Robust Offline Reinforcement Learning with Heavy-Tailed Rewards

Jin Zhu LSE

Runzhe Wan Amazon

Zhengling Qi George Washington University

Shikai Luo Bytedance

Chengchun Shi LSE

Abstract

This paper endeavors to augment the robustness of offline reinforcement learning (RL) in scenarios laden with heavy-tailed rewards, a prevalent circumstance in real-world applications. We propose two algorithmic frameworks, `ROAM` and `ROOM`, for robust off-policy evaluation and offline policy optimization (OPO), respectively. Central to our frameworks is the strategic incorporation of the median-of-means method with offline RL, enabling straightforward uncertainty estimation for the value function estimator. This not only adheres to the principle of pessimism in OPO but also adeptly manages heavy-tailed rewards. Theoretical results and extensive experiments demonstrate that our two frameworks outperform existing methods on the logged dataset exhibits heavy-tailed reward distributions. The implementation of the proposal is available at https://github.com/Mamba413/ROOM.

1 INTRODUCTION

In reinforcement learning (RL, Sutton and Barto, 2018), evaluating and optimizing policies without accessing the environment becomes crucial nowadays, because frequently interacting with the environment could be prohibitively expensive or even impractical in many real-world applications such as robotics, healthcare, education, autonomous driving, and so on. This leads to a surge of interest in offline RL (Levine et al., 2020; Uehara et al., 2022), which aims to leverage only logged data for policy evaluation and optimization.

The success of offline RL so far crucially relies on that the reward distribution is well-behaved. However, in a number of real-world applications, the reward distribution is usually heavy-tailed[1]. Heavy-tailed rewards can be generated by various real-world decision-making systems, such as the stock market, networking routing, scheduling, hydrology, image, audio, and localization errors, etc (Georgiou et al., 1999; Hamza and Krim, 2001; Huang and Zhang, 2017; Ruotsalainen et al., 2018).

The heavy-tailedness pose great challenges to existing offline RL methods. We first illustrate this via a fundamental problem in offline RL: off-policy evaluation (OPE). OPE aims to evaluate the value of policies using only logged data. One classic algorithm is fitted-Q evaluation (FQE), where each step is solving a regression problem with the response variable being the observed reward plus some estimated long-term values. Yet, it is well-known that the performance of standard regression methods is very sensitive to heavy-tailed responses (Lugosi and Mendelson, 2019) and will have a much slower convergence rate. Consequently, this issue will degrade the performance of policy evaluation.

As for offline policy optimization (OPO), the heavy-tailed rewards pose even more challenges because the issue of overestimation in standard RL algorithms could be aggravated. We elaborate this with a bandit example shown in Figure 1, a special case of RL. In this example, the large variance in estimating the expected reward causes a non-negligible probability of selecting the sub-optimal arm. In settings with heavy-tailed rewards, the empirical mean of the sub-optimal arm is subject to an even larger variance, leading to a higher probability of selecting the sub-optimal arm.

To accommodate the heavy-tailed rewards in offline RL, we propose new frameworks for both OPE and OPO by leveraging the median-of-means (MM) estimator in robust statistics (Nemirovskij and Yudin, 1983; Alon et al., 1996). Specifically, we design frameworks that can effectively robustify existing RL algorithms against the heavy-tailed rewards. The frameworks are simple and easy-to-implement. Moreover, the proposed approach also provides a natural way for qualifying

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

[1] A random variable is called heavy-tailed when its tail distribution is heavier than the exponential distribution, and sometimes even its variance is not well defined.

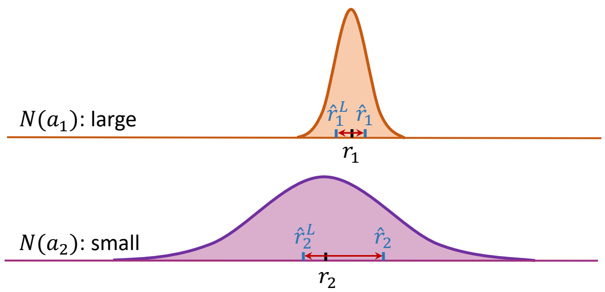

Figure 1: Reward distributions in a two-armed bandit example. The oracle expected rewards for the two arms a_k are given by r_k (for $k = 1, 2$). $N(a_k)$ denotes the number of reward observations for the k-arm. The expected rewards estimator is given by $\widehat{r}_k$. Due to the limited sample size for the second sub-optimal arm, its estimated expected reward $\widehat{r}_2$ suffers from a large variance. Consequently, there's a non-negligible probability of $\widehat{r}_2 > \widehat{r}_1$. By penalizing the uncertainty of reward estimation, a pessimistic estimation $\widehat{r}_k^L$ lowers bound the reward, leading to $\widehat{r}_2^L < \widehat{r}_1^L$, yielding the optimal action.

the uncertainty of value estimation, which is crucial in both OPE and OPO.

1.1 Contribution

The contribution of this paper is three-fold. First, we propose a general and unified framework to improve the robustness of existing OPE and OPO methods against heavy-tailed rewards. By leveraging MM, our approach naturally allows uncertainty quantification of the estimated values. This is critical for OPE in high-risk applications (e.g., healthcare) and also for OPO to incorporate the principle of pessimism (Jin et al., 2021; Bai et al., 2022) to address insufficent data coverage.

Second, we provide rigorous theoretical analyses on our OPE and OPO algorithms and clearly demonstrate their advantages over existing solutions that overlook the heavy-tailed issue. In particular, our analysis only requires the reward to have finite $(1 + \alpha)$-th moment. On the contrary, most of the existing methods require the rewards to be bounded (or sub-Gaussian) to achieve similar performance.

Finally, on a couple of benchmark OpenAI environments, we observe the superiority of the proposed algorithms against existing ones when the rewards are heavy-tailed. In particular, for OPE, our methods are 1.5 to 30 times more accurate than the non-robust algorithms in terms of rooted MSE; on several D4RL benchmarks for OPO, the score of the robust version is about 1.3 to 3 times higher than those of the vanilla version of state-of-the-art (SOTA) algorithm.

2 RELATED WORKS

Off-Policy Evaluation. In the literature, there are three commonly-used approaches for OPE. The first one is the direct method (DM), which evaluates the target policy by estimating its Q-function (Bertsekas, 2012; Farajtabar et al., 2018; Le et al., 2019; Duan et al., 2020; Luckett et al., 2020; Liao et al., 2021; Hao et al., 2021; Shi et al., 2022a). Importance sampling (IS) is another popular OPE approach (Precup, 2000; Thomas et al., 2015; Liu et al., 2018; Hanna et al., 2019; Nachum et al., 2019; Xie et al., 2019; Dai et al., 2020; Wang et al., 2021a), motivated by the change of measure theorem. Sequential IS gives an unbiased estimator but has an exponentially large variance with respect to the horizon. Liu et al. (2018); Xie et al. (2019) developed marginal IS estimators to break this curse of horizon. The last approach aims to exploit the advantages of both DM and IS, by combining them to derive a doubly robust (DR) estimator (Thomas and Brunskill, 2016; Jiang and Li, 2016; Farajtabar et al., 2018; Tang et al., 2019; Kallus and Uehara, 2020; Liao et al., 2022). We refer to Uehara et al. (2022) for a comprehensive review for OPE. However, to our knowledge, most existing methods cannot handle the heavy-tailed rewards.

Offline Policy Optimization. It is well-known that standard OPO methods (e.g., Ernst et al., 2005) may fail to converge and produce unstable solutions due to the distributional mismatch in the offline setting (Wang et al., 2021b). To address this limitation, one possible approach is to force the learned policy to choose actions close to the observed ones in the offline data (Wu et al., 2020; Brandfonbrener et al., 2021; Fujimoto and Gu, 2021; Kostrikov et al., 2021; Dadashi et al., 2021). Recently, there is a streamline of research utilizing the principle of pessimism to address the insufficient data coverage issue (e.g., Kumar et al., 2020; An et al., 2021; Jin et al., 2021; Xie et al., 2021; Yin et al., 2021; Yu et al., 2021; Bai et al., 2022; Fu et al., 2022b; Kostrikov et al., 2022; Guo et al., 2022; Shi et al., 2022b; Uehara and Sun, 2022; Lyu et al., 2022; Fu et al., 2022a; Zhou et al., 2023; Xu et al., 2023; Zhang et al., 2023; Zhou, 2023; Chen et al., 2023). We refer interested readers to Prudencio et al. (2023) for a recent survey on OPO. However, existing OPO methods cannot handle heavy-tailed rewards. In addition, recent OPO methods proposed to use confidence intervals (CIs) to quantify the uncertainty of the estimated Q-function (Jin et al., 2021; Bai et al., 2022). These CIs could be unreasonably wide due to the heavy-tailed rewards.

Robust RL. Most existing works on handling heavy-tailed rewards are only designed for bandits, a special case of RL. Various robust mean estimators are proposed for designing algorithms in finding an optimal arm in the online setting (e.g., Bubeck et al., 2013; Shao et al., 2018; Lu et al., 2019; Zhong et al., 2021). However, less attention has been paid to heavy-tailed rewards when there has a state transition. To the best of our knowledge, Zhuang and Sui (2021); Rowland et al. (2023); Liu et al. (2023) are the most related papers studying this issue. They focus on an online setting which is substantially different from our offline setting.

We remark that there has another line of research on robust offline RL (Chen et al., 2021; Lykouris et al., 2021; Mo et al., 2021; Si et al., 2020; Zhang et al., 2021; Kallus et al., 2022; Xu et al., 2022; Zhang et al., 2022), which mainly focuses on robust decision making under the uncertainty of the changing environment. Another stream studies OPE/OPO under the robust Markov decision process (Nilim and El Ghaoui, 2005) by exploiting prior distributional information allow uncertainty quantification (Mannor et al., 2016; Wiesemann et al., 2013; Wang et al., 2022; Goyal and Grand-Clement, 2023). In summary, the goals of these research are different from ours.

3 PRELIMINARIES

Markov decision process. We consider an infinite-horizon discounted stationary Markov Decision Process (MDP, Sutton and Barto, 2018), which is defined by a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, the transition kernel $\mathcal{P}(\bullet|S_{t-1}, A_{t-1})$ specifies the probability mass (or density) function of S_t by taking action A_{t-1} at a state S_{t-1}, and similarly $\mathcal{R}$ specifies the reward. The constant $\gamma \in [0, 1)$ is the discount factor. We denote the initial state distribution as $\mathbb{G}$. For simplicity of notations, we assume $\mathbb{G}$ is pre-specified in this paper. $\mathbb{G}$ can be estimated from the empirical initial state distribution in practice.

In the existing literature, the reward is assumed to be uniformly bounded or at least sub-Gaussian (Thomas and Brunskill, 2016; Fan et al., 2020; Chen and Qi, 2022; Shi et al., 2023). However, as discussed in the introduction, such an assumption could be violated in many real applications. In this paper, we consider a much milder assumption, that is, the reward distribution $\mathcal{R}$ has finite $(1 + \alpha)$-th moments for some $\alpha \in (0, 1]$. Then, the mean reward function $r(s, a) = \mathbb{E}(R_t|S_t = s, A_t = a)$ exists. No other assumptions are imposed and the reward distribution can be arbitrarily heavy-tailed. Let $\pi(a|s) : \mathcal{S} \to \mathcal{A}$ be a given policy that specifies the conditional distribution of the action given the state. We next the value function $V^\pi(s) := \mathbb{E}^\pi [\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s]$ and the Q-function $Q^\pi(s, a) := \mathbb{E}^\pi [\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s, A_0 = a]$. Let $\mathbb{E}_{\mathcal{D}}[\cdot]$ denote the expectation taken with respect to the empirical measure over the offline data $\mathcal{D}$.

