Privacy Preserving Reinforcement Learning with One-Sided Feedback

Lin Cong[1,3*], **Guangyan Gan**[1*], **Hanzhang Qin**[2*] and **Zhenzhen Yan**[1*]
[1]Nanyang Technological University
[2]National University of Singapore
[3] Cornell SC Johnson College of Business
will.cong@ntu.edu.sg, GUANGYAN001@e.ntu.edu.sg, hzqin@nus.edu.sg, yanzz@ntu.edu.sg

Abstract

We study reinforcement learning (RL) in multi-dimensional continuous state and action spaces with one-sided feedback, where the agent receives partial observations of the state and obtains reward information for only a subset of the state-action space at each time step. This setting introduces substantial challenges in both learning efficiency and privacy preservation. To address these challenges, we propose POOL, a novel privacy-preserving RL algorithm. We conduct a comprehensive theoretical analysis of POOL, deriving a sample complexity bound of $\tilde{O}((1 + E_\rho)H^3\alpha^{-2})$, which matches the known lower bounds for non-private RL. Here, E_ρ denotes the privacy parameter, H is the time horizon, and α is optimality-gap parameter. Our findings show that it is possible to enforce strong privacy guarantees while maintaining high learning efficiency, marking a significant step toward practical, privacy-aware RL in multi-dimensional environments with one-sided feedback.

1 Introduction

Reinforcement learning (RL) with one-sided feedback refers to a learning setting within a Markov Decision Process (MDP) where the agent receives feedback (e.g., rewards or next-state observations) only for a restricted subset of state-action pairs. Specifically, feedback is observable only when the state-action pair falls within a predefined observable boundary, while no information is revealed otherwise. This partial observability introduces structural challenges in learning transition dynamics and estimating value functions, distinguishing it from traditional RL settings with full feedback. These challenges make standard RL techniques less effective, motivating the study of RL methods under one-sided feedback.

In many real-world applications, online data collection is infeasible or expensive [Levine *et al.*, 2020], which motivates the use of offline RL. The objective of offline RL is to learn an optimal policy from pre-collected datasets without further interactions with the environment. This setting is particularly relevant in domains such as robotic control, clinical trials, and education [Zhou *et al.*, 2024; Rengarajan *et al.*, 2024; Nie, 2025]. Consequently, in this paper, we focus on offline RL under one-sided feedback.

While offline RL has been extensively studied, research on RL with one-sided feedback remains limited and predominantly confined to one-dimensional settings [Qin *et al.*, 2023; Gong and Simchi-Levi, 2023; Gong and Simchi-Levi, 2020]. Moreover, many theoretical studies focus on tabular MDPs with finite, one-dimensional state and action spaces. In contrast, real-world applications such as marketing, autonomous systems, and healthcare involve *continuous* and *multi-dimensional* state-action spaces and often operate under partial or censored feedback. Bridging this gap requires scalable RL algorithms capable of handling complex, multi-dimensional environments with one-sided feedback.

Furthermore, many applications involve sensitive data, making *privacy* a critical concern [Qiao and Wang, 2023; Qiao and Wang, 2024]. For example, in healthcare decision-making, RL algorithms are often used to optimize sequential treatment policies based on patient data [Raghu *et al.*, 2017]. Without proper safeguards, such data usage can lead to privacy breaches and regulatory noncompliance. Existing research on differentially private RL has largely been limited to single-dimensional, tabular RL with full information on state–action pairs. Our work addresses this dual challenge: one-sided feedback in multi-dimensional continuous RL while ensuring differential privacy.

We propose POOL (Privacy-Oriented One-Sided Learning), a novel privacy-preserving RL algorithm for multi-dimensional continuous MDPs with one-sided feedback. Prior works have either focused on privacy in tabular RL or on one-sided feedback without privacy considerations [Qiao and Wang, 2023; Qiao and Wang, 2024; Garcelon *et al.*, 2021; Hossain and Lee, 2023; Qin *et al.*, 2023]. To the best of our knowledge, POOL is the first algorithm to jointly address both challenges in continuous, multi-dimensional environments.

Technical innovations. Extending private RL frameworks from tabular full-feedback settings [Qiao and Wang, 2024] to continuous, multi-dimensional environments with one-sided feedback introduces significant challenges. These arise from the high computational complexity of infinite state and action spaces and the systematic bias induced by partial observabil-

*Authors are listed in alphabetical order.

ity. We tackle these challenges via a carefully designed *partial discretization strategy* and *multi-dimensional piecewise-linear approximation* techniques. Together, these innovations form the first scalable, privacy-preserving RL approach with provable guarantees for continuous, multi-dimensional settings under one-sided feedback. Table 1 compares our results with existing work, highlighting improvements in dimensionality, privacy, and feedback handling.

