following error bounds on the value function.

Theorem 2 *With the notations of Lemma 1,*

$$
\| v^{\pi^\star} - \breve{v}^{\breve{\pi}^\star} \|_\infty \leq \frac{c_\eta}{(1-\gamma)^2 c_\pi (1-c_\eta)}, \tag{8}
$$

$$
\| v^{\pi^\star} - v^{\breve{\pi}^\star} \|_\infty \leq \frac{2c_\eta}{(1-\gamma)^2 c_\pi (1-c_\eta)}. \tag{9}
$$

Remark 1 The first error bound is between the optimal policy $\pi^\star$ of the true model (P, r) applied to (P, r) and the optimal policy $\breve{\pi}^\star$ of the polluted model $(\breve{P}, \breve{r})$ applied to $(\breve{P}, \breve{r})$. On the other hand, the second error bound is between the optimal policy $\pi^\star$ of (P, r) applied to (P, r) and the optimal policy $\breve{\pi}^\star$ of $(\breve{P}, \breve{r})$ also applied to (P, r). Both are under control when the misclassification rate c_η is small and c_π is bounded from below.

Proof To prove (8), from the solution of two MDPs (S, A, P, r, γ) and $(S, A, \breve{P}, \breve{r}, \gamma)$ we have

$$
\begin{aligned}
v_s^{\pi^\star} &= \max_a [r_{sa} + \gamma P_{sa\cdot} v^{\pi^\star}], \\
\breve{v}_s^{\breve{\pi}^\star} &= \max_a [\breve{r}_{sa} + \gamma \breve{P}_{sa\cdot} \breve{v}^{\breve{\pi}^\star}].
\end{aligned} \tag{10}
$$

Denote the optimal actions of (10) by a and $\breve{a}$, respectively. By dropping the superscripts in both for notational simplicity, we obtain

$$\begin{aligned} v_s &= r_{sa} + \gamma P_{sa\cdot} v, \\ \breve{v}_s &= \breve{r}_{s\breve{a}} + \gamma \breve{P}_{s\breve{a}\cdot} \breve{v}. \end{aligned}$$

By the optimality of $\breve{a}$, we have

$$v_s - \breve{v}_s = r_{sa} + \gamma P_{sa\cdot} v - (\breve{r}_{s\breve{a}} + \gamma \breve{P}_{s\breve{a}\cdot}) \le r_{sa} + \gamma P_{sa\cdot} v - (\breve{r}_{sa} + \gamma \breve{P}_{sa\cdot} \breve{v}).$$

Rearranging the terms, we have

$$\begin{aligned} v_s - \breve{v}_s &\le (r_{sa} - \breve{r}_{sa}) + \gamma (P_{sa\cdot} v - \breve{P}_{sa\cdot} \breve{v}) \\ &= (r_{sa} - \breve{r}_{sa}) + \gamma P_{sa\cdot} (v - \breve{v}) + \gamma (P_{sa\cdot} - \breve{P}_{sa\cdot}) \breve{v} \\ &= (r_{sa} - \breve{r}_{sa}) + \gamma P_{sa\cdot} (v - \breve{v}) + \gamma (P_{sa\cdot} - \breve{P}_{sa\cdot}) \left(\breve{v} - \frac{1}{2(1-\gamma)} \mathbf{1} \right), \end{aligned}$$

where the last equality results from the fact that $P_{sa\cdot}\mathbf{1} = \breve{P}_{sa\cdot}\mathbf{1} = 1$. Then by Lemma 1,

$$\begin{aligned} v_s - \breve{v}_s &\le \max_{s,a} |r_{sa} - \breve{r}_{sa}| + \gamma \| P_{sa\cdot} \|_1 \| v - \breve{v} \|_\infty + \gamma \| P_{sa\cdot} - \breve{P}_{sa\cdot} \|_1 \left\| \breve{v} - \frac{1}{2(1-\gamma)} \mathbf{1} \right\|_\infty \\ &\le \frac{c_\eta}{c_\pi (1 - c_\eta)} + \gamma \| v - \breve{v} \|_\infty + \gamma \frac{2 c_\eta}{c_\pi (1 - c_\eta)} \cdot \frac{1}{2(1-\gamma)} \\ &= \frac{c_\eta}{c_\pi (1 - c_\eta)} + \gamma \| v - \breve{v} \|_\infty + \frac{\gamma c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)} \\ &= \gamma \| v - \breve{v} \|_\infty + \frac{c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)}. \end{aligned}$$

Similarly, with the optimality of a and the result of Lemma 1 we have,

$$\begin{aligned} \breve{v}_s - v_s &\le \breve{r}_{s\breve{a}} + \gamma \breve{P}_{s\breve{a}\cdot} \breve{v} - (r_{s\breve{a}} + \gamma P_{s\breve{a}\cdot} v) \\ &= (\breve{r}_{s\breve{a}} - r_{s\breve{a}}) + \gamma (\breve{P}_{s\breve{a}\cdot} \breve{v} - P_{s\breve{a}\cdot} v) \\ &= (\breve{r}_{s\breve{a}} - r_{s\breve{a}}) + \gamma \breve{P}_{s\breve{a}\cdot} (\breve{v} - v) + \gamma (\breve{P}_{s\breve{a}\cdot} - P_{s\breve{a}\cdot}) v \\ &= (\breve{r}_{s\breve{a}} - r_{s\breve{a}}) + \gamma \breve{P}_{s\breve{a}\cdot} (\breve{v} - v) + \gamma (\breve{P}_{s\breve{a}\cdot} - P_{s\breve{a}\cdot}) \left(v - \frac{1}{2(1-\gamma)} \mathbf{1} \right) \\ &\le \max_{s,a} |\breve{r}_{sa} - r_{sa}| + \gamma \| \breve{P}_{s\breve{a}\cdot} \|_1 \| \breve{v} - v \|_\infty + \gamma \| \breve{P}_{s\breve{a}\cdot} - P_{s\breve{a}\cdot} \|_1 \| v \|_\infty \\ &\le \frac{c_\eta}{c_\pi (1 - c_\eta)} + \gamma \| \breve{v} - v \|_\infty + \frac{\gamma c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)} \\ &= \gamma \| v - \breve{v} \|_\infty + \frac{c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)}. \end{aligned}$$

As a result, $\| v - \breve{v} \|_\infty \le \gamma \| v - \breve{v} \|_\infty + \frac{c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)}$. This leads to $\| v - \breve{v} \|_\infty \le \frac{c_\eta}{(1-\gamma)^2 c_\pi (1 - c_\eta)}$, which finishes the proof.