Problem Formulation. We assume that an agent interacts with the environment $\mathcal{M}$ and collects a series of random tuples in the form of (S, A, R, S') using one behavior policy. The offline dataset $\mathcal{D}$ consists of all tuples with form (S, A, R, S'). There are two main tasks in offline RL as follows.

- Off-policy evaluation (OPE): given the offline dataset $\mathcal{D}$ and a given target policy π, OPE estimates its value $J^\pi := \mathbb{E}_{s\sim\mathbb{G}, a\sim\pi(\bullet|s)} Q^\pi(s, a)$.
- Offline policy optimization (OPO): given the offline dataset $\mathcal{D}$, OPO aims to learn an optimal policy $\pi^* = \arg\max_\pi J^\pi$.

Most existing methods for the two tasks crucially rely on the assumption that the rewards are uniformly bounded, yet simply employing them cannot address the challenges posed by heavy-tailed rewards, as illustrated in the example below.

Failure of standard direct methods. To illustrate, we mainly focus on DM for OPE, which has shown promising performances from theory and empirical studies (Duan et al., 2020; Voloshin et al., 2021). A DM-type OPE algorithm first estimates the Q-function as $\widehat{Q}$ and then estimate the value of η^π by constructing a plug-in estimator for $\widehat{J}^\pi = \mathbb{E}_{s\sim\mathbb{G}, a\sim\pi(\bullet|s)} \widehat{Q}^\pi(s, a)$.

To see the failure of DM, we first present the connection between Q-function estimation and population mean estimation. Define the conditional discounted state-action visitation distribution of the tuple (s, a) following policy π starting from (s_0, a_0) as $d^\pi(a, s|a_0, s_0) = (1 - \gamma)\{\mathbb{I}(a = a_0, s = s_0) + \sum_{t=1}^{\infty} \gamma^t p_t^\pi(a, s \mid a_0, s_0)\}$. Then,

$$
\begin{aligned}
&Q^\pi(s_0, a_0) \\
=&(1 - \gamma)^{-1} \mathbb{E}_{(S_t, A_t)\sim d^\pi(a,s|a_0,s_0), R_t\sim\mathcal{R}(S_t,A_t)} R_t.
\end{aligned}
$$

In other words, the Q-value $Q^\pi(s_0, a_0)$ is the population mean of the stochastic rewards under the corresponding state-action visitation distribution induced by policy π starting from (s_0, a_0). The heavy tailedness[2] of R_t typically will carry over to the distribution

[2]The heavy-tailedness can be caused by either the heavy-tailedness of $\mathcal{R}$ (i.e., $R_t - r(S_t, A_t)$, the randomness of the stochastic rewards) or that of $r(S_t, A_t)$ (i.e., the distribution of the mean reward following some policy). Almost all of our discussions can accommodate both sources simultaneously.

of $\sum_{t=0}^{\infty}\gamma^t R_t$ conditioned on $\{S_0 = s, A_0 = a\}$ and following π. In this sense, the estimation of $Q^\pi(s_0, a_0)$ will face the same challenge as the population mean estimation with heavy-tailed noises[3]. At this case, the estimation error of the sample mean $\bar{R}$ can be upper bounded by $|\bar{R} - \mu| \le C'\delta^{-\frac{1}{1+\alpha}} n^{-\frac{\alpha}{1+\alpha}}$ with probability at least $1-\delta$, for a constant $C' > 0$. To ensure a high-probability result, δ shall inversely scale polynomially in n, causing the error bound may scale polynomially in n, which dominate the (constant) reward means.

Median-of-mean method. The key tool in our algorithms is the MM estimator (Nemirovskij and Yudin, 1983; Alon et al., 1996) in robust statistics. Due to its flexibility and that it is straightforward to produce uncertainty quantification, MM has also been employed in robust linear regressions as well (Zhang and Liu, 2021; Minsker, 2015). We present its form in population mean estimation and related property below.

Definition 1 (Population mean estimation via MM)**.** *Let $R_1, \ldots, R_n$ be n i.i.d. real-valued heavy-tailed observations under a distribution F. To estimate the population mean, we first partition $[n] = \{1, \ldots, n\}$ into $K \in \mathbb{N}^+$ blocks $B_1, \ldots, B_K$, each of size $|B_i| \ge \lfloor n/K \rfloor \ge 2$. We compute the sample mean in each block as $Z_k = \frac{1}{|B_k|}\sum_{i \in B_k} R_i$. The MM estimator for the mean value of F is defined as* $\mathrm{Median}(\{Z_1, \ldots, Z_K\})$.

Proposition 1 (Lugosi and Mendelson (2019), Theorem 3)**.** *Suppose $R_1, \ldots, R_n$ are i.i.d. with mean μ and the $(1+\alpha)$ th moment. For any $\delta \in (0,1)$, by setting $K = \lceil 8\log(2/\delta)\rceil$, we have with probability at least $1-\delta$ that*

$$|\widehat{\mu}_n - \mu| \le C[\log(1/\delta)]^{\frac{\alpha}{1+\alpha}} n^{-\frac{\alpha}{1+\alpha}}$$

for some constant $C > 0$.

Comparing sample mean and the MM, we easily see that sample mean's dependence on the confidence parameter δ is exponentially worse than that of MM. Indeed, a sub-Gaussian assumption is typically required for sample mean to enjoy the same property as MM estimator. Based on the aforementioned discussion, we will borrow ideas from the MM to improve the robustness of OPE and OPO.

4 MM FOR ROBUST OFFLINE RL

In this section, we start by presenting our proposal for OPE to illustrate the main idea of utilizing MM in offline RL to address the heavy-tailed rewards. We then extend the idea to OPO in Section 4.2.

[3] One may also refer to Theorem 4 in Gerstenberg et al. (2022) for a sufficient condition for the cumulative reward to be heavy-tailed.

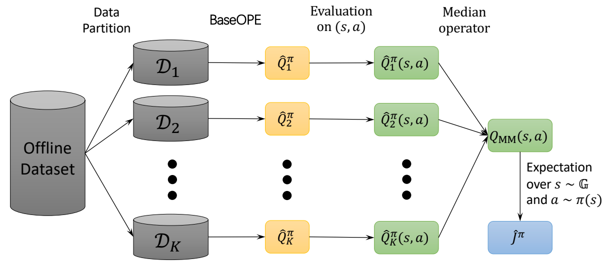

Figure 2: Graphical illustration for `ROAM`. $Q_{\mathrm{MM}}(s,a)$ is equal to $\mathrm{Median}(\{\widehat{Q}_k^\pi(s,a)\}_{k=1}^K)$.

4.1 MM for OPE

This section introduces the Robust Off-policy evaluAtion via Median-of-means (`ROAM`) framework. For ease of exposition, we first focus on the DM in this paper. Specifically, the discussions above motivate us to consider leveraging the MM scheme for robust estimation of Q^π (see an illustration in Figure 2). We first split all trajectories $\mathcal{D}$ into K partitions $\{\mathcal{D}_k\}_{k=1}^K$. Notice that data subsets across the K splits are *i.i.d.* Next, with any given DM-type OPE algorithm `BaseOPE`, we obtain K *i.i.d.* estimates $\{\widehat{Q}_k^\pi\}_{k=1}^K$ for Q^π. However, with heavy-tailed rewards, these estimates may also have large errors and distributed with heavy tails around Q^π. As discussed above, this is similar to the sample average in every split for population mean estimation. Therefore, we propose to extend MM to OPE by first applying the median operator to the K Q-functions and then calculate the integrated value estimate as $\widehat{J}^\pi = \mathbb{E}_{s\sim\mathbb{G}, a\sim\pi(\bullet|s)}\mathrm{Median}(\{\widehat{Q}_k^\pi(s,a)\}_{k=1}^K)$. We summarize the procedure in Algorithm 1. Notably, our approach employs a split-and-aggregation strategy to estimate a robust Q function, which is markedly different from the standard MM method for estimating a scalar. As such, verifying the robustness of the estimated Q function necessitates a non-trivial analysis of the proposed procedure.

Algorithm 1 Robust Off-policy Evaluation via Median-of means based Direct Method (`ROAM-DM`)

input Policy π, data $\mathcal{D}$, data partitions number K, decay rate γ, base DM-type OPE algorithm `BaseOPE`
1: Partition trajectories $\mathcal{D}$ into K disjoint parts: $\mathcal{D}_1, \ldots, \mathcal{D}_K$.
2: **for** $k = 1, \ldots, K$ **do**
3: $\quad \widehat{Q}_k^\pi \leftarrow$ `BaseOPE` $(\pi, \mathcal{D}_k, \gamma)$
4: **end for**
5: $\widehat{J}^\pi \leftarrow \mathbb{E}_{s\sim\mathbb{G}, a\sim\pi(\bullet|s)}\mathrm{Median}(\{\widehat{Q}_k^\pi(s,a)\}_{k=1}^K)$.
output $\widehat{J}^\pi$.

Uncertainty quantification. In many high-risk applications such as mobile health studies, in addition to a point estimate on a target policy's value, it is crucial to quantify the uncertainty of the value estimates, which has attracted increasing attention in recent years (Dai et al., 2020; Shi et al., 2021; Liao et al., 2021, 2022; Kallus et al., 2022). One prominent advantage of leveraging MM in OPE is that it is straightforward to produce uncertainty quantification. Specifically, with $\{\widehat{Q}_k^{\pi}\}_{k=1}^{K}$, we can have K integrated value estimators as $\{\widehat{J}_k^{\pi}\}_{k=1}^{K}$. Notice that $\{\widehat{J}_k^{\pi}\}_{k=1}^{K}$ are *i.i.d.*. When each $\widehat{J}_k^{\pi}$ is unbiased, $\mathcal{Q}_q(\{\widehat{J}_k^{\pi}\}_{k=1}^{K})$ is a natural $1-q$ lower confidence bound for η^{π}, where $\mathcal{Q}_q(\cdot)$ returns the q-th lower quantile value among a set. In contrast, it is nontrivial to obtain uncertainty quantification with other robust estimators like the truncated mean.

Variants. Our proposal is general and has a few theoretical guaranteed variants. First, instead of applying the median operator to the Q-values in Step 5 in Algorithm 1, we can apply the median operator to the estimated integral value $\widehat{J}_k^{\pi} = \mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\bullet|s)}\widehat{Q}_k^{\pi}(s,a)$ to obtain $\widehat{J}^{\pi} = \text{Median}(\{\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\bullet|s)}\widehat{Q}_k^{\pi}(s,a)\}_{k=1}^{K})$. We study this variant, called `ROAM-Variant`, by empirical studies.

Next, we can extend the framework of MM to give a robust IS estimator. To illustrate this extension, we take the marginal important sampling (MIS) estimator (Liu et al., 2018; Xie et al., 2019) as our example. The MIS estimator first estimates the state-action density ratio $\omega^{\pi}(s,a) := d^{\pi}(s)\pi(a|s)/b(s,a)$ as $\widehat{\omega}^{\pi}(s,a)$. Here, $b(s,a)$ is the state-action density of behavior policy, $d^{\pi}(s)$ is the average visitation distribution, defined as $d^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_t^{\pi}(s)$ where $d_t^{\pi}(s)$ is the distribution of state s_t when we execute policy π. Then, the MIS estimates the value of π as $\widehat{J}^{\pi} := \mathbb{E}_{\mathcal{D}}\left[\widehat{\omega}^{\pi}(S,A)R\right]$. To apply the MM procedure, we partition $\mathcal{D}$ into K disjoint parts $\mathcal{D}_1,\ldots,\mathcal{D}_K$; then for each $\mathcal{D}_k$, we estimate ratios $\widehat{\omega}_k^{\pi}(s,a)$ and compute $\widehat{J}_k^{\pi} := \mathbb{E}_{\mathcal{D}_k}\left[\widehat{\omega}_k^{\pi}(S,A)R\right]$. Finally, we define the robust IS estimator as $\widehat{J}^{\pi} = \text{Median}(\{\widehat{J}_k^{\pi}\}_{k=1}^{K})$. We summarize our method in Algorithm 3 in Appendix A2.1, which we refer to as `ROAM-MIS`. Our method can be similarly extended to accommodate doubly robust methods (Thomas and Brunskill, 2016; Kallus and Uehara, 2020).