Our contributions. Building on these innovations, our key contributions are summarized as follows:

- Problem formulation: We present the *first* privacy-preserving RL framework for multi-dimensional continuous MDPs under one-sided feedback, addressing a previously unexplored gap in the literature.
- Algorithmic innovation: We introduce POOL, a scalable privacy-preserving RL algorithm that integrates partial discretization with multi-dimensional piecewise-linear approximation. This design enables efficient learning in continuous, multi-dimensional spaces while maintaining strong privacy guarantees and outperforming existing approaches under one-sided feedback.
- Theoretical guarantees: We prove that POOL satisfies ρ-zero-concentrated differential privacy (ρ-zCDP) (Definition 4) and establish a sample complexity upper bound of $\tilde{O}\left((1+E_\rho)H^3\alpha^{-2}\right)$, which matches the information-theoretic lower bound while ensuring rigorous privacy protection (Theorem 2). This provides formal assurances on both privacy and learning performance.
- Empirical validation: We conduct extensive experiments in the inventory control problem to demonstrate that POOL not only preserves privacy but also achieves superior performance compared to existing baselines under one-sided feedback.

Related work. Recent years have witnessed significant advancements in privacy-preserving RL [Vietri *et al.*, 2020; Shariff and Sheffet, 2018; Guha Thakurta and Smith, 2013]. [Qiao and Wang, 2023] introduced near-optimal private policy learning for offline RL, while [Qiao and Wang, 2024] developed differentially private exploration strategies that achieve sublinear regret in online RL. However, these methods are restricted to discrete settings under full feedback, leaving the more complex continuous environments with one-sided feedback largely unexplored.

In parallel, a substantial body of work has focused on RL with continuous state and action spaces, which are crucial in various control applications [Munos and Bourgine, 1997; Jia and Zhou, 2022; Zhao *et al.*, 2024; Wang *et al.*, 2020; Antos *et al.*, 2007; Carrara *et al.*, 2019]. Discretization and function approximation methods [Qin *et al.*, 2023; Gong and Simchi-Levi, 2020] have shown promise in these settings. However, these advances either operate in restrictive one-dimensional environments or fail to consider privacy concerns [Qin *et al.*, 2023; Gong and Simchi-Levi, 2020]. To date, no method simultaneously addresses privacy, multi-dimensionality, and continuous control.

RL with one-sided feedback has gained substantial attention due to its practical relevance in a wide range of problems [Gong and Simchi-Levi, 2020; Zhao and Chen, 2019; Yuan *et al.*, 2021]. This includes complex scenarios like second-price auction design [Haoyu and Wei, 2020], inventory control in lost-sales settings [Qin *et al.*, 2023; Gong and Simchi-Levi, 2023], and dynamic product pricing [Cohen *et al.*, 2020]. For example, [Feng *et al.*, 2018] explores learning in repeated auction environments where bidders do not know their valuation for the items being auctioned, leveraging partial feedback structures. [Zhao and Chen, 2019] addresses one-sided information feedback in bandit problems, while [Lobel *et al.*, 2018] studies multidimensional search problems under one-sided feedback. [Gong and Simchi-Levi, 2020] introduces an efficient Q-learning framework for one-sided feedback. However, these works focus primarily on one-dimensional settings and neglect privacy concerns. To the best of our knowledge, we are the first to study multi-dimensional continuous RL with one-sided feedback while incorporating rigorous privacy guarantees.

This paper closes a crucial gap. To the best of our knowledge, this paper is the first to develop scalable privacy-preserving RL algorithms for multi-dimensional continuous MDPs with one-sided feedback. Our framework introduces a scalable partial discretization strategy and a piecewise-linear approximation scheme, enabling rigorous privacy guarantees in complex environments that were previously considered intractable.

Notation. Throughout this paper, we denote the ℓ_p norm of a vector $\boldsymbol{a}$ by $\|\boldsymbol{a}\|_p$. For any $H \in \mathbb{N}_+$, let $[H]$ denote the set $\{1, 2, \ldots, H\}$. The natural logarithm is denoted by $\log(\cdot)$. Let $\mathbf{1} = [1, 1, \ldots, 1] \in \mathbb{R}^{w+d}$ denote the all-one vector of dimension $w + d$. We use the standard asymptotic notations $O(\cdot)$ and $\Omega(\cdot)$ as defined in [Cormen *et al.*, 2022]. When logarithmic factors are omitted, we use $\tilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$, respectively.

2 Preliminaries

We first outline the framework of multi-dimensional RL with one-sided feedback in Section 2.1, followed by a brief overview of differential privacy in Section 2.2.