To prove (9), from (8) we have

$$\| v^{\pi^\star} - v^{\breve{\pi}^\star} \| \le \| v^{\pi^\star} - \breve{v}^{\breve{\pi}^\star} \| + \| \breve{v}^{\breve{\pi}^\star} - v^{\breve{\pi}^\star} \| \le \frac{c_\eta}{(1-\gamma)^2 c_\pi (1 - c_\eta)} + \| \breve{v}^{\breve{\pi}^\star} - v^{\breve{\pi}^\star} \|. \tag{11}$$

Dropping the superscripts for simplicity, we have

$$
\begin{aligned}
|\check{v}_s - v_s| &= |\check{r}_{s\check{a}} + \gamma \check{P}_{s\check{a}\cdot}\check{v} - (r_{s\check{a}} + \gamma P_{s\check{a}\cdot}v)| \\
&= |(\check{r}_{s\check{a}} - r_{s\check{a}}) + \gamma(\check{P}_{s\check{a}\cdot}\check{v} - P_{s\check{a}\cdot}v)| \\
&= |(\check{r}_{s\check{a}} - r_{s\check{a}}) + \gamma \check{P}_{s\check{a}\cdot}(\check{v} - v) + \gamma(\check{P}_{s\check{a}\cdot} - P_{s\check{a}\cdot})v| \\
&\le |\check{r}_{s\check{a}} - r_{s\check{a}}| + \gamma|\check{P}_{s\check{a}\cdot}(\check{v} - v)| + \gamma \left|(\check{P}_{s\check{a}\cdot} - P_{s\check{a}\cdot})\left(v - \frac{1}{2(1-\gamma)}\mathbf{1}\right)\right| \\
&\le \max_{s,\check{a}}|\check{r}_{s\check{a}} - r_{s\check{a}}| + \gamma\|\check{P}_{s\check{a}\cdot}\|_1\|\check{v} - v\|_\infty + \gamma\|\check{P}_{s\check{a}\cdot} - P_{s\check{a}\cdot}\|_1 \left\|v - \frac{1}{2(1-\gamma)}\mathbf{1}\right\|_\infty \\
&\le \frac{c_\eta}{c_\pi(1-c_\eta)} + \gamma\|\check{v} - v\|_\infty + \gamma\frac{2c_\eta}{c_\pi(1-c_\eta)} \cdot \frac{1}{2(1-\gamma)} \\
&= \frac{c_\eta}{c_\pi(1-c_\eta)} + \gamma\|\check{v} - v\|_\infty + \frac{\gamma c_\eta}{(1-\gamma)c_\pi(1-c_\eta)} \\
&= \gamma\|v - \check{v}\|_\infty + \frac{c_\eta}{(1-\gamma)c_\pi(1-c_\eta)}.
\end{aligned}
$$

Maximizing over s and rearranging the terms gives

$$
\|\check{v} - v\|_\infty \le \frac{c_\eta}{(1-\gamma)^2 c_\pi(1-c_\eta)}.
$$

Combining this inequality with (11) yields (9). □

3.2 Finite-Horizon

The above analysis can be readily extended to finite-horizon MDPs. With $\bar{\pi}$ denoting the behavior policy, we first define the polluted rewards and transition matrices in the finite-horizon setting

$$
\begin{aligned}
\check{r}_{hsa} &:= \frac{\bar{\pi}_{hsa}\eta_{saa}r_{hsa} + \sum_{j\neq a}\bar{\pi}_{hsj}\eta_{sja}r_{hsj}}{\bar{\pi}_{hsa}\eta_{saa} + \sum_{j\neq a}\bar{\pi}_{hsj}\eta_{sja}}, \\
\check{P}_{hsa\cdot} &:= \frac{\bar{\pi}_{hsa}\eta_{saa}P_{hsa\cdot} + \sum_{j\neq a}\bar{\pi}_{hsj}\eta_{sja}P_{hsj\cdot}}{\bar{\pi}_{hsa}\eta_{saa} + \sum_{j\neq a}\bar{\pi}_{hsj}\eta_{sja}}.
\end{aligned}
$$

The following lemma gives the error of the MDP model.

Lemma 3 *Define* $c_\pi := \min_{hsa}\bar{\pi}_{hsa} > 0$ *and* $c_\eta := \max_{sa}\sum_{j\neq a}\eta_{saj} < 1$. *Then*

$$
\max_{h,s,a}|r_{hsa} - \check{r}_{hsa}| \le \frac{c_\eta}{c_\pi(1-c_\eta)}, \tag{12}
$$

$$
\max_{h,s,a}\|P_{hsa\cdot} - \check{P}_{hsa\cdot}\|_1 \le \frac{2c_\eta}{c_\pi(1-c_\eta)}. \tag{13}
$$

The proof is omitted since for each fixed h it is the same as the one for Lemma 1. We define

- $\check{\pi}^\star$ to be the optimal policy for the polluted MDP $(S, A, \{\check{P}_h\}, \{\check{r}_h\}, H)$,
- $\check{v}^{\check{\pi}^\star}$ to be the value function of $\check{\pi}^\star$ on $(S, A, \{\check{P}_h\}, \{\check{r}_h\}, H)$, i.e.,

$$
\check{v}_{hs}^{\check{\pi}^\star} := \max_a\left[\check{r}_{hsa} + \gamma\check{P}_{hsa\cdot}\check{v}_{h+1}^{\check{\pi}^\star}\right], \quad h = 0, 1, \ldots, H-1. \tag{14}
$$

Based on Lemma 3, we have the following value function error bound for finite-horizon MDPs.

Theorem 4 *With the notations of Lemma 3,*

$$\|v_h^{\pi^\star} - \check{v}_h^{\check{\pi}^\star}\|_\infty \le \frac{(H-h)(H-h+3)c_\eta}{2c_\pi(1-c_\eta)}, \tag{15}$$

$$\|v_h^{\pi^\star} - v_h^{\check{\pi}^\star}\|_\infty \le \frac{(H-h)(H-h+3)c_\eta}{c_\pi(1-c_\eta)}. \tag{16}$$

Proof To prove (15), we have from the optimality conditions of $(S, A, \{P_h\}, \{r_h\}, H)$ and $(S, A, \{\check{P}_h\}, \{\check{r}_h\}, H)$

$$v_{hs}^{\pi^\star} = \max_a [r_{hsa} + \gamma P_{hsa\cdot} v_{h+1}^{\pi^\star}],$$
$$\check{v}_{hs}^{\check{\pi}^\star} = \max_a [\check{r}_{hsa} + \gamma \check{P}_{hsa\cdot} \check{v}_{h+1}^{\check{\pi}^\star}].$$

Let the optimal action of the true model and estimated model be a and $\check{a}$, respectively. Again, dropping the superscripts for simplicity leads to

$$v_{hs} = r_{hsa} + P_{hsa\cdot} v_{h+1},$$
$$\check{v}_{hs} = \check{r}_{hs\check{a}} + \check{P}_{hs\check{a}\cdot} \check{v}_{h+1}.$$

The derivation of the key inequalities is similar to the ones in Theorem 2. We first use the optimality of $\check{a}$ to rearrange the terms and arrive at

$$\begin{aligned} v_{hs} - \check{v}_{hs} &\le r_{hsa} + P_{hsa\cdot} v_{h+1} - (\check{r}_{hsa} + \check{P}_{hsa\cdot} \check{v}_{h+1}) \\ &= (r_{hsa} - \check{r}_{hsa}) + (P_{hsa\cdot} v_{h+1} - \check{P}_{hsa\cdot} \check{v}_{h+1}) \\ &= (r_{hsa} - \check{r}_{hsa}) + P_{hsa\cdot}(v_{h+1} - \check{v}_{h+1}) + (P_{hsa\cdot} - \check{P}_{hsa\cdot})\check{v}_{h+1} \\ &= (r_{hsa} - \check{r}_{hsa}) + P_{hsa\cdot}(v_{h+1} - \check{v}_{h+1}) + (P_{hsa\cdot} - \check{P}_{hsa\cdot})\left(\check{v}_{h+1} - \frac{H-h}{2}\mathbf{1}\right). \end{aligned}$$