4.2 MM for OPO with Pessimism

In this section, we introduce the extension of our proposal to OPO, called Robust OPO via Median-of-means (`ROOM`). To illustrate, we focus on value-based OPO algorithms in this section. A value-based OPO algorithm typically first estimates the optimal Q-function as $\widehat{Q}^*$, and then derives the corresponding optimal policy as either $\widehat{\pi}^*(s) = \arg\max_a \widehat{Q}^*(s,a)$ or $\widehat{\pi}^*(s) = \arg\max_{\pi\in\Pi}\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\bullet|s)}\widehat{Q}^*(s,a)$ (when a policy class Π is prespecified). Popular methods include fitted Q-iteration (FQI, Ernst et al., 2005), LSTD Q-learning (LSTD-Q, Lagoudakis and Parr, 2003), etc.

To design a robust value-based OPO algorithm, we can follow a similar procedure for OPE as in Section 4.1. Specifically, we can first split $\mathcal{D}$ into K folds to estimate K independent optimal Q-functions $\{\widehat{Q}_k^*\}_{k=1}^{K}$, then output a policy $\widehat{\pi}^*(s) = \arg\max_a \text{Median}(\{\widehat{Q}_k^*(s,a)\})$. We present this algorithm in Algorithm 2, which we call `ROOM` for Value-based Method (`ROOM-VM`).

Algorithm 2 Robust OPO via Median-of-means for Value-based Method (`ROOM-VM`)

input Data $\mathcal{D}$, decay rate γ, number of data partitions K, base value-based OPO algorithm `BaseOPO`
1: Partition data $\mathcal{D}$ into K parts: $\mathcal{D}_1,\ldots,\mathcal{D}_K$.
2: **for** $k=1,\ldots,K$ **do**
3: $\quad\widehat{Q}_k^* \leftarrow$ `BaseOPO` $(\mathcal{D}_k,\gamma)$
4: **end for**
5: $\widehat{\pi}^*(s) \leftarrow \arg\max_a \text{Median}(\{\widehat{Q}_k^*(s,a)\}_{k=1}^{K})$ for any s.
output Policy $\widehat{\pi}^*$.

Pessimism for robust OPO. Insufficient data coverage is known as a critical issue to offline RL (Levine et al., 2020; Xu et al., 2023). When some state-action pairs are less explored, the related value estimation tends to have high variance and hence classical RL algorithms may produce a sub-optimal policy. The pessimistic principle effectively mitigates this issue, by taking the uncertainty of the value function estimation into consideration; see Figure 1 for an illustration.

Employing the pessimism principle relies crucially on the construction of the uncertain set for the value function. However, as pointed out by Zhou et al. (2023), it is often challenging to derive a credible lower bound in general, and the tuning is typically sensitive. The issue becomes more serious when there exist heavy-tailed rewards. One prominent advantage of our MM procedure is that the pessimism mechanism can be naturally and efficiently incorporated. Specifically, we can replace the Median$(\cdot)$ operator in the Step 5 of Algorithm 2 by:

$$\widehat{\pi}^*(s) \leftarrow \arg\max_a \mathcal{Q}_q(\{\widehat{Q}_k^*(s,a)\}_{k=1}^{K}) \text{ for any } s.$$

By choosing $q<0.5$, we naturally obtain a pessimistic estimation for addressing insufficient data coverage. Moreover, quantile optimization is known to be robust (Wang et al., 2018) against heavy-tailed rewards as well (notice that the median operator is just a special case of quantile operators). Therefore, by this design, one can expect our method can address both insufficient data coverage and heavy-tailed rewards.

We conclude this section with the comparison the standard methods. Standard methods construct the lower confidence bound of $Q^*(s,a)$ rely on subtracting the sample standard deviation (multiplied by a factor) from the sample mean, both are obtained from bootstrapping or concentration inequalities (see, e.g., Kumar et al. (2019, 2020); Bai et al. (2022)). However, as Hall (1990) points out, non-parametric bootstrap of the sample mean for heavy-tailed variables may not lead to a Gaussian asymptotic distribution. Therefore, pessimistic RL algorithms based on bootstrapping may not work in heavy-tailed environments. Similar challenges apply to concentration inequality-based methods, especially when the variance does not exist.

5 THEORY

In this section, we focus on deriving the statistical properties of `ROAM`, designed for OPE. Meanwhile, our analysis can be easily extended to obtain the upper error bound of the estimated Q-function via `ROOM-VM` for OPO. We begin with a set of technical assumptions.

Assumption 1 (Independent transitions). $\mathcal{D}$ contains n i.i.d. copies of (S, A, R, S').

Assumption 2 (Heavy-tailed rewards) There exists some $\alpha \in (0,1]$ such that $\mathbb{E}[|R|^{1+\alpha}] < \infty$.

Assumption 3 (Sequential overlap). ω^π is bounded away from 0.

We make a few remarks. First, the independence condition in Assumption 1 is commonly imposed in the literature to simplify the theoretical analysis (see e.g., Sutton et al., 2008; Chen and Jiang, 2019; Fan et al., 2020; Uehara et al., 2020). It can be relaxed by imposing certain mixing conditions on the underlying MDP (Kallus and Uehara, 2022; Chen and Qi, 2022; Bhandari et al., 2021). Second, as we have commented earlier, nearly most existing works require the rewards to be uniformly bounded. On the contrary, Assumption 2 requires a very mild moment condition. When $\alpha < 1$, this assumption even allows the variance of the reward to be infinity. Therefore, it is commonly used in the robust learning for bandits/RL literature (Bubeck et al., 2013; Zhuang and Sui, 2021). Assumption 3 corresponds to the sequential overlap condition that is widely imposed in the OPE literature (see e.g., Kallus and Uehara, 2020; Shi et al., 2021). It essentially requires the support of the stationary state-action distribution under the behavior policy to cover that of the discounted state-action visitation distribution under π.

We first study the theoretical properties of `ROAM-MIS`.

Theorem 1. *Assume Assumptions 1-3 hold. Then for any $\delta > 0$, by setting $K = \lceil 8\log(2/\delta)\rceil$, we have with probability $1-\delta$ that $|\widehat{J}^\pi_{MIS} - J^\pi|$ is of the order of magnitude*

$$M_R^{(1+\alpha)}\Big[\|\widehat{\omega}^\pi - \omega^\pi\|_\infty + \|\omega^\pi\|_\infty [\log(1/\delta)]^{\frac{\alpha}{1+\alpha}} n^{-\frac{\alpha}{1+\alpha}}\Big],$$

where $M_R^{(1+\alpha)} = (\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}$ and ℓ_∞-norm of any function ω is defined as $\|\omega\|_\infty \coloneqq \sup_x |\omega(x)|$.

According to Theorem 1, the estimation error of the proposed `ROAM-MIS` estimator can be decomposed into the sum of two terms. The first term depends crucially on the estimation error of the MIS ratio. By definition, the MIS ratio is independent of the reward distribution. As such, existing solutions are applicable to compute $\widehat{\omega}^\pi$ to obtain a tight estimation error bound. The second term depends on δ only through $(\log(1/\delta))^{\frac{\alpha}{1+\alpha}}$, which demonstrates the advantage of MM. It also relies on $\|\omega^\pi\|_\infty$, which measures the degree of distribution shift due to the discrepancy between the behavior and target policies. Finally, notice that both terms are proportional to $(\mathbb{E}[|R|^{1+\alpha}])^{\frac{1}{1+\alpha}}$. Compared with Proposition 1 for the classical MM estimation, Theorem 1 shows `ROAM-MIS` presents a much greater challenge. Although this may seem intuitive, our quantitative analysis reveals two key insights: (i) the error of the estimated ratio has an additive effect on the OPE error, and (ii) the moment of reward introduces an additional scaling effect on the OPE error.

We next study `ROAM-DM`. For illustration purposes, we focus on a particular `BaseOPE` algorithm, the LSTD algorithm (Bradtke and Barto, 1996; Boyan, 1999). In particular, let $\phi(S,A)$ denote a set of uniformly bounded basis functions, we parameterize the Q-function $Q^\pi(s,a) \approx \phi^\top(s,a)\theta^*$ for some θ^* and estimate this parameter by solving the following estimating equation with respect to θ,

$$\mathbb{E}_{\mathcal{D}}\Big\{\sum_a \phi(S,A)[R + \phi^\top(S',a)\theta - \phi^\top(s,a)\theta]\Big\} = 0.$$

Denote the resulting estimator as $\widehat{\theta}$. Let $\xi_\pi(S,A) = \sum_a \mathbb{E}[\pi(a|S')\phi(S',a)|S,A]$. We impose some additional assumptions.

Assumption 4 (Realizability) There exists some θ^* such that $Q^\pi(s,a) = \phi^\top(s,a)\theta^*$ for any s and a.

Assumption 5 (Invertibility) The minimum eigenvalue of $[\mathbb{E}\phi(S,A)\phi^\top(s,a) - \gamma\mathbb{E}\xi_\pi(S,A)\xi_\pi^\top(S,A)]$, denoted by $\lambda_{\min}$, is strictly positive.

We again, make some remarks. First, Assumption 4 is a widely-used condition in the OPO literature to simplify the theoretical analysis (Duan et al., 2020; Jiang and Huang, 2020; Min et al., 2021; Zhan et al., 2022). It can be relaxed by allowing the Q-function to be misspecified, i.e., $\inf_\theta \mathbb{E}|Q^\pi(s,a) - \phi^\top(s,a)\theta|^2 > 0$ (see, e.g., Chen and Jiang, 2019). Second, Assumption 5 is commonly imposed in the literature (Luckett et al., 2020; Perdomo et al., 2022; Shi et al., 2022a). It can be viewed as a version of the Bellman completeness assumption when specialized to linear models (Munos and Szepesvári, 2008). Moreover, according to Theorem 3 in Perdomo et al. (2022), invertibility is a necessary condition for solving any OPE problem using a broad class of linear estimators such as LSTD.

Theorem 2. *Assume Assumptions 1, 2, 4 and 5 hold. Then for any* $\delta > 0$ *such that* $\lambda_{\min} \gg (n/\log\delta^{-1})^{-1/2}$*, by setting* $K = \lceil 8\log(2/\delta)\rceil$*, we have with probability* $1-\delta$ *that* $|\widehat{J}^{\pi}_{DM} - J^{\pi}|$ *is of the order of magnitude*

$$\lambda_{\min}^{-1}(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\left(\log(1/\delta)\right)^{\frac{\alpha}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}.$$

Using similar arguments in the proof of Theorem 2 (see Appendix A1.2), we can show that the estimation error of the standard LSTD estimator grows at a polynomial order of δ^{-1}. This again demonstrates the advantage of our proposal. Moreover, Theorem 2 shows that, $\lambda_{\min}^{-1}$ serves a unique factor when compared to the classical MM estimator. Like Theorem 1, the $(1+\alpha)$-moment of the reward have a scaling effect on the OPE error. And thus, Theorem 2 highlights the challenges of robustifying the Q function and quantitatively illustrates the crucial terms for controlling the error of MM-based LSTD. To the best of our knowledge, this has largely unexplored in literature. Lastly, from the proof of Theorem 2, we can prove `ROAM-Variant` also possesses the same order of error bound.

Finally, the subsequent theorem shows that `ROOM` derives a "robusified" pointwise lower bound of $Q^*(s,a)$. For illustration purposes, we concentrate on a specific `BaseOPO` method, the LSTD-Q algorithm.