Then we apply Lemma 3 to get

$$\begin{aligned} &\le \max_{h,s,a} |r_{hsa} - \check{r}_{hsa}| + \|P_{hsa\cdot}\|_1 \|v_{h+1} - \check{v}_{h+1}\|_\infty + \|P_{hsa\cdot} \\ &\quad - \check{P}_{hsa\cdot}\|_1 \left\| \check{v}_{h+1} - \frac{H-h}{2}\mathbf{1} \right\|_\infty \\ &\le \frac{c_\eta}{c_\pi(1-c_\eta)} + \|v_{h+1} - \check{v}_{h+1}\|_\infty + \frac{2c_\eta}{c_\pi(1-c_\eta)} \cdot \frac{H-h}{2} \\ &\le \|v_{h+1} - \check{v}_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1), \end{aligned}$$

where we have used the direct upper bound $v_h \le (H+1-h)$. By symmetry,

$$\begin{aligned} \check{v}_{hs} - v_{hs} &\le \max_{s,a} |\check{r}_{hsa} - r_{hsa}| + \|\check{P}_{hs\check{a}\cdot}\|_1 \|\check{v}_{h+1} - v_{h+1}\|_\infty + \|\check{P}_{hs\check{a}\cdot} \\ &\quad - P_{hs\check{a}\cdot}\|_1 \left\| v_{h+1} - \frac{H-h}{2}\mathbf{1} \right\|_\infty \\ &\le \|\check{v}_{h+1} - v_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1). \end{aligned}$$

As a result, $\|v_h - \check{v}_h\|_\infty \le \|v_{h+1} - \check{v}_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1)$.

Summing this estimate over h and using the fact that $\check{v}_{H\cdot} = v_{H\cdot}$, we obtain

$$\begin{aligned}\|v_h - \check{v}_h\|_\infty &\le \sum_{i=h}^{H-1} \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1) + \|v_H - \check{v}_H\|_\infty \\ &= \frac{(H-h)((H-h+1)+2)c_\eta}{2c_\pi(1-c_\eta)} + 0 \\ &= \frac{(H-h)(H-h+3)c_\eta}{2c_\pi(1-c_\eta)}\end{aligned}$$

which is (15).

To prove (16), notice from (15) that

$$\|v_h^{\pi^\star} - v_h^{\check{\pi}^\star}\| \le \|v_h^{\pi^\star} - \check{v}_h^{\check{\pi}^\star}\| + \|\check{v}_h^{\check{\pi}^\star} - v_h^{\check{\pi}^\star}\| \le \frac{(H-h)(H-h+3)c_\eta}{2c_\pi(1-c_\eta)} + \|v_h^{\check{\pi}^\star} - \check{v}_h^{\check{\pi}^\star}\|. \tag{17}$$

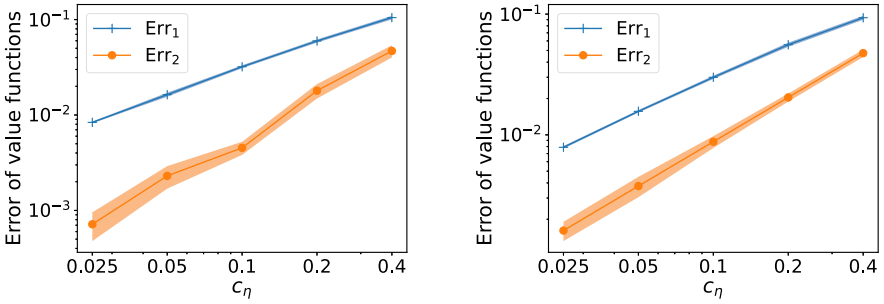
Dropping the superscript $\check{\pi}^\star$ for simplicity gives

$$\begin{aligned}|v_{hs} - \check{v}_{hs}| &= |r_{hs\check{a}} + P_{hs\check{a}\cdot} v_{h+1} - (\check{r}_{hs\check{a}} + \check{P}_{hs\check{a}\cdot}\check{v}_{h+1})| \\ &= |(r_{hs\check{a}} - \check{r}_{hs\check{a}}) + (P_{hs\check{a}\cdot} v_{h+1} - \check{P}_{hs\check{a}\cdot}\check{v}_{h+1})| \\ &= |(r_{hs\check{a}} - \check{r}_{hs\check{a}}) + P_{hs\check{a}\cdot}(v_{h+1} - \check{v}_{h+1}) + (P_{hs\check{a}\cdot} - \check{P}_{hs\check{a}\cdot})\check{v}_{h+1}| \\ &\le |r_{hs\check{a}} - \check{r}_{hs\check{a}}| + |P_{hs\check{a}\cdot}(v_{h+1} - \check{v}_{h+1})| + \left|(P_{hs\check{a}\cdot} - \check{P}_{hs\check{a}\cdot})\left(\check{v}_{h+1} - \frac{H-h}{2}\mathbf{1}\right)\right| \\ &\le \max_{h,s,\check{a}} |r_{hs\check{a}} - \check{r}_{hs\check{a}}| + \|P_{hs\check{a}\cdot}\|_1 \|v_{h+1} - \check{v}_{h+1}\|_\infty \\ &\quad + \|P_{hs\check{a}\cdot} - \check{P}_{hs\check{a}\cdot}\|_1 \left\|\check{v}_{h+1} - \frac{H-h}{2}\mathbf{1}\right\|_\infty \\ &\le \frac{c_\eta}{c_\pi(1-c_\eta)} + \|v_{h+1} - \check{v}_{h+1}\|_\infty + \frac{2c_\eta}{c_\pi(1-c_\eta)} \cdot \frac{H-h}{2} \\ &\le \|v_{h+1} - \check{v}_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1).\end{aligned}$$

Maximizing over s and combining with (17) gives (16). □

3.3 Numerical Results

In the first example, we consider the infinite-horizon case, where the model $(\check{P}, \check{r})$ is given in (3). The true MDP used in this experiment is adapted from [42], where $|S| = 200$, $|A| = 50$, and the transition probabilities and rewards are randomly generated in the following way. The transition probabilities are defined as $P_{sas'} = 1/20$ for 20 states s' chosen uniformly randomly from S and zero for other states. The true reward r_{sa} is defined as $U_{sa}U_s$, where U_{sa} and U_s are independently uniformly sampled from $[0, 1]$. The behavior policy is assumed to be the uniform policy, thus $c_\pi = 1/50$. The discount rate γ is set as 0.9. Figure 1a plots the relative errors of the value function as a function of the action noise measure c_η (defined in Lemma 1), where the reference solution is calculated with a policy gradient method starting from the uniform policy. The shaded area around the two curves gives the standard deviation calculated from 10 randomly generated η. The action noise tensor η is generated as follows. For each



(a) The errors of the value functions in the infinite-horizon case.

(b) The errors of the value functions in the finite-horizon case.