Theorem 3. *If Assumptions 1, 2, 4 hold and Assumption 5 holds for* π^**, then for any* $(s,a) \in \mathcal{S}\times\mathcal{A}$*, the following event*

$$Q^*(s,a) - \lambda_{\min}^{-1}(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \geq \mathcal{Q}_q(\{\widehat{Q}_k(s,a)\}_{k=1}^K)$$

hold with probability at least $1-\exp(-2K(2q-1)^2)$.

Notably, the gap between $Q^*(s,a)$ and pessimistic Q function estimation is proportional to $\mathbb{E}|R|^{1+\alpha}$, which exists even when rewards are heavy-tailed. Therefore, compared to pessimistic methods based on subtracting standard deviation, our proposal provides a robust lower bound and addresses heavy-tailedness and data coverage simultaneously.

6 EXPERIMENTS

Experiments for OPE. We first describe the procedure to generate the offline dataset $\mathcal{D}$ with heavy-tailed rewards. We first train a policy online under $\mathcal{M}$ using PPO (Schulman et al., 2017) and denote it as π^*, the optimal policy. Let the behaviour policy be a ϵ-greedy policy based on π^*, i.e., $\pi^b := (1-\epsilon)\pi^* + \epsilon\pi^r$ where π^r is a random policy. We set $\epsilon = 0.05$ in our experiments. We use π^b to interact with the environment to collect an offline dataset $\mathcal{D}'$ with 100 episodes. In $\mathcal{D}'$, we add *i.i.d.* zero-mean heavy-tailed random variables ν_{df} to the rewards and obtain the dataset $\mathcal{D}$. We set ν_{df} as a scaled t_{df} random variable, i.e., $\nu_{\mathrm{df}} := \kappa\nu'_{\mathrm{df}}/\sigma^2$, in which ν'_{df} comes from a t_{df} distribution and σ is the standard deviation of α-truncated t_{df} random variable where $\alpha = 0.02$. The degree of freedom (df) controls the degree of heavy-tailedness, and κ controls the impact of heavy-tailed noises.

With the offline data $\mathcal{D}$, we investigate the performance of `ROAM-DM`, `ROAM-MIS`, and `ROAM-Variant` on estimating the value of π^* by comparing with the vanilla FQE algorithm on a classical OpenAI Gym tasks, Cartpole. To ablate the effect of computing K functions, we also compare with mean-aggregation-based methods, named `MA-DM` and `MA-MIS`. Moreover, considering the truncated mean (TM) as a useful technique in robust statistics, we also take it into consideration by implementing TM upon MIS, denoted by `TM-MIS`. Finally, since the FQE algorithm iteratively performs temporal difference (TD) updates, it is natural to leverage the structure of the MDP by applying MM to the TD updates. We formulate this idea for the FQE in Algorithm 4 in Appendix A2.2 and denote this algorithm as `ROAM-FQE`. For all methods, we use a linear model $\phi^\top(s,a)\theta$ to model $Q^{\pi^*}(s,a)$, where $\phi(s,a)$ includes the main effects and two-order interactions of the feature vector $(s^\top, a^\top)^\top$, generated by the `PolynomialFeatures` method of `scikit-learn` (Pedregosa et al., 2011).

Given the ground truth J^{π^*} computed via Monte Carlo, the mean squared errors of all methods are reported in Figure 3a, aggregated over 100 replicates in each case. From Figure 3a, all methods' MSEs reasonably decrease as the degree of freedom increases. The vanilla FQE method is outperformed by our methods, due to that it is not robust to heavy-tailed noises. The differences between `FQE` and our methods diminish when df increases, which is reasonable. Although `TM-MIS` does outperform the vanilla MIS, it is generally surpassed by `ROAM-MIS`. Additionally, `ROAM-MIS`, `ROAM-DM` and `ROAM-Variant` have a tiny difference, while `ROAM-FQE` surpasses all of them. This implies using MM at each iteration of FQE largely mitigates the negative impact of heavy-tailed rewards such that the whole dataset can be fully utilized during iterations. Finally, the comparison between `ROAM-DM` (or `ROAM-MIS`) and `MA-DM` (or `MA-MIS`) reveals the mean-ensemble strategy cannot handle heavy-tailed rewards.

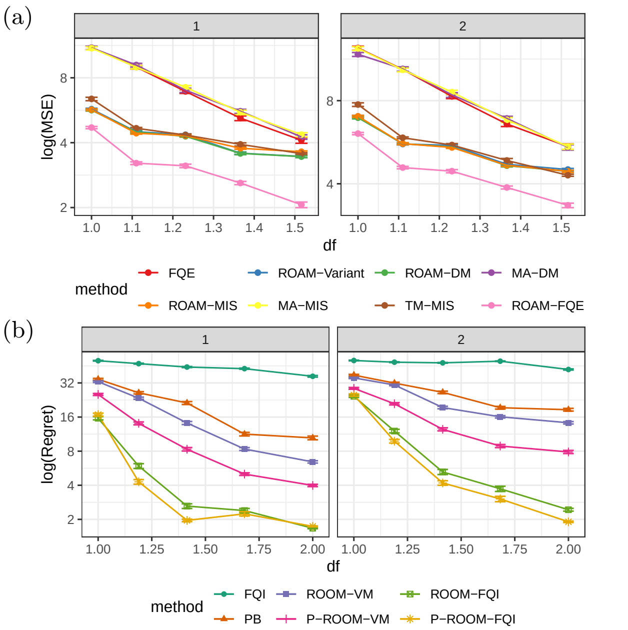

Figure 3: (a) OPE task: the trend of log(MSE) with the degree of freedom (DF). (b) OPO task: The trend of regret with respect to the DF. κ takes value 1.0 (Left panel) and 2.0 (Right panel) in each subfigure. The error bar corresponds to 95% CI.

Compared with bootstrap-based `ROAM`. Since our proposal can be interpreted as the ensemble of Q functions with a Median operator, another heuristic variant is using bootstrap instead of data partition. We conducted a comparison between this variant (denoted with the suffix "`B-`") and `ROAM`-type methods, adopting the same settings as in the previous section. The results are visualized in Figure 4. It is evident that the vertical gap between `ROAM-DM` and `B-ROAM-DM` is negligible. Furthermore, they perform significantly better than the corresponding `MA`-type methods. This phenomenon also holds for MIS-type methods presented in the right panel. Therefore, we can conclude that bootstrap serves as an alternative implementation for the proposed procedure.

Experiments for OPO (Cartpole environment). We study our algorithms: (i) `ROOM-VM` and (ii) `ROOM-VM` with pessimism (denoted as `P-ROOM-VM`), where `BaseOPO` are set as `FQI`. We compare two benchmark algorithms: `FQI` and the pessimistic-bootstrapping (`PB`) methods (Bai et al., 2022). See Appendix A3.2.1 for detailed implementations. Motivated by the powerful performance of `ROAM-FQE`, we also consider employing MM (and its pessimism version) in the TD update of `FQI`. We name these algorithms `ROOM-FQI` and `P-ROOM-FQI`, and defer their definitions to Algorithm 5 in the Appendix. We generate 400 episodes to form an offline dataset following the same

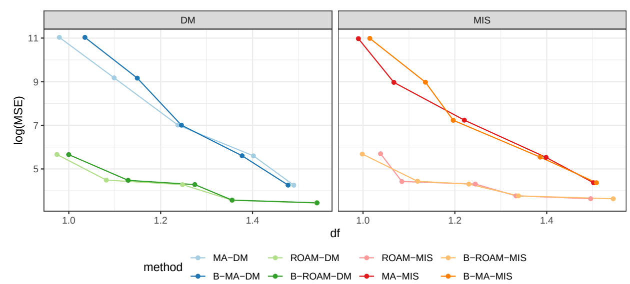

Figure 4: The left panel presents the results for DM methods, and the right one displays the results for MIS methods. To prevent point overlap, random noise has been added to each point on the x-axis.

procedure described in experiments for OPE. To evaluate the performance of a learned policy $\widehat{\pi}^*$, we compute its regret compared with the optimal policy.

We report the numerical results of 100 replications in Figure 3b. We see that the regret of `FQI` reasonably decreases when df goes up, but it decreases more slowly when κ enlarges. We can also observe that `PB` improves over `FQI`, because it can properly address the insufficient data coverage issue in OPO. However, due to the existence of heavy-tailed rewards, `ROOM-VM` and `ROOM-FQI` can outperform `PB`. Even though we have no theoretical guarantee for `ROOM-FQI`, it shows a better numerical performance compared with `ROOM-VM`. Finally, we turn to `P-ROOM-VM` and `P-ROOM-FQI` in Figure 3b. As expected, `P-ROOM-VM` (or `P-ROOM-FQI`) performs better than `ROOM-VM` (or `ROOM-FQI`) because it addresses the insufficient data coverage issue and the heavy-tailedness simultaneously.

Experiments for OPO (D4RL datasets). We evaluate our proposed approach on the MuJoCo and Kitchen environments in the D4RL benchmarks (Fu et al., 2020), which cover diverse dataset settings and domains. We generate heavy-tailed datasets by adding *i.i.d.* noises into the reward observations, similar as the previous part. To show that the generality of our framework, in these datasets, we use another SOTA algorithm, sparsity Q-learning (`SQL`, Xu et al. (2023)), as our `BaseOPO` algorithm. For the sake of fairness, we also take into account the mean aggregation (denoted as `MA`), which replaces the `Median` operator in Step 5 of Algorithm 2 with the `Mean` operator. Setting the discount factor $\gamma = 0.99$, we train each algorithm for one million time steps and evaluate it every 5000 time steps. Each evaluation consists of 10 episodes.

We report the performance in Figure 5 and show learning curves in Figures A1 and A2. It is worth noting that, in all cases, our methods are superior to the vanilla `SQL` algorithm. The superiority can be highly signifi-

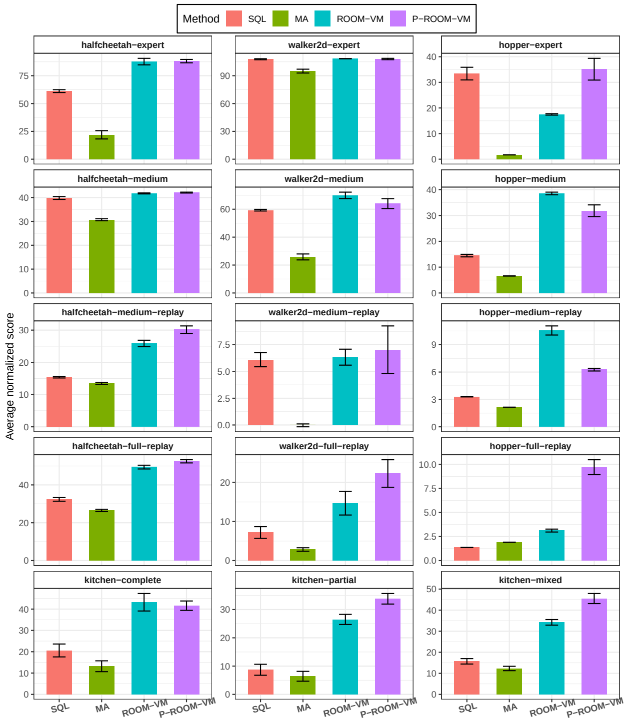

Figure 5: Results on D4RL datasets. Each bar corresponds to the average normalized score that is taken over the final 10 evaluations and 5 seeds. The error bar captures the 2 times standard error over 5 seeds.

cant. For example, on the halfcheetah-expert dataset, `ROOM-VM` and `P-ROOM-VM` achieve a 30% improvement over SQL, and in the kitchen environment, our proposal shows a 200% improvement in `SQL`'s returns. This again shows that our proposal can effectively address the challenge of heavy-tailed rewards even in complex environments. The superiority of our proposal over `MA` reveals that the mean ensemble cannot handle heavy-tailed rewards but harm numerical performance because it has to use fewer samples to learn Q functions. Notably, in almost all cases, `P-ROOM-VM` surpasses `ROOM-VM` because `P-ROOM-VM` can more effectively address the issue of severe data insufficiency coverage. Furthermore, we also study the cases where `BaseOPO` is `IQL` (Kostrikov et al., 2022). The results reported in Appendix A4.2 again show that our proposal has better performance than vanilla `IQL`, reflecting the versatility of our proposal. Finally, motivated by the success of the bootstrap-based variant and `P-ROOM-FQI`, we can further extend our proposal to the actor-critic paradigm and train agent with the whole dataset. Additionally, by setting $q = 0.0$, this heuristic implementation leads to the exact `SAC-N` proposed by An et al. (2021), which is shown to be robust on heavy-tailed rewards in Figure A3 in the Appendix A4.3.