Fig. 1 The relative errors of the value functions caused by action noise. The blue and orange curves show how the errors change as a function of c_η. The shaded area around the curves indicates the standard deviation. The logarithmic scale is used for both axes. Here $\mathrm{Err}_1 = \|\check{v}^{\check{\pi}^\star} - v^{\pi^\star}\|/\|v^{\pi^\star}\|$ and $\mathrm{Err}_2 = \|v^{\check{\pi}^\star} - v^{\pi^\star}\|/\|v^{\pi^\star}\|$, where $\|\cdot\|$ is the infinity norm. Both errors decrease when c_η goes to zero, where the trend coincides with the theoretical predictions in Theorem 2 and Theorem 4 (Color figure online)

(s, a), η_{saa} is uniformly sampled from $[1 - c_\eta, 1]$, and then each η_{saj} is first generated from a uniform random variable on $[0, 1]$ and then normalized so that $\sum_j \eta_{saj} = 1$. We generate η in the same way for all examples in this paper. The two curves match the theoretical estimates (8) and (9), respectively. When the action noise goes to zero, the value function errors vanish.

In the second example, we consider the finite-horizon case. Similar to the first example, we define the transition probabilities as $P_{hsas'} = 1/20$ for s' belonging to a uniformly randomly chosen subset with 20 elements and zero everywhere else. The reward r_{hsa} is defined as the product $U_{hsa}U_sU_h$ where U_{hsa}, U_s and U_h are independent samples of the uniform random variable on $[0, 1]$, and the terminal reward r_{Hs} is also independently uniformly sampled from $[0, 1]$. The horizon H is chosen as 10. Figure 1b shows the relative errors of the value function, with the standard deviation obtained from a collection of 10 random η. Here the reference solution is calculated with a linear programming method. Both types of errors decrease monotonically when c_η decreases, compatible with the result of Theorem 4.

4 Finite-Sample Analysis with Action Noise

In the previous section, we considered the error bounds in the infinite-sample limit, where the error only comes from the action noise and distributional shift. In practice, the rewards and transition rates need to be estimated from a finite number of samples, which introduces an additional source of statistical error.

4.1 Infinite-Horizon

To model the finite-sample case, we denote by $\hat{P}_{sas'}$ and $\hat{r}_{sa}$ the transitions and rewards estimated from the finite-sample dataset. Here, we assume that a few samples are collected for each state-action pair (s, a) independently according to the behavior policy. There are two key quantities.

- N_{spl} is the minimum of the number of transitions from (s, a) over all state-action pairs.

- For each pair (s, a), let S_{sa} be the set of states s' such that $\breve{P}_{sas'} > 0$. Define $N_{\text{nb}} := \max_{s,a} |S_{sa}|$.

It is worth noticing that these quantities are defined for the polluted dynamics $\breve{P}$ rather than the true dynamics. The following lemma is concerned with the statistical error in $\hat{P}_{sas'}$ and $\hat{r}_{sa}$.

Lemma 5 *With probability at least* $1 - \delta$,

$$\max_{s,a} |\breve{r}_{sa} - \hat{r}_{sa}| \le \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \tag{18}$$

and

$$\max_{s,a} \|\breve{P}_{sa\cdot} - \hat{P}_{sa\cdot}\|_1 \le 2\sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}. \tag{19}$$

Proof By Hoeffding's inequality

$$\mathbb{P}\left(|\breve{r}_{sa} - \hat{r}_{sa}| \ge \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}}\right) \le 2\exp\left(-2N_{\text{spl}} \cdot \left(\sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}}\right)^2\right)$$
$$= 2\exp\left(-\log(4|S||A|/\delta)\right) = \frac{\delta}{2|S||A|}.$$

Then, by the union bound,

$$\mathbb{P}\left(\max_{s,a} |\breve{r}_{sa} - \hat{r}_{sa}| \le \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}}\right) \le \sum_{s,a} \mathbb{P}\left(|\breve{r}_{sa} - \hat{r}_{sa}| \ge \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}}\right)$$
$$\le |S||A| \cdot \frac{\delta}{2|S||A|} = \frac{\delta}{2},$$

which proves that the event in (18) holds with probability at least $1 - \delta/2$. Therefore, it suffices to show that the event in (19) holds with probability at least $1 - \delta/2$. The proof is inspired by the argument of [12]. Consider the set $\mathcal{U}_{sa}$ of vectors $u \in \mathbb{R}^{|S|}$ such that $|u_j| = 1$ for state $j \in S_{sa}$ and $u_j = 0$ for the remaining coordinates. Then the ℓ_1 norm in (19) can be expressed by

$$\max_{u \in \mathcal{U}_{sa}} u^\top \left(\hat{P}_{sa\cdot} - \breve{P}_{sa\cdot}\right).$$

Notice that $u^\top \hat{P}_{sa\cdot}$ is the finite-sample average of i.i.d. random variables $u^\top e_{s'}$, where $e_{s'}$ is the s'-th standard basis vector, and $u^\top \breve{P}_{sa\cdot}$ is the expectation. Applying Hoeffding's inequality and the union bound shows that with probability at least $1 - \delta/2$

$$\max_{s,a} \max_{u \in \mathcal{U}_{sa}} u^\top \left(\hat{P}_{sa\cdot} - \breve{P}_{sa\cdot}\right) \le 2\sqrt{\frac{\log(2|S||A|2^{N_{\text{nb}}}/\delta)}{2N_{\text{spl}}}} = 2\sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}.$$

Therefore, (18) and (19) hold simultaneously with probability at least $1 - \delta$. □

Given $\hat{r}$ and $\hat{P}$, the MDP polluted both by the action noise and the finite-sample statistical error is $(S, A, \hat{P}, \hat{r}, \gamma)$. We define

- $\hat{\pi}^\star$ to be the optimal policy for $(S, A, \hat{P}, \hat{r}, \gamma)$, and
- $\hat{v}^{\hat{\pi}^\star}$ to be the value function of $\hat{\pi}^\star$ on $(S, A, \hat{P}, \hat{r}, \gamma)$, i.e.,

$$\hat{v}_s^{\hat{\pi}^\star} := \max_a \left[\hat{r}_{sa} + \gamma \mathbb{E}_{s' \sim \hat{P}_{sa\cdot}} \hat{v}_{s'}^{\hat{\pi}^\star} \right]. \tag{20}$$

Equipped with the statistical error bound in Lemma 5, we have the following error bounds for the value functions.

Theorem 6 *Under the assumptions of Lemma 1, with probability at least* $1 - \delta$

$$\begin{aligned} \|v^{\pi^\star} - \hat{v}^{\hat{\pi}^\star}\|_\infty \le & \frac{c_\eta}{(1-\gamma)^2 c_\pi (1 - c_\eta)} + \frac{1}{1-\gamma} \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\ & + \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}, \end{aligned} \tag{21}$$

and with probability at least $1 - 2\delta$

$$\begin{aligned} \|v^{\pi^\star} - v^{\hat{\pi}^\star}\|_\infty \le & \frac{2c_\eta}{(1-\gamma)^2 c_\pi (1 - c_\eta)} + \frac{2}{1-\gamma} \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\ & + \frac{2\gamma}{(1-\gamma)^2} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}. \end{aligned} \tag{22}$$

Proof From the optimal solutions of (S, A, P, r, γ) and $(S, A, \hat{P}, \hat{r}, \gamma)$,

$$\begin{aligned} v_s^{\pi^\star} &= \max_a [r_{sa} + \gamma P_{sa\cdot} v^{\pi^\star}], \\ \hat{v}_s^{\hat{\pi}^\star} &= \max_a [\hat{r}_{sa} + \gamma \hat{P}_{sa\cdot} \hat{v}^{\hat{\pi}^\star}]. \end{aligned}$$