Trade-off: Computation and robustness. We close this section with an examination of the trade-off between computation and robustness as K varies. The results in Figure 6 present the performance of `ROAM-FQE` in the Cartpole environment. As anticipated, the runtime of our proposal scales linearly with K. Yet, the computation is not demanding and can terminate in less than one second on a personal laptop. In terms of statistical performance, we observed that a moderate K — which well balances accuracy within each fold and accuracy of MM operator — achieves the highest accuracy while requiring no more than half a second to terminate.

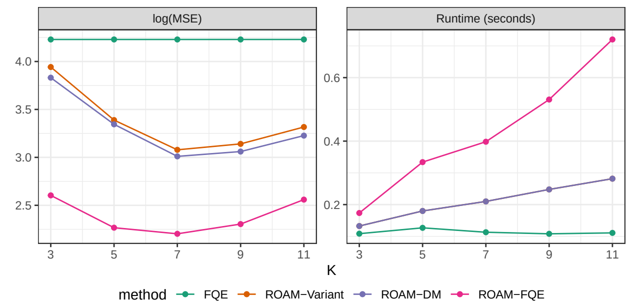

Figure 6: The log(MSE) (left) and runtime (right) of Algorithm 1 as K increases.

7 CONCLUSIONS AND FUTURE WORKS

Motivated by the real needs for robust offline RL methods against heavy-tailed rewards, we leverage the MM method in robust statistics to design a new frameworks that can robustify existing OPE and OPO algorithms. Our key insight is that employing MM to offline RL does more than just tackle heavy-tailed rewards–it offers valid uncertainty quantification to easily address insufficient coverage issue in offline RL as well. This insight is highly novel and, to our knowledge, has not been previously introduced in literature. Theoretical analysis demonstrates the advantages of our methods and extensive numerical studies support the empirical performance of our methods. An interesting future research direction is exploring the idea of leveraging the nature of heavy-tailed rewards and extending algorithms in Dubey and Pentland (2019) to the RL setting.

Acknowledgements

Jin Zhu and Runzhe Wan contribute equally to this article. Zhu's and Shi's research were supported by the EPSRC grant EP/W014971/1. The authors express their gratitude for the insightful feedback provided by the referees and the area chair, which significantly enhanced the initial version of this paper.

References

Alon, N., Matias, Y., and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29.

An, G., Moon, S., Kim, J.-H., and Song, H. O. (2021). Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447.

Bai, C., Wang, L., Yang, Z., Deng, Z.-H., Garg, A., Liu, P., and Wang, Z. (2022). Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*.

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific.

Bhandari, J., Russo, D., and Singal, R. (2021). A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973.

Boyan, J. A. (1999). Least-squares temporal difference learning. In *ICML*, pages 49–56.

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57.

Brandfonbrener, D., Whitney, W., Ranganath, R., and Bruna, J. (2021). Offline rl without off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:4933–4946.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Chen, X. and Qi, Z. (2022). On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pages 3558–3582. PMLR.

Chen, X., Qi, Z., and Wan, R. (2023). Steel: Singularity-aware reinforcement learning. *arXiv:2301.13152*.

Chen, Y., Du, S., and Jamieson, K. (2021). Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, pages 1561–1570. PMLR.

Dadashi, R., Rezaeifar, S., Vieillard, N., Hussenot, L., Pietquin, O., and Geist, M. (2021). Offline reinforcement learning with pseudometric learning. In *International Conference on Machine Learning*, pages 2307–2318. PMLR.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvari, C., and Schuurmans, D. (2020). Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725.

Duan, Y., Jia, Z., and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR.

Dubey, A. and Pentland, A. S. (2019). Thompson sampling on symmetric alpha-stable bandits. pages 5715–5721. International Joint Conferences on Artificial Intelligence Organization.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning. *arXiv:2004.07219*.

Fu, Y., Wu, D., and Boulet, B. (2022a). A closer look at offline rl agents. *Advances in Neural Information Processing Systems*, 35:8591–8604.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. (2022b). Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv:2209.08666*.

Fujimoto, S. and Gu, S. S. (2021). A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145.

Georgiou, P. G., Tsakalides, P., and Kyriakakis, C. (1999). Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE transactions on multimedia*, 1(3):291–301.

Gerstenberg, J., Neininger, R., and Spiegel, D. (2022). On solutions of the distributional bellman equation. *arXiv:2202.00081*.

Goyal, V. and Grand-Clement, J. (2023). Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226.

Guo, K., Shao, Y., and Geng, Y. (2022). Model-based offline reinforcement learning with pessimism-modulated dynamics belief. *NeurIPS*.

Hall, P. (1990). Asymptotic properties of the bootstrap for heavy-tailed distributions. *The Annals of Probability*, pages 1342–1360.

Hamza, A. B. and Krim, H. (2001). Image denoising: A nonlinear robust statistical approach. *IEEE transactions on signal processing*, 49(12):3045–3054.

Hanna, J., Niekum, S., and Stone, P. (2019). Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. (2021). Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR.

Huang, Y. and Zhang, Y. (2017). A new process uncertainty robust student'st based kalman filter for sins/gps integration. *IEEE Access*, 5:14391–14404.

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

Kallus, N., Mao, X., Wang, K., and Zhou, Z. (2022). Doubly robust distributionally robust off-policy evaluation and learning. pages 10598–10632.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167).

Kallus, N. and Uehara, M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Kostrikov, I., Fergus, R., Tompson, J., and Nachum, O. (2021). Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR.

Kostrikov, I., Nair, A., and Levine, S. (2022). Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191.

Lagoudakis, M. G. and Parr, R. (2003). Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv:2005.01643*.

Liao, P., Klasnja, P., and Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391.

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. (2022). Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31.

Liu, W., Tu, J., Zhang, Y., and Chen, X. (2023). Online estimation and inference for robust policy evaluation in reinforcement learning. *arXiv:2310.02581*.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163. PMLR.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2021). Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR.

Lyu, J., Ma, X., Li, X., and Lu, Z. (2022). Mildly conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*.

Mannor, S., Mebel, O., and Xu, H. (2016). Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509.

Min, Y., Wang, T., Zhou, D., and Gu, Q. (2021). Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34:7598–7610.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335.

Mo, W., Qi, Z., and Liu, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116(534):659–674.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5).

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32.

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Nilim, A. and El Ghaoui, L. (2005). Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Perdomo, J. C., Krishnamurthy, A., Bartlett, P., and Kakade, S. (2022). A sharp characterization of linear estimators for offline policy evaluation. *arXiv:2203.04236*.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Prudencio, R. F., Maximo, M. R. O. A., and Colombini, E. L. (2023). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–0.

Rowland, M., Tang, Y., Lyle, C., Munos, R., Bellemare, M. G., and Dabney, W. (2023). The statistical benefits of quantile temporal-difference learning for value estimation. In *International Conference on Machine Learning*. PMLR.

Ruotsalainen, L., Kirkko-Jaakkola, M., Rantanen, J., and Mäkelä, M. (2018). Error modelling for multi-sensor measurements in infrastructure-free indoor navigation. *Sensors*, 18(2):590.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv:1707.06347*.

Shao, H., Yu, X., King, I., and Lyu, M. R. (2018). Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31.

Shi, C., Qi, Z., Wang, J., and Zhou, F. (2023). Value enhancement of reinforcement learning via efficient and robust trust region optimization. *arXiv preprint arXiv:2301.02220*.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. (2021). Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591. PMLR.

Shi, C., Zhang, S., Lu, W., and Song, R. (2022a). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):765–793.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022b). Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *International Conference on Machine Learning*.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*, pages 8884–8894. PMLR.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008). A convergent o(n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21):1609–1616.

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv:1910.07186*.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. (2015). High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR.

Uehara, M., Shi, C., and Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. *arXiv:2212.06355*.

Uehara, M. and Sun, W. (2022). Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. (2021). Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Wang, J., Gao, R., and Zha, H. (2022). Reliable off-policy evaluation for reinforcement learning. *Operations Research*.

Wang, J., Qi, Z., and Wong, R. K. (2021a). Projected state-action balancing weights for offline reinforcement learning. *arXiv:2109.04640*.

Wang, Q., Xiong, J., Han, L., Liu, H., Zhang, T., et al. (2018). Exponentially weighted imitation learning for batched historical data. *Advances in Neural Information Processing Systems*, 31.

Wang, R., Wu, Y., Salakhutdinov, R., and Kakade, S. (2021b). Instabilities of offline rl with pre-trained neural representation. In *International Conference on Machine Learning*, pages 10948–10960. PMLR.

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183.

Wu, Y., Tucker, G., and Nachum, O. (2020). Behavior regularized offline reinforcement learning.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32.

Xu, H., Jiang, L., Li, J., Yang, Z., Wang, Z., Chan, V. W. K., and Zhan, X. (2023). Offline RL with no OOD actions: In-sample learning via implicit value regularization. In *The Eleventh International Conference on Learning Representations*.

Xu, Y., Shi, C., Luo, S., Wang, L., and Song, R. (2022). Quantile off-policy evaluation via deep conditional generative learning. *arXiv:2212.14466*.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal offline reinforcement learning via double variance reduction. In *Advances in Neural Information Processing Systems*.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. (2021). Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR.

Zhang, H., Mao, Y., Wang, B., He, S., Xu, Y., and Ji, X. (2023). In-sample actor critic for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2021). Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, pages 12391–12401. PMLR.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2022). Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5757–5773. PMLR.

Zhang, Y. and Liu, P. (2021). Median-of-means approach for repeated measures data. *Communications in Statistics-Theory and Methods*, 50(17):3903–3912.

Zhong, H., Huang, J., Yang, L., and Wang, L. (2021). Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 34:15710–15720.

Zhou, W. (2023). Bi-level offline policy optimization with limited exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhou, Y., Qi, Z., Shi, C., and Li, L. (2023). Optimizing pessimism in dynamic treatment regimes: A bayesian learning approach. In *International Conference on Artificial Intelligence and Statistics*.

Zhuang, V. and Sui, Y. (2021). No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 3385–3393. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Robust Offline Reinforcement Learning with Heavy-Tailed Rewards: Supplementary Materials

A1 THEORETICAL PROOF

We use c and C to denote some general constants whose values are allowed to vary over time. Under Assumption 1, let $\{(S_i, A_i, R_i, S_i')\}_i$ denote the *i.i.d.* transition tuples.

A1.1 Proof of Theorem 1

Proof. Let $m = n/K$. The key to prove Theorem 1 is to show that for some properly chosen constant $c(\alpha)$ that depends only on α and

$$\Delta \geq c(\alpha)(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\Big[\|\widehat{\omega}^\pi - \omega^\pi\|_\infty + \|\omega^\pi\|_\infty\Big(\frac{1}{m}\Big)^{\frac{\alpha}{1+\alpha}}\Big], \tag{1}$$

then

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{i=1}^m \widehat{\omega}^\pi(S_i, A_i)R_i - J^\pi\Big| \geq \Delta\Big) \leq 0.2. \tag{2}$$

The rest of the proof can similarly be established based on the arguments in the proof of Theorem 1 of Lugosi and Mendelson (2019) and we omit the details to save space.