Let the optimal actions be $a, \hat{a}$, respectively. Dropping the superscripts in both for simplicity leads to

$$\begin{aligned} v_s &= r_{sa} + \gamma P_{sa\cdot} v, \\ \hat{v}_s &= \hat{r}_{s\hat{a}} + \gamma \hat{P}_{s\hat{a}\cdot} \hat{v}. \end{aligned}$$

By Lemmas 1, 5, and the triangle inequality,

$$\max_{s,a} |\hat{r}_{sa} - r_{sa}| \le \frac{c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)} + \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}}, \tag{23}$$

and

$$\max_{s,a} \|\hat{P}_{sa\cdot} - P_{sa\cdot}\|_1 \le \frac{2c_\eta}{(1-\gamma) c_\pi (1 - c_\eta)} + 2\sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}, \tag{24}$$

with probability at least $1-\delta$. Thus, with probability at least $1-\delta$

$$
\begin{aligned}
v_s - \hat{v}_s &\le r_{sa} + \gamma P_{sa\cdot} v - (\hat{r}_{sa} + \gamma \hat{P}_{sa\cdot} \hat{v}) \\
&= r_{sa} - \hat{r}_{sa} + \gamma (P_{sa\cdot} v - \hat{P}_{sa\cdot} \hat{v}) \\
&= r_{sa} - \hat{r}_{sa} + \gamma P_{sa\cdot} (v - \hat{v}) + \gamma (P_{sa\cdot} - \hat{P}_{sa\cdot}) \hat{v} \\
&= r_{sa} - \hat{r}_{sa} + \gamma P_{sa\cdot} (v - \hat{v}) + \gamma (P_{sa\cdot} - \hat{P}_{sa\cdot}) \left(\hat{v} - \frac{1}{2(1-\gamma)} \mathbf{1} \right) \\
&\le \max_{s,a} |r_{sa} - \hat{r}_{sa}| + \gamma \| P_{sa\cdot} \|_1 \| v - \hat{v} \|_\infty + \gamma \| P_{sa\cdot} - \hat{P}_{sa\cdot} \|_1 \left\| \hat{v} - \frac{1}{2(1-\gamma)} \mathbf{1} \right\|_\infty \\
&\le \frac{c_\eta}{c_\pi (1-c_\eta)} + \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} + \gamma \| v - \hat{v} \|_\infty \\
&\quad + \frac{\gamma c_\eta}{(1-\gamma) c_\pi (1-c_\eta)} + \frac{\gamma}{1-\gamma} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}} \\
&= \gamma \| v - \hat{v} \|_\infty + \frac{c_\eta}{(1-\gamma) c_\pi (1-c_\eta)} + \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\
&\quad + \frac{\gamma}{1-\gamma} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}.
\end{aligned} \tag{25}
$$

Similarly, we have

$$
\begin{aligned}
\hat{v}_s - v_s &\le \gamma \| \hat{v} - v \|_\infty + \frac{c_\eta}{(1-\gamma) c_\pi (1-c_\eta)} + \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\
&\quad + \frac{\gamma}{1-\gamma} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}}.
\end{aligned}
$$

As a result,

$$
\begin{aligned}
\| v - \hat{v} \|_\infty &\le \gamma \| v - \hat{v} \|_\infty + \frac{c_\eta}{(1-\gamma) c_\pi (1-c_\eta)} + \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\
&\quad + \frac{\gamma}{1-\gamma} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}},
\end{aligned}
$$

and

$$
\begin{aligned}
\| v - \hat{v} \|_\infty &\le \frac{c_\eta}{(1-\gamma)^2 c_\pi (1-c_\eta)} + \frac{1}{1-\gamma} \sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\
&\quad + \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{N_{\text{nb}} \log 2 + \log(2|S||A|/\delta)}{2N_{\text{spl}}}},
\end{aligned}
$$

which proves (21).

Based on this inequality, with probability at least $1-\delta$

$$\begin{aligned}\|v^{\pi^\star}-v^{\hat{\pi}^\star}\| &\le \|v^{\pi^\star}-\hat{v}^{\hat{\pi}^\star}\|+\|\hat{v}^{\hat{\pi}^\star}-v^{\hat{\pi}^\star}\| \\ &\le \frac{c_\eta}{(1-\gamma)^2 c_\pi(1-c_\eta)}+\frac{1}{1-\gamma}\sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\ &\quad+\frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{N_{\text{nb}}\log 2+\log(2|S||A|/\delta)}{2N_{\text{spl}}}}+\|\hat{v}^{\hat{\pi}^\star}-v^{\hat{\pi}^\star}\|.\end{aligned} \tag{26}$$

Moreover, using the same argument as (25) but with a replaced by $\hat{a}$, one has that with probability at least $1-\delta$

$$\begin{aligned}|v_s-\hat{v}_s| &= |r_{s\hat{a}}+\gamma P_{s\hat{a}\cdot}v-(\hat{r}_{s\hat{a}}+\gamma\hat{P}_{s\hat{a}\cdot}\hat{v})| \\ &\le \gamma\|v-\hat{v}\|_\infty+\frac{c_\eta}{(1-\gamma)c_\pi(1-c_\eta)}+\sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\ &\quad+\frac{\gamma}{1-\gamma}\sqrt{\frac{N_{\text{nb}}\log 2+\log(2|S||A|/\delta)}{2N_{\text{spl}}}}.\end{aligned}$$

Here, we again dropped the superscripts for simplicity. Therefore, with probability at least $1-\delta$

$$\begin{aligned}\|\hat{v}^{\hat{\pi}^\star}-v^{\hat{\pi}^\star}\| &\le \frac{c_\eta}{(1-\gamma)^2 c_\pi(1-c_\eta)}+\frac{1}{1-\gamma}\sqrt{\frac{\log(4|S||A|/\delta)}{2N_{\text{spl}}}} \\ &\quad+\frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{N_{\text{nb}}\log 2+\log(2|S||A|/\delta)}{2N_{\text{spl}}}}.\end{aligned}$$

Combining this inequality with (26) and the union bound yields (22) with probability at least $1-2\delta$. □

4.2 Finite-Horizon

For the finite-horizon case, we similarly introduce $\hat{P}_{hsas'}$ and $\hat{r}_{hsa}$ as the transitions and rewards estimated from a finite number of samples. Assuming that for each triple (h,s,a) there exists a number of recorded transitions from state s under action a at epoch h, we denote

- N_{spl} to be the minimum of the number of transitions from (h,s,a) over all triples.
- $N_{\text{nb}}=\max_{h,s,a}|S_{hsa}|$, where S_{hsa} is the set of states s' such that $P_{hsas'}>0$.

The following lemma on the statistical error is needed in the analysis afterward.

Lemma 7 *With probability at least* $1-\delta$,

$$\max_{h,s,a}|\check{r}_{hsa}-\hat{r}_{hsa}|\le\sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}},\ \max_{s}|\check{r}_{Hs}-\hat{r}_{Hs}|\le\sqrt{\frac{\log(8|S|/\delta)}{2N_{\text{spl}}}}, \tag{27}$$

and

$$\max_{h,s,a}\|\check{P}_{hsa\cdot}-\hat{P}_{hsa\cdot}\|_1\le 2\sqrt{\frac{N_{\text{nb}}\log 2+\log(4H|S||A|/\delta)}{2N_{\text{spl}}}}. \tag{28}$$

Proof By Hoeffding's inequality and the union bound, the event in (27) holds with probability at least $1-\delta/2$, It then suffices to show that the event in (28) holds with probability at least $1-\delta/2$. Using the proof of Lemma 5, we know that

$$\max_{s,a} \|\breve{P}_{hsa\cdot} - \hat{P}_{hsa\cdot}\|_1 \leq 2\sqrt{\frac{N_{\text{nb}}\log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}}$$

holds with probability at least $1-\delta/(2H)$. A union bound over the horizons shows that (28) holds with probability at least $1-\delta/2$ according to the union bound. □

Given the reward $\hat{r}$ and the transition $\hat{P}$, the MDP with both the action noise and the finite-sample statistical error is $(S, A, \{\hat{P}_h\}, \{\hat{r}_h\}, H)$. We now define

- $\hat{\pi}^\star$ to be the optimal policy for the MDP $(S, A, \{\hat{P}_h\}, \{\hat{r}_h\}, H)$, and
- $\hat{v}^{\hat{\pi}^\star}$ to be the value function of $\hat{\pi}^\star$ on $(S, A, \{\hat{P}_h\}, \{\hat{r}_h\}, H)$, i.e.,

$$\hat{v}_{hs}^{\hat{\pi}^\star} := \max_a \left[\hat{r}_{hsa} + \gamma \hat{P}_{hsa\cdot}\hat{v}_{h+1}^{\hat{\pi}^\star}\right]. \tag{29}$$

With the preparation in Lemma 7, we have the following theorem that concerns the finite-horizon setting.

Theorem 8 *Under the assumptions of Lemma 3, with probability at least* $1-\delta$

$$\begin{aligned}\|v_h^{\pi^\star} - \hat{v}_h^{\hat{\pi}^\star}\|_\infty \leq{}& \frac{(H-h+1)(H-h+3)c_\eta}{2c_\pi(1-c_\eta)} + (H+1)\sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} \\ &+ \frac{(H-h)(H-h+1)}{2}\sqrt{\frac{N_{\text{nb}}\log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}},\end{aligned} \tag{30}$$

and with probability at least $1-2\delta$

$$\begin{aligned}\|v_h^{\pi^\star} - v_h^{\hat{\pi}^\star}\|_\infty \leq{}& \frac{(H-h+1)(H-h+3)c_\eta}{c_\pi(1-c_\eta)} + 2(H+1)\sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} \\ &+ (H-h)(H-h+1)\sqrt{\frac{N_{\text{nb}}\log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}}.\end{aligned} \tag{31}$$

Proof From the Bellman optimality conditions of the two MDPs, we have

$$\begin{aligned} v_{hs}^{\pi^\star} &= \max_a [r_{hsa} + \gamma P_{hsa\cdot} v_{h+1}^{\pi^\star}], \\ \hat{v}_{hs}^{\hat{\pi}^\star} &= \max_a [\hat{r}_{hsa} + \gamma \hat{P}_{hsa\cdot}\hat{v}_{h+1}^{\hat{\pi}^\star}]. \end{aligned}$$

Let the optimal actions be a and $\hat{a}$, respectively. Dropping the superscripts gives

$$\begin{aligned} v_{hs} &= r_{hsa} + P_{hsa\cdot} v_{h+1}, \\ \hat{v}_{hs} &= \hat{r}_{hs\hat{a}} + \hat{P}_{hs\hat{a}\cdot}\hat{v}_{h+1}. \end{aligned}$$

Therefore, with probability at least $1-\delta$

$$
\begin{aligned}
v_{hs} - \hat{v}_{hs} &\le r_{hsa} + P_{hsa\cdot} v_{h+1} - \hat{r}_{hsa} - \hat{P}_{hsa\cdot} \hat{v}_{h+1} \\
&= (r_{hsa} - \hat{r}_{hsa}) + (P_{hsa\cdot} v_{h+1} - \hat{P}_{hsa\cdot} \hat{v}_{h+1}) \\
&= (r_{hsa} - \hat{r}_{hsa}) + P_{hsa\cdot}(v_{h+1} - \hat{v}_{h+1}) + (P_{hsa\cdot} - \hat{P}_{hsa\cdot})\hat{v}_{h+1} \\
&= (r_{hsa} - \hat{r}_{hsa}) + P_{hsa\cdot}(v_{h+1} - \hat{v}_{h+1}) + (P_{hsa\cdot} - \hat{P}_{hsa\cdot})\left(\hat{v}_{h+1} - \frac{H-h}{2}\mathbf{1}\right) \\
&\le \max_{h,s,a} |r_{hsa} - \hat{r}_{hsa}| + \|P_{hsa\cdot}\|_1 \|v_{h+1} - \hat{v}_{h+1}\|_\infty + \|P_{hsa\cdot} - \hat{P}_{hsa\cdot}\|_1 \left\|\hat{v}_{h+1} - \frac{H-h}{2}\mathbf{1}\right\|_\infty \\
&\le \frac{c_\eta}{c_\pi(1-c_\eta)} + \sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} + \|v_{h+1} - \hat{v}_{h+1}\|_\infty + \\
&\quad \frac{H-h}{2} \cdot \left(\frac{2c_\eta}{c_\pi(1-c_\eta)} + 2\sqrt{\frac{N_{\text{nb}} \log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}} \right) \\
&\le \|v_{h+1} - \hat{v}_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1) \\
&\quad + \sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} + (H-h)\sqrt{\frac{N_{\text{nb}} \log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}},
\end{aligned}
\tag{32}
$$

where we have used the direct upper bound $v_h \le (H-h+1)$ and the inequalities (23) and (24). By symmetry,

$$
\begin{aligned}
\hat{v}_{hs} - v_{hs} &\le \max_{h,s,a} |\hat{r}_{hsa} - r_{hsa}| + \|\hat{P}_{hs\hat{a}\cdot}\|_1 \|\hat{v}_{h+1} - v_{h+1}\|_\infty \\
&\quad + \|\hat{P}_{hs\hat{a}\cdot} - P_{hs\hat{a}\cdot}\|_1 \|v_{h+1} - \frac{H-h}{2}\mathbf{1}\|_\infty \\
&\le \|\hat{v}_{h+1} - v_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1) + \sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} \\
&\quad + (H-h)\sqrt{\frac{N_{\text{nb}} \log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}}.
\end{aligned}
$$

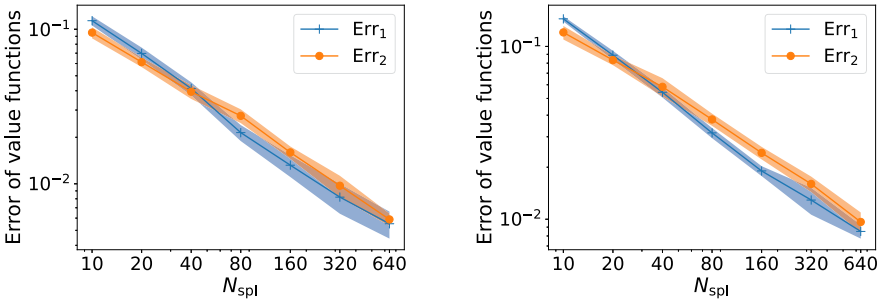
As a result,

$$
\begin{aligned}
\|v_h - \hat{v}_h\|_\infty &\le \|v_{h+1} - \hat{v}_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1) + \sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} \\
&\quad + (H-h)\sqrt{\frac{N_{\text{nb}} \log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}}.
\end{aligned}
$$

Finally, summing over h proves (30).