We focus on proving (2) below. We begin by considering the following decomposition,

$$\frac{1}{m}\sum_{i=1}^m \widehat{\omega}^\pi(S_i, A_i)R_i - J^\pi = \frac{1}{m}\sum_{i=1}^m \omega^\pi(S_i, A_i)R_i - J^\pi + \frac{1}{m}\sum_{i=1}^m [\widehat{\omega}^\pi(S_i, A_i) - \omega^\pi(S_i, A_i)]R_i.$$

For the first term, under Assumptions 2 and 3, the $(1+\alpha)$th moment of $\omega^\pi(S, A)R$ is upper bounded by $\|\omega^\pi\|_\infty \mathbb{E}|R|^{1+\alpha}$. Using the results in Bubeck et al. (2013) and Devroye et al. (2016), we can show that there exists some constant $c > 0$ that depends only α such that

$$\mathbb{P}\Big\{\Big|\frac{1}{m}\sum_{i=1}^m \omega^\pi(S_i, A_i)R_i - J^\pi\Big| \geq c(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\|\omega^\pi\|_\infty\Big(\frac{1}{m}\Big)^{\frac{\alpha}{1+\alpha}}\Big\} \leq 0.1. \tag{3}$$

As for the second term, it is upper bounded by $\|\widehat{\omega}^\pi - \omega^\pi\|_\infty(m^{-1}\sum_{i=1}^m |R_i|)$. Consider $m^{-1}\sum_{i=1}^m |R_i|$. We decompose it into the sum of $m^{-1}\sum_{i=1}^m(|R_i| - \mathbb{E}|R|)$ and $\mathbb{E}|R|$. Similar to (3), we can show $m^{-1}\sum_{i=1}^m(|R_i| - \mathbb{E}|R|)$ satisfies the following,

$$\mathbb{P}\Big\{\Big|\frac{1}{m}\sum_{i=1}^m |R_i| - \mathbb{E}|R|\Big| \geq c(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\Big(\frac{1}{m}\Big)^{\frac{\alpha}{1+\alpha}}\Big\} \leq 0.1. \tag{4}$$

In addition, according to Hölder's inequality, we have $\mathbb{E}|R| \leq (\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}$. This together with (4) implies that the second term can be upper bounded by $C\|\widehat{\omega}^\pi - \omega^\pi\|_\infty(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}$ for some constant $C > 0$, with probability at least 0.9. Combining this together with (1) and (3) yields (2). $\square$

A1.2 Proof of Theorem 2

Proof. Similar to the proof of Theorem 1, it suffices to show that each base OPE estimator satisfies (2) for any

$$\Delta \geq c(\alpha)\lambda_{\min}^{-1}(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\left(\frac{1}{m}\right)^{\frac{\alpha}{1+\alpha}}.$$

Under the realizability assumption and the boundedness assumption on ϕ, the estimation error of the base OPE estimator is of the same order of magnitude as that of the based LSTD estimator $\widehat{\theta}$. It suffices to show that each base $\widehat{\theta}$ satisfies (2) for any

$$\Delta \geq c\lambda_{\min}^{-1}(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\left(\frac{1}{m}\right)^{\frac{\alpha}{1+\alpha}},$$

where c denotes some positive constant that depends only on α.

By definition, $\widehat{\theta}-\theta^*$ equals

$$\begin{aligned} &\Big[\tfrac{1}{m}\textstyle\sum_{i=1}^m \phi(S_i,A_i)\{\phi(S_i,A_i)-\gamma\sum_a \pi(a|S_i')\phi(S_i',a)\}^\top\Big]^{-1} \\ \times\quad &\Big[\tfrac{1}{m}\textstyle\sum_{i=1}^m \phi(S_i,A_i)\{R_i+\gamma\sum_a \pi(a|S_i')Q^\pi(S_i',a)-Q^\pi(S_i,A_i)\}\Big]. \end{aligned}$$

Under the matrix invertibility assumption, using similar arguments in the proof of Lemma 3 of Shi et al. (2022a), we can show that the ℓ_2 norm of the matrix

$$\Big[\frac{1}{m}\sum_{i=1}^m \phi(S_i,A_i)\{\phi(S_i,A_i)-\gamma\sum_a \pi(a|S_i')\phi(S_i',a)\}^\top\Big]^{-1}$$

can be upper bounded by $C\lambda_{\min}^{-1}$ with probability approaching 1. As for the second term, using the results in Bubeck et al. (2013) and Devroye et al. (2016), we can show that its ℓ_2 norm is upper bounded by $\mathbb{E}(|R|^{1+\alpha})^{\frac{1}{1+\alpha}}m^{-\frac{\alpha}{1+\alpha}}$ with probability at least 0.9. Combining these results yield that

$$\mathbb{P}\Big(\|\widehat{\theta}-\theta^*\|_2 \geq c(\mathbb{E}|R|^{1+\alpha})^{\frac{1}{1+\alpha}}\left(\frac{1}{m}\right)^{\frac{\alpha}{1+\alpha}}\Big) \leq 0.2,$$

for some constant $c>0$. The proof is hence completed. $\square$

A1.3 Proof of Theorem 3

We first the following Lemma that would be used in our proof.

Lemma 1 (Bubeck et al. (2013))**.** *Let $\epsilon \in (0,1]$ and $X_1,\ldots,X_n$ be i.i.d. random variable with mean μ and $(1+\epsilon)$-th moment $\mathbb{E}|X-\mu|^{1+\epsilon}=v$. Suppose each fold has N observations such that $n=NK$, then for each $l\in\{1,\ldots,K\}$, the sample mean $\widehat{\mu}_l=\frac{1}{N}\sum\limits_{t=(l-1)N+1}^{lN}X_t$ satisfies*

$$\mathbb{P}\left(|\mu-\widehat{\mu}_l|\geq J\right)\leq \frac{6v}{n^{1+\epsilon}}$$

where for any $J>0$.

Next, we prove Lemma 2, once it holds, we can follow a very similar proof for Theorem 2 to obtain the conclusion in Theorem 3.

Lemma 2. *Under the same notations and conditions in Lemma 1, then $\widehat{\mu}_q^{\mathcal{Q}}$, the q-th quantile of $\{\widehat{\mu}_1,\ldots,\widehat{\mu}_K\}$, satisfies*

$$\mathbb{P}\left(\widehat{\mu}-(12v)^{\frac{1}{1+\epsilon}}\left(\frac{1}{N}\right)^{\frac{\epsilon}{1+\epsilon}}>\widehat{\mu}_q\right)\geq 1-\exp(-2K(2q-1)^2).$$

Proof. According to Lemma 1, for each $l \in \{1, \ldots, K\}$, we have

$$\mathbb{P}\left(\mu - J \leq \widehat{\mu}_l\right) \geq \mathbb{P}\left(|\widehat{\mu}_l - \mu| \leq J\right) \geq 1 - \frac{6v}{n^{\epsilon} J^{1+\epsilon}}.$$

Let $Y_l = I(\mu - J \leq \widehat{\mu}_l)$ with $p := \mathbb{E}(Y_l) \geq 1 - \frac{6v}{n^{\epsilon} J^{1+\epsilon}}$. Then, according to the definition of $\widehat{\mu}_q$, we have

$$\mathbb{P}\left(\mu - J \leq \widehat{\mu}_q\right) = \mathbb{P}\left(\sum_{l=1}^{K} Y_l \geq qK\right) \leq \exp\left(-2K(q-p)^2\right), \tag{5}$$

where the second inequality comes from the Hoeffding inequality. Note that for

$$J = \left(\frac{6v}{q}\right)^{\frac{1}{1+\epsilon}} \left(\frac{1}{N}\right)^{\frac{\epsilon}{1+\epsilon}},$$

we can easily see that $p \geq 1 - q \geq q$ (due to $q \leq 0.5$), and thus, Equation (5) can be simplified to:

$$\mathbb{P}\left(\mu - J \leq \widehat{\mu}_q\right) \leq \exp\left(-2K(2q-1)^2\right).$$

□

A2 ALGORITHM DETAILS

A2.1 The ROAM-MIS Algorithm

Algorithm 3 Robust Off-policy Evaluation via Median-of-means based Marginal Important Sampling (`ROAM-MIS`)

input Policy π, data $\mathcal{D}$, number of data partitions K, decay rate γ, base marginal important sampling ratio estimation algorithm `BaseMIS`

Partition trajectories $\mathcal{D}$ into K parts: $\mathcal{D}_1, \ldots, \mathcal{D}_K$.
for $k = 1, \ldots, K$ **do**
  $\widehat{\omega}_k^{\pi} \leftarrow$ `BaseMIS`$(\pi, \mathcal{D}_k, \gamma)$
end for
$\widehat{J}^{\pi} \leftarrow \text{Median}(\{\mathbb{E}_{\mathcal{D}_k}[\widehat{\omega}_k^{\pi}(S, A)R]\}_{k=1}^{K})$

output $\widehat{J}^{\pi}$

A2.2 The ROOM-FQE Algorithm

Algorithm 4 derives robust intermediate estimators by replacing the heavy-tailed target $Y = R + \gamma \mathbb{E}_{a \sim \pi(S')} Q\left(S', a\right)$ with a MM-type target $Y = R + \gamma \mathbb{E}_{a \sim \pi(S')} \text{Median}(\{\widehat{Q}_k^{\pi}\left(S', a\right)\}_{k=1}^{K})$. However, one issue is that, these estimators $\{\widehat{Q}_k^{\pi}\}_{k=1}^{K}$ (and all estimators after this update including the final ones) in Algorithm 4 are not independent any longer. Therefore, it is not clear whether or not MM can still have theoretical benefits. Thus we only study this variant empirically.

Algorithm 4 Robust Off-policy Evaluation via Median-of-means based Fitted Q-Evaluation (`ROAM-FQE`)

input Policy π, Data $\mathcal{D}$, decay rate γ, number of iterations M, number of partitions K.

Partition data $\mathcal{D}$ into K disjoint parts: $\mathcal{D}_1, \ldots, \mathcal{D}_K$
Initialize K Q-functions $\widehat{Q}_1^{\pi}, \ldots, \widehat{Q}_K^{\pi}$ with corresponding parameters $w_1, \ldots, w_K$
for $m = 1, \ldots, M$ **do**
  for $k = 1, \ldots, K$ **do**
    For each $(S, A, R, S') \in \mathcal{D}_k$, compute: $Y \leftarrow R + \gamma\, \text{Median}(\{\mathbb{E}_{a \sim \pi(S')} \widehat{Q}_k^{\pi}\left(S', a\right)\}_{k=1}^{K})$
    Update $\widehat{Q}_k^{\pi}$ by: $w_k \leftarrow \arg\min_{w_k} \mathbb{E}_{\mathcal{D}_k}(Y - \widehat{Q}_k^{\pi}(S, A; w_k))^2$
  end for
end for
$\widehat{J}^{\pi} \leftarrow \mathbb{E}_{s \sim \mathbb{G}, a \sim \pi(s)} \text{Median}(\{\widehat{Q}_k^{\pi}(s, a)\}_{k=1}^{K})$

output $\widehat{J}^{\pi}$

A2.3 The ROOM-FQI Algorithm

Analogous to Algorithm 4, for iterative OPO algorithms, we can apply MM in every TD update. Take FQI as an example, we replace the definition of Y in the Step 5 of Algorithm 4 by: $Y \leftarrow R + \gamma \,\text{Median}(\{\max_a \widehat{Q}^*_k(S', a)\}_{k=1}^K)$, then we can obtain a robust FQI algorithm.

Algorithm 5 Robust Offline-Policy Optimization via Median-of-mean based Fitted Q-Iteration (`ROOM-FQI`)

input Data $\mathcal{D}$, decay rate γ, number of iterations M, number of data partitions K.
1: Partition data $\mathcal{D}$ into K disjoint parts: $\mathcal{D}_1, \ldots, \mathcal{D}_K$.
2: Initialize Q-functions $\widehat{Q}^*_1, \ldots, \widehat{Q}^*_K$ with parameters $w_1, \ldots, w_K$.
3: **for** $m = 1, \ldots, M$ **do**
4: **for** $k = 1, \ldots, K$ **do**
5: For each $(S, A, R, S') \in \mathcal{D}_k$, compute: $Y = R + \gamma \max_a \text{Median}(\{\widehat{Q}^*_k(S', a)\}_{k=1}^K)$.
6: Update Q_k: $w_k \leftarrow \arg\min_{w_k} \mathbb{E}_{\mathcal{D}_k}(Y - \widehat{Q}^*_k(S, A; w_k))^2$.
7: **end for**
8: **end for**
9: $\widehat{\pi}^*(s) \leftarrow \arg\max_a \text{Median}(\{\widehat{Q}^*_k(s, a)\}_{k=1}^K)$.
output Policy $\widehat{\pi}^*$

Moreover, we also consider its pessimistic variant by using a quantile operator $\mathcal{Q}_q(\cdot)$ rather than the median operator in Step 5 of Algorithm 5. We denoted such a variant as `P-ROOM-FQI`.

A3 EXPERIMENTS: DETAILS

A3.1 Settings for OPE

In the experiments for OPE (see Section 6), we implement the minimax-optimal off-policy evaluation algorithm (Duan et al., 2020) as the benchmarke `FQI` algorithm. Specifically, we use `Ridge` in `scikit-learn` with ℓ_2-penalty fixed at 0.01. The implemented `FQI` algorithm serves as the `BaseOPE` algorithm for `ROAM-DM`. The implementation of `ROAM-FQE` also uses the same `Ridge` to update Q_{w_k} in the Step 6 in Algorithm 4. The maximum number of iterations of all algorithms in Section 6 are fixed at 100.

Next, we discuss the tuning of our algorithms. The only additional tuning parameter of `ROAM`-type algorithms is the number of partitions K, compared with its corresponding base algorithm. In our experiments, fixing $K = 5$ already provides the desired performance and maintains a high computational efficiency. In Appendix A4.4, we try a range of values for K and find that our algorithms are insensitive to this tuning parameter. One may choose this parameter via an adaptive method (Lugosi and Mendelson, 2019) as well.

A3.2 Settings for OPO

A3.2.1 Cartpole environment

For the experiments for OPO at Section 6, we implement the ridge-regression-based `FQI` algorithm as the benchmarked algorithm and the `BaseOPO` algorithm for `ROOM-VM`. The `FQI` uses `Ridge` in `scikit-learn` to solve the optimal Q-function. We implement the `ROOM-FQI` with the same ridge regression with the same settings. For pessimistic variant of `ROOM`-type algorithms, we need an additional tuning parameter q, i.e., the quantile level. As argued in Zhou et al. (2023), the fact that one quantile *explicitly* corresponds to one confidence level makes the tuning much easier than most existing methods where the relationship between the pessimism parameters and the confidence level is implicit and unknown. According to empirical results in Appendix A4.5, we find $q \in [0.1, 0.4]$ perform fairly well. We fix $q = 0.1$ in our experiments.

We also implement a pessimistic-bootstapping OPO method, `PB`, to give a more comprehensive comparison. It is the same as `ROOM-VM` except the two modifications:

- the Step 1 in Algorithm 2 is changed to: "Sample K bootstrapped samples from $\mathcal{D}$: $\mathcal{D}_1, \ldots, \mathcal{D}_K$";

- the Step 5 in Algorithm 2 is modified to:

$$\widehat{\pi}(s) \leftarrow \arg\max_{a} \left[\mathrm{Mean}(\{\widehat{Q}_k^*(s,a)\}_{k=1}^K) - 2 \times \mathrm{Std}(\{\widehat{Q}_k^*(s,a)\}_{k=1}^K) \right] \text{ for any } s.$$

A3.2.2 Mujoco environments

Datasets. All D4RL datasets (Fu et al., 2020) on MuJoCo environments in the experiments are the "v2" version. The datasets on the Kitchen environment are the "v0" version.

Network architecture. The implementations of `SQL` is based on an open-source implementations from GitHub[4], which largely reproduce the results in Kostrikov et al. (2022). Following the same architecture in SQL, both the critic and value networks are two-layer multi-layer perceptron (MLP) with 256 hidden nodes and ReLU activations. We recruit a deterministic policy network whose architecture is the same as critic network.

The implementation of `N-SAC` is upon a public Github repository for SAC[5]. Our implementation completely adopt the same actor-critic architecture in An et al. (2021). Specifically, the critic network is a three-layer MLPs with 256 hidden nodes and ReLU activations. The policy in `SAC-N` is a Gaussian policy network which enables automatic entropy tuning. As for `SAC-N (STD)` to be introduced in Section A4.3, it inherits the same architecture and hyperparameters as `SAC-N`.

Hyperparameters. For the behavior-regularized term α in `SQL`, we set $\alpha = 10$ since Table 3 in Xu et al. (2023) reports $\alpha = 10$ leads to the best average result in MuJoCo environment. For the remaining parameters in `SQL`, we set them as their default hyperparameters, which are listed in Table A1. Notice that, once we complete training `ROOM-VM`, the learned critics can be reused for `MA` and `P-ROOM-VM`. We recruit this programming trick to reduce the time for experiments.

Table A1: The hyperparameters of `SQL` used in the experiments for D4RL tasks.

	Hyperparameter	Value
`SQL`	Optimizer	Adam (Kingma and Ba, 2014)
	Value/Critic learning rate	3^{-4}
	Actor learning rate	Cosine schedule (Loshchilov and Hutter, 2017)
	Critic target update rate	5×10^{-3}
	Mini-batch size	256
	behavior-regularized α	10
`ROOM-VM` & `MA`	Data partition K	5
`P-ROOM-VM`	Quantile q	0.0

We summarized the hyperparameters for train `SAC-N` in Table A2.

Table A2: The hyperparameters of `SAC-N` used in the experiments for D4RL tasks.

Hyperparameter	Value
Critic/actor learning rate	3^{-4}
Critic target update rate	5×10^{-3}
Mini-batch size	256
Ensemble number (a.k.a., K)	10
Temperature	0.2

[4]`https://github.com/gwthomas/IQL-PyTorch`
[5]`https://github.com/pranz24/pytorch-soft-actor-critic`

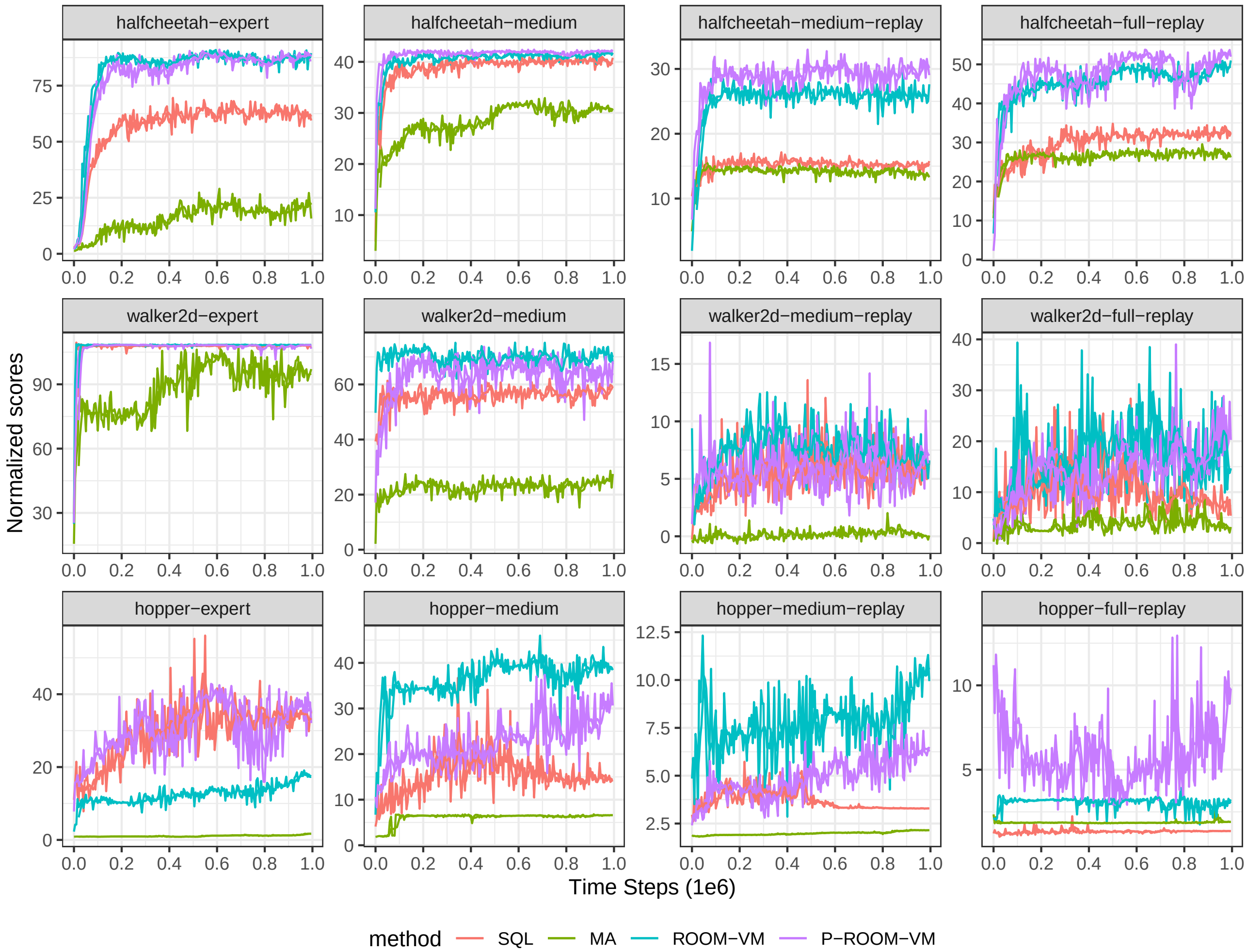

Figure A1: Learning curves of `SQL`, `MA`, `ROOM-VM`, `P-ROOM-VM` on D4RL MuJoCo locomotion datasets.

A3.3 Computation Details

Hardware infrastructure. The experiments in Carpole environment will finish in 5 hours on a personal laptop with 2.6 GHz 6-Core Intel Core i7 and 16 GB memory.

As for the experiments for D4RL datasets, `ROOM-VM` generally consumes 14 hours to train a agent on a task with a machine with GPU Tesla P-100, while `SAC-N` roughly takes round 30 hours to train on the same device.

Time complexity analysis. Let the computational cost of the base algorithm be $c(N)$ for a dataset with N transition tuples. Our algorithm in general yields $\mathcal{O}(K \times c(N/K))$. For those based methods that have a linear computational cost in N (e.g., FQE and FQI; see derivations in Shi et al. (2021)), our computational cost is at the same order. Moreover, our method is embarrassingly parallel.

A4 ADDITIONAL EXPERIMENTS AND RESULTS

A4.1 Learning Curves of `SQL`

Figures A1 and A2 present the learning curves on MuJoCo and Kitchen environments, where the evaluations is conducted every 5000 training steps. Each curve is averaged over 5 seeds and is smoothed by simple moving averages over three periods.