Using this inequality, we know that with probability at least $1-\delta$,

$$
\|v_h^{\pi^\star} - v_h^{\hat{\pi}^\star}\| \le \|v_h^{\pi^\star} - \hat{v}_h^{\hat{\pi}^\star}\| + \|\hat{v}_h^{\hat{\pi}^\star} - v_h^{\hat{\pi}^\star}\|. \tag{33}
$$



(a) The errors of the value functions in the infinite-horizon case.

(b) The errors of the value functions in the finite-horizon case.

Fig. 2 The relative errors of the value functions in the finite-sample case. The blue and orange curves show the error as a function of N_{spl}. The shaded area around them indicates the standard deviation calculated from 10 randomly generated offline datasets. The logarithmic scale is used for both axes. Here $\text{Err}_1 = \|\hat{v}^{\hat{\pi}^\star} - v^{\pi^\star}\|/\|v^{\pi^\star}\|$ and $\text{Err}_2 = \|v^{\hat{\pi}^\star} - v^{\pi^\star}\|/\|v^{\pi^\star}\|$, where $\|\cdot\|$ is the infinity norm. Both errors decrease when N_{spl} increases, where the trend coincides with the results from Theorem 6 and Theorem 8 (Color figure online)

Using the derivation of (32) with a replaced by $\hat{a}$, with probability at least $1-\delta$

$$
\begin{aligned}
&|v_{hs} - \hat{v}_{hs}| = |r_{hs\hat{a}} + P_{hs\hat{a}}.v_{h+1} - \hat{r}_{hs\hat{a}} - \hat{P}_{hs\hat{a}}.\hat{v}_{h+1}|,\\
&\le \|v_{h+1} - \hat{v}_{h+1}\|_\infty + \frac{c_\eta}{c_\pi(1-c_\eta)}(H-h+1)\\
&+ \sqrt{\frac{\log(8H|S||A|/\delta)}{2N_{\text{spl}}}} + (H-h)\sqrt{\frac{N_{\text{nb}}\log 2 + \log(4H|S||A|/\delta)}{2N_{\text{spl}}}},
\end{aligned}
$$

where we drop the superscripts for simplicity. By applying the union bound and combining with (33), we obtain (31) with probability at least $1-2\delta$. □

4.3 Numerical Results

The first example considers the infinite-horizon problem with finite samples. The behavior policy is a uniform policy, and the model is estimated from the offline dataset from the same MDP used in Sect. 3.3. The reference solution is also computed with the policy gradient method as in Sect. 3.3. The action noise tensor η is also randomly generated such that the assumptions of Lemma 1 are satisfied with the action noise level $c_\eta = 0.01$. From the model description, $N_{\text{nb}} = 20$, which is significantly smaller than $|S| = 500$, justifying the necessity of the improvement in Lemma 5 over the original result in [12]. Figure 2(a) presents the convergence of the two relative value function errors as a function of N_{spl}, with the standard deviation (shown by the shaded area around the curve) calculated from 10 randomly generated datasets. When the number of samples is relatively smaller, the error is significantly larger than the one obtained by the infinite sample limit model, i.e., the statistical error dominates the action noise error. When the number of samples increases, both types of error decrease to the same level as the infinite-sample error.

The second example considers the finite-horizon case. Similar to the first example, we adopt the MDP in Sect. 3.3. We also use the linear programming method to compute the reference solutions. The action noise tensor η is also randomly generated with $c_\eta = 0.01$.

Figure 2b show in the logarithmic scale that Err_1 and Err_2 decrease monotonically with respect to N_{spl}, verifying the theoretical results of Theorems 6 and 8.

5 Conclusion and Discussion

This paper focused on the offline model-based RL with action noise in the tabular setting. As mentioned in Sect. 1.2, the application of RL algorithms in real-world problems is hindered by three components simultaneously, namely the statistical error caused by the finite size of the samples, the distributional shift caused by the difference between the behavior policy and the optimal policy, and the pollution in action information. The statistical error and the distributional shift have been extensively studied in the RL literature. In this work, we emphasize the analyses of the action noise. We first analyzed the infinite-sample case and illustrated that the combined effect of the distributional shift and the action noise can be captured by the term $\frac{c_\eta}{c_\pi(1-c_\eta)}$, and then move on to the finite-sample case with an additional statistical estimation error. In both cases, the error bounds of the estimated model and the value function are derived. The error estimations are also verified with numerical examples.

For the theoretical analysis, we have shown that the action noise interacts with the distributional shift via the term c_η/c_π, while the error caused by a finite sample size is added as an orthogonal component in the error. An interesting direction is to explore more sophisticated ways to integrate these three sources of error (for example, one can potentially use the techniques in [20, 21]). In our theoretical analyses, our prior knowledge of the dataset is represented by the misclassification tensor η. In practice, the parameter η can often be estimated depending on the specific context. For example, in robotic datasets, discrepancies between the intended actions and those actually executed by the robots may arise due to malfunctions of devices. In such cases, η could potentially be estimated using the error rate of the device responsible for executing the actions. Similarly, in medical or biological datasets, action misspecification can occur due to different definitions of treatments, which is particularly relevant when data from one institution (e.g., institution A) are used in a study conducted by another institution (e.g., institution B). In these scenarios, η could be estimated by comparing the different treatment definitions.

Different kinds of pessimism techniques can be used to derive novel algorithms for coping with action noise as well. There are several potential ways to incorporate the uncertainty in the model caused by action noise into the design of pessimism schemes. For example, one can include an additional term in the bonus function used in MILO [4] or in the constraint bound in [34]. One can also introduce a perturbed reward as in [21] or a Bernstein-style confidence penalty as in [20]. Another possibility is to use the analysis presented in Sect. 3 to derive a bound for the MOReL algorithm [14]. We expect that with a few necessary changes, most of the analyses in these papers will continue to hold.

It is also natural to extend the study of action noise to the continuous action setting. With additional assumptions such as the locality or sparsity of the action noise, similar error bounds can be derived for several continuous models, such as the linear model, the kernelized nonlinear regulator (KNR), etc.

Acknowledgements The authors thank Subhajit Sanyal, Branislav Kveton, Sumeet Katariya, Anshuka Rangi, Masatoshi Uehara, Sanath Kumar Krishnamurthy, and Tesi Xiao for the helpful discussion and constructive comments.

Funding The authors have received research support from Amazon Search.

Data Availability The data that support the findings of this study are available from Amazon Search but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Amazon Search.

Declarations

Competing interests The authors have not disclosed any competing interests.

References

1. Agarwal, A., Jiang, N., Kakade, S.M., Sun, S.: Reinforcement Learning: Theory and Algorithms (2019)
2. Agarwal, R., Schuurmans, D., Norouzi, M.: An optimistic perspective on offline reinforcement learning. In: International Conference on Machine Learning, pp. 104–114 (2020)
3. Brandfonbrener, D., Whitney, W., Ranganath, R., Bruna, J.: Offline RL without off-policy evaluation. Adv. Neural Inf. Process. Syst. **34**, 4933–4946 (2021)
4. Chang, J., Uehara, M., Sreenivas, D., Kidambi, R., Sun, W.: Mitigating covariate shift in imitation learning via offline data with partial coverage. Adv. Neural Inf. Process. Syst. **34**, 965–979 (2021)
5. Chen, X., Ghadirzadeh, A, Tianhe, Y., Gao, Y., Wang, J., Li, W., Liang, B., Finn, C., Zhang, C.: Latent-variable advantage-weighted policy optimization for offline RL. arXiv preprint arXiv:2203.08949 (2022)
6. Diehl, C., Sievernich, T., Krüger, M., Hoffmann, F., Bertran, T.: Umbrella: uncertainty-aware model-based offline reinforcement learning leveraging planning. arXiv preprint arXiv:2111.11097 (2021)
7. Duan, Y., Jia, Z., Wang, M.: Minimax-optimal off-policy evaluation with linear function approximation. In: International Conference on Machine Learning, pp. 2701–2709 (2020)
8. Eysenbach, B., Khazatsky, A., Levine, S., Salakhutdinov, R.: Mismatched no more: joint model-policy optimization for model-based RL. arXiv preprint arXiv:2110.02758 (2021)
9. Fujimoto, S., Shixiang, S.G.: A minimalist approach to offline reinforcement learning. Adv. Neural Inf. Process. Syst. **34**, 20132–20145 (2021)
10. Fujimoto, S., Meger, D., Precup, D.: Off-policy deep reinforcement learning without exploration. In: International Conference on Machine Learning, pp. 2052–2062 (2019)
11. Ghasemipour, K., Gu, S.S., Nachum, O.: Why so pessimistic? Estimating uncertainties for offline RL through ensembles, and why their independence matters. arXiv preprint arXiv:2205.13703 (2022)
12. Jiang, N.: Notes on tabular methods (2018)
13. Jin, Y., Yang, Z., Wang, Z.: Is pessimism provably efficient for offline RL? In: International Conference on Machine Learning, pp. 5084–5096 (2021)
14. Kidambi, R., Rajeswaran, A., Netrapalli, P., Joachims, T.: MOReL: model-based offline reinforcement learning. Adv. Neural Inf. Process. Syst. **33**, 21810–21823 (2020)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
16. Kostrikov, I., Nair, A., Levine, S.: Offline reinforcement learning with implicit Q-learning. arXiv preprint arXiv:2110.06169 (2021)
17. Kumar, A., Justin, F., Soh, M., Tucker, G., Levine, S.: Stabilizing off-policy Q-learning via bootstrapping error reduction. Adv. Neural Inf. Process. Syst. **32** (2019)
18. Kumar, A., Zhou, A., Tucker, G., Levine, S.: Conservative Q-learning for offline reinforcement learning. Adv. Neural Inf. Process. Syst. **33**, 1179–1191 (2020)
19. Levine, S., Kumar, A., Tucker, G., Fu, J.: Offline reinforcement learning: tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643 (2020)
20. Li, G., Shi, L., Chen, Y., Chi, Y., Wei, Y.: Settling the sample complexity of model-based offline reinforcement learning. Ann. Stat. **52**(1), 233–260 (2024)
21. Li, G., Wei, Y., Chi, Y., Chen, Y.: Breaking the sample size barrier in model-based reinforcement learning with a generative model. Oper. Res. **72**(1), 203–221 (2024)
22. Li, H., Yu, H.F., Ying L., Dhillon, I.: Accelerating primal-dual methods for regularized Markov decision processes (2022)
23. Lyu, J., Li, X., Lu, Z.: Double check your state before trusting it: confidence-aware bidirectional offline model-based imagination. arXiv preprint arXiv:2206.07989 (2022)
24. Lyu, J., Ma, X., Li, X., Lu, Z.: Mildly conservative q-learning for offline reinforcement learning. arXiv preprint arXiv:2206.04745 (2022)
25. Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., Peters, J.: Robust reinforcement learning: a review of foundations and recent advances. Mach. Learn. Knowl. Extr. **4**(1), 276–315 (2022)

26. Peng, X.B., Kumar, A., Zhang, G., Levine, S.: Advantage-weighted regression: simple and scalable off-policy reinforcement learning. arXiv preprint arXiv:1910.00177 (2019)
27. Prudencio, R.F., Maximo, M.R., Colombini, E.L.: A survey on offline reinforcement learning: taxonomy, review, and open problems. arXiv preprint arXiv:2203.01387 (2022)
28. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken (1994)
29. Rigter, M., Lacerda, B., Hawes, N.: RAMBO-RL: robust adversarial model-based offline reinforcement learning. arXiv preprint arXiv:2204.12581 (2022)
30. Singh, A., Yu, A., Yang, J., Zhang, J., Kumar, A., Levine, S.: COG: connecting new skills to past experience with offline reinforcement learning. arXiv preprint arXiv:2010.14500 (2020)
31. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
32. Swazinna, P., Udluft, S., Runkler, T.: Overcoming model bias for robust offline deep reinforcement learning. Eng. Appl. Artif. Intell. **104**, 104366 (2021)
33. Tessler, C., Efroni, Y., Mannor, S.: Action robust reinforcement learning and applications in continuous control. In: International Conference on Machine Learning, pp. 6215–6224 (2019)
34. Uehara, M., Sun, W.: Pessimistic model-based offline reinforcement learning under partial coverage. arXiv preprint arXiv:2107.06226 (2021)
35. Wang, J., Li, W., Jiang, H., Zhu, G., Li, S., Zhang, C.: Offline reinforcement learning with reverse model-based imagination. Adv. Neural Inf. Process. Syst. **34**, 29420–29432 (2021)
36. Wiesemann, W., Kuhn, D., Rustem, B.: Robust Markov decision processes. Math. Oper. Res. **38**(1), 153–183 (2013)
37. Xie, T., Jiang, N.: Batch value-function approximation with only realizability. In: International Conference on Machine Learning, pp. 11404–11413 (2021)
38. Yang, W., Zhang, L., Zhang, Z.: Toward theoretical understandings of robust Markov decision processes: sample complexity and asymptotics. Ann. Stat. **50**(6), 3223–3248 (2022)
39. Ying, L., Zhu, Y.: A note on optimization formulations of Markov decision processes. arXiv preprint arXiv:2012.09417 (2020)
40. Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., Finn, C.: COMBO: conservative offline model-based policy optimization. Adv. Neural Inf. Process. Syst. **34**, 28954–28967 (2021)
41. Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J.Y., Levine, S., Finn, C., Ma, T.: MOPO: model-based offline policy optimization. Adv. Neural Inf. Process. Syst. **33**, 14129–14142 (2020)
42. Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J.D., Chi, Y.: Policy mirror descent for regularized reinforcement learning: a generalized framework with linear convergence (2021)
43. Zhou, W., Bajracharya, S., Held, D.: Plas: Latent action space for offline reinforcement learning. In: Conference on Robot Learning, pp. 1719–1735 (2021)