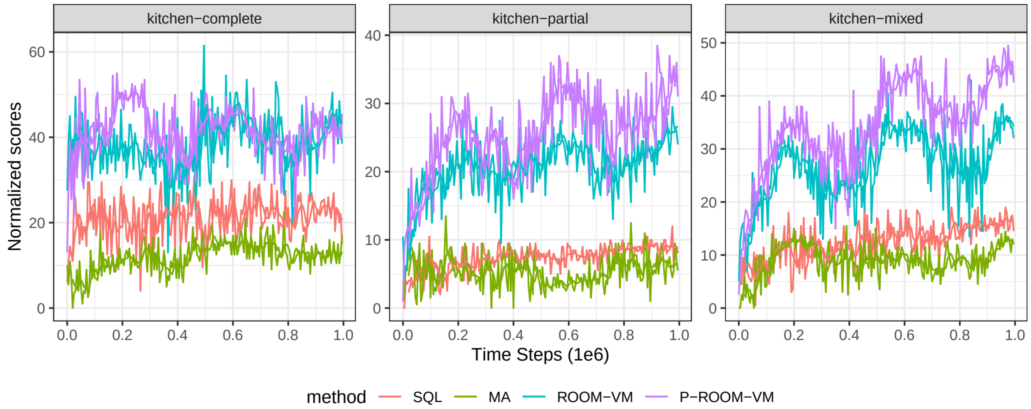

Figure A2: Learning curves of `SQL`, `MA`, `ROOM-VM`, `P-ROOM-VM` on D4RL Kitchen datasets.

A4.2 Comparison with `IQL`

To show that our framework is general, in these datasets, we use another SOTA algorithm, implicit Q-learning (`IQL`, Kostrikov et al. (2022)), as our `BaseOPO` algorithm. We run IQL using the open-source implementation. Notably, the simplicity of ROOM-VM requires minimal modifications of the existing implementation. We adopt the data generation and model evaluation in Section 6.

We report the performance in Table A3. In almost all cases, our method is significantly superior to `IQL`. Notably, `ROOM-VM` can tackle the insufficient coverage challenge by employing `IQL` that can cope with insufficient coverage (Xu et al., 2023). Moreover, the superiority of `ROOM-VM` and `P-ROOM-VM` are clearer in the walker and hopper environments, because these two environments are more challenging than others. For example, the returns of `ROOM-VM` and `P-ROOM-VM` on walker-medium are 20% higher than that of `IQL`, and the returns of `ROOM-VM` and `P-ROOM-VM` on hopper-expert are about 200% of `IQL`'s returns. It is also worth noting that `P-ROOM-VM` shows comparable performance with `ROOM-VM` in most cases, and `P-ROOM-VM` has superior performance in the expert generating datasets because this setup has more severe data insufficient coverage issue.

Table A3: Results for D4RL datasets. Each result is the division of average normalized score of `ROOM-VM` (or `P-ROOM-VM`) and `IQL`. We takes over the final 10 evaluations and 5 seeds. ± captures the 2 times standard error over 5 seeds.

Task Name	`ROOM-VM`	`P-ROOM-VM`
halfcheetah-expert	1.04 ± 0.02	1.03 ± 0.02
walker2d-expert	1.02 ± 0.02	1.02 ± 0.02
hopper-expert	1.89 ± 0.12	2.14 ± 0.13
halfcheetah-medium	0.99 ± 0.01	0.98 ± 0.01
walker2d-medium	1.31 ± 0.08	1.40 ± 0.06
hopper-medium	1.32 ± 0.03	1.28 ± 0.02
halfcheetah-medium-replay	1.04 ± 0.04	1.02 ± 0.05
walker2d-medium-replay	1.56 ± 0.26	0.95 ± 0.16
hopper-medium-replay	1.15 ± 0.14	1.20 ± 0.11

A4.3 Robustness of `SAC-N`

It's noteworthy that `SAC-N` (An et al., 2021) can be interpret to `P-ROOM-FQI`, as it assesses uncertainty by taking the pointwise minimum (i.e., setting $q = 0.0$) of multiple Q-functions but trained on the entire dataset with an soft-actor-critic (SAC) paradigm. Hence, we can be regarded as a heuristic implementation of our approach, and we can expect that it enjoys robustness on heavy-tailed rewards.

To illustrate, we implement `SAC-N` and compare with `SAC-N (STD)`, a method that achieves pessimistic estimation for Q function by pointwisely subtracting two times standard deviation of N functions. To rephrase, `SAC-N (STD)` replaces the pointwise quantile $\mathcal{Q}_q(\{Q_k(s,a)\}_{k=1}^N)$ with $\mathrm{Mean}(\{Q_k(s,a)\}_{k=1}^N) - 2 \times \mathrm{Std}(\{Q_k(s,a)\}_{k=1}^N)$. We demonstrate the numerical performance `SAC-N` and `SAC-N (STD)` on halfcheetah-medium-v2 in Figure A3. From the left panel of Figure A3, we can see that the results of `SAC-N` are highly resembles to the results of Figure 1 in An et al. (2021). More importantly, despite `SAC-N` and `SAC-N (STD)` shares almost the same learning behavior when datasets has no heavy-tailed rewards (see left panel of Figure A3), we can see that `SAC-N` is shown to be much robust to the heavy-tailed noises while `SAC-N (STD)` completely fails at this case. Therefore, it is also highly recommended to use `SAC-N` in environments with heavy-tailed rewards.

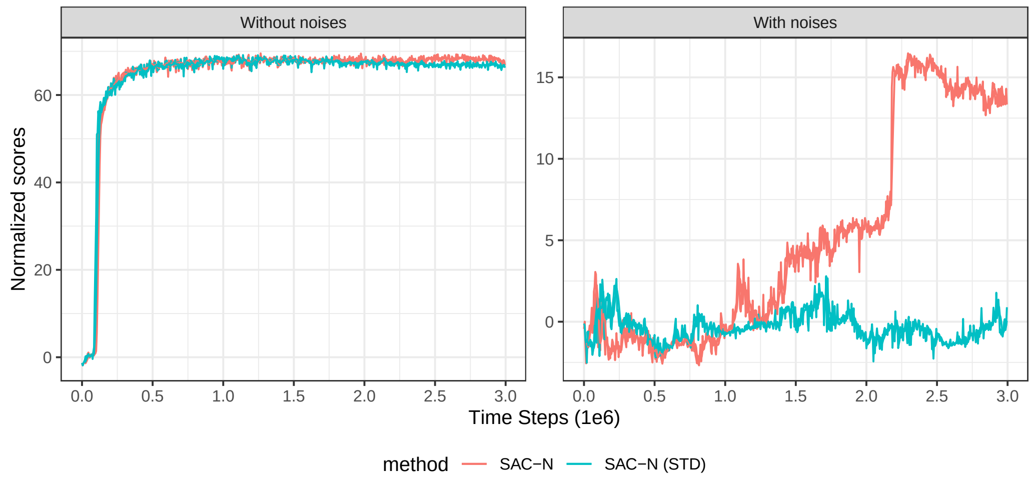

Figure A3: Learning curves comparing the performance of `SAC-N` against `SAC-N-STD` on the halfcheetah-medium-v2 dataset where the left panel corresponds to the case that heavy-tailed noises are not taken into consideration and the right one vice versa. Curves are averaged over 5 seeds and are smoothed by simple moving averages over three periods.

A4.4 Selection of K

In this part, we aim to study how the selection of K influence the performance of `ROAM-DM`. Out of simplicity, we consider values for $K \in \{3, 4, \dots, 10\}$, while the other settings adopt that in Section 6. The estimation error of each algorithm is visualized in Figure A4. From Figure A4, we can see that, our methods exceeds `FQE` for all $K \in \{3, 4, \dots, 10\}$. This implies that, whatever K is taken, our methods are more suitable than `FQE` for offline data with heavy-tailed rewards. It is also worthy to note that, the optimal value of K varies across algorithms and the degree of freedom of the heavy-tailed rewards. In terms of degree of freedom, since it is unknown in practice, there has no general criteria to decide the optimal K. We suggest $K = 5$ as a rule-of-thumb selection for all of our methods because this selection can result in a good performance. Notice that the comparison between ROAM-based methods and mean aggregation methods reveals taking the median operator is crucial for robustness — the mean aggregation achieves a poor performance.

A4.5 Selection of Pessimistic

In this part, we aim to study how the selection of q influence the performance of `ROAM-DM` and `ROOM-FQI`. We consider values for $q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, while the other settings adopt that in the experiments on Cartpole environment. We visualize the regret of each algorithm in Figure A4, from which we see that, our methods surpass `FQI` whatever the value of q. Besides, like the choice for K, both algorithms and the degree of freedom of the heavy-tailed rewards have an impact on the optimal value of q. Figure A5 shows $q \in [0.1, 0.4]$ is a rule-of-thumb criterion for the guarantee of admirable numerical performance.

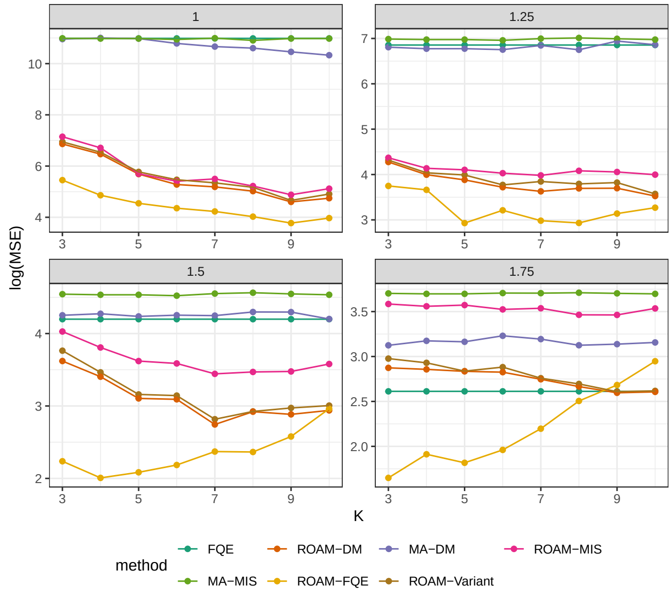

Figure A4: The Ablation on the tuning parameter K for the OPE problem at the Cartpole environment. Each panel corresponds to a certain degree of freedom.

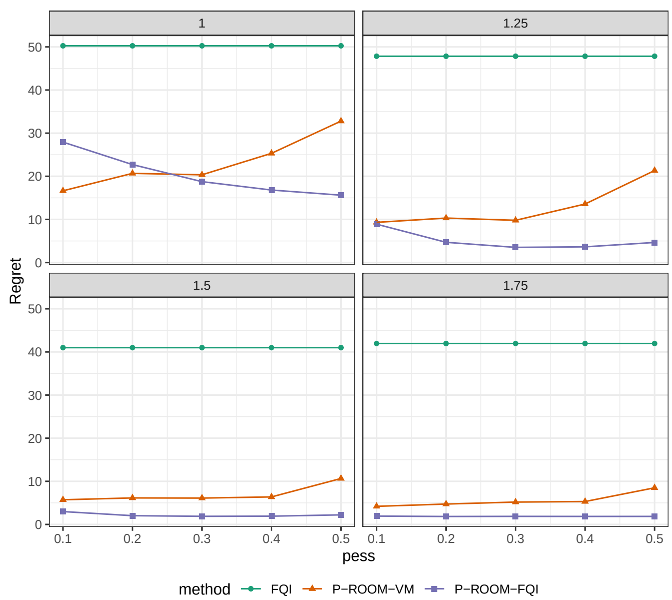

Figure A5: The ablation on the tuning parameter q for the OPE problem at the Cartpole environment. Each panel corresponds to a certain degree of freedom.

A5 BROADER IMPACT STATEMENT

Our approach provides offline RL methods to be applied to systems with heavy-tailed rewards. While our method can properly handle heavy-tailed rewards, it may also neglect the potential societal impact. For instance, heavy-tailed rewards in finance system may arise from fraudulent transactions or attacks on banking systems, which deserves adequate attention. One possible approach to monitor heavy-tailed rewards involves measures the gap between the two sides of Bellman equation. If the resulting value exhibits high variance, then the rewards warrants monitoring.