2.1 Multi-dimensional RL with One-Sided Feedback

We consider a finite-horizon episodic MDP with continuous, multi-dimensional state and action spaces, specified by the tuple $\left(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H, d_1\right)$, where $\mathcal{S} = [0,1]^w$ denotes the w-dimensional state space, $\mathcal{A} = [0,1]^d$ denotes the d-dimensional action space, and H is the episode horizon. In contrast to classical tabular MDPs with discrete state-action spaces, we consider a setting where both states and actions lie in continuous spaces. Such settings frequently arise in real-world applications. For instance, in inventory management, where the inventory level represents the state and the order quantity serves as the action, both of which are naturally continuous. Other examples include energy management systems, where battery levels and power output represent the

Literature	Finite/Continuous	Dimension	Private	One-Sided	Sample Complexity
[Qin *et al.*, 2023]	Continuous	One	✗	✓	$\tilde{O}\left(H^3\alpha^{-2}\right)$
[Sidford *et al.*, 2018]	Finite	One	✗	✗	$\Omega\left(H^{-3}\alpha^{-2}\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\right)$
[Qiao and Wang, 2024]	Finite	One	✓	✗	$\tilde{O}\left((1+E_\rho)H^3\alpha^{-2}\right)$
This paper	Continuous	Multi	✓	✓	$\tilde{O}\left((1+E_\rho)H^3\alpha^{-2}\right)$

Table 1: Comparison of our results to existing work. Here, H denotes the episode length, α is the optimality gap, E_ρ is a privacy-related parameter whose precise definition will be provided in Theorem 2, and $|\mathcal{S}|$ and $|\mathcal{A}|$ represent the cardinalities of the state and action spaces, respectively.

state and action, respectively; financial portfolio optimization, with asset prices as states and allocation proportions as actions; and auction design, where the bidder value distributions form the state and the reserve price serves as the action. At each time step $h \in [H]$, the (potentially non-stationary) transition kernel $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ maps each state-action pair (s_h, a_h) to a probability distribution $P_h(\cdot \mid s_h, a_h)$ over next states. The reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ is assumed to be known.[1] The initial state distribution is denoted by d_1. A policy $\pi = (\pi_1, \ldots, \pi_H)$ consists of stochastic decision rules, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ maps a state to a distribution over actions at step h. At each step, the agent observes s_h, samples $a_h \sim \pi_h(\cdot \mid s_h)$, receives reward $r_h(s_h, a_h)$, and transitions to $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$. The episode terminates after H steps. A trajectory $(s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1})$ is generated according to: $s_1 \sim d_1$, $a_h \sim \pi_h(\cdot \mid s_h)$, $r_h = r_h(s_h, a_h)$, and $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$ for each $h \in [H]$. For notational clarity, we omit explicit arguments when there is no risk of ambiguity. For example, we write P_h instead of $P_h(s_{h+1} \mid s_h, a_h)$, and similarly simplify other functions such as V_h^π for $V_h^\pi(s_h)$ and $Q_h^\pi = Q_h^\pi(s_h, a_h)$, when the context is clear.

Value functions and bellman equations. The value and Q-value functions under policy π are defined by: $V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=h}^H r_t \mid s_h = s\right]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=h}^H r_t \mid s_h = s, a_h = a\right]$, for all $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. We use the Bellman equations to characterize the system dynamics: $Q_h^\pi = r_h + P_h V_{h+1}^\pi$, where $V_h^\pi = \mathbb{E}_{a\sim\pi_h}[Q_h^\pi]$. The value and Q-functions under the optimal policy are given by: $Q_h^\star = r_h + P_h V_{h+1}^\star$, $V_h^\star = \max_a Q_h^\star(\cdot, a)$. The performance metric of each policy π is defined by $v^\pi := \mathbb{E}_{d_1}[V_1^\pi] = \mathbb{E}_{\pi,d_1}\left[\sum_{t=1}^H r_t\right]$. We consider an offline RL setting and employ a backward induction approach to solve the optimization problem, initializing with $V_{H+1}^\pi = 0$ for any π.

Full-Feedback. Under the full-feedback setting, both rewards and transitions for *all* state–action pairs are fully observable, making them entirely learnable.

One-Sided feedback. We first introduce observable boundaries, denoted by $\boldsymbol{\lambda}_h = [\lambda_h^1, \lambda_h^2, \ldots, \lambda_h^{w+d}]$ to capture the feedback structure. This notion is motivated by [Qin *et al.*, 2023], which focuses on one-dimensional RL without addressing privacy guarantee.

Definition 1 (Observable Boundary Comparison)**.** *For any state-action pair $b = (s,a) \in \mathcal{S} \times \mathcal{A}$ and observable boundary $\boldsymbol{\lambda} = [\lambda^1, \ldots, \lambda^{w+d}]$, we define: (1) $b < \boldsymbol{\lambda}$ if $b^i < \lambda^i$ for all $i \in [w+d]$; (2) $b \geq \boldsymbol{\lambda}$ if there exists $i \in [w+d]$ such that $b^i \geq \lambda^i$.*

We now formally introduce one-sided feedback:

Definition 2 (One-Sided Feedback)**.** *At period h, the agent observes rewards and transitions only for state-action pairs lying on* one side *of the observable boundary $\boldsymbol{\lambda}_h$. Specifically: (1) Under lower one-sided feedback, rewards and transitions are revealed for state-action pairs satisfying $b_h = (s_h, a_h) < \boldsymbol{\lambda}_h$; (2) Under higher one-sided feedback, rewards and transitions are revealed for state-action pairs satisfying $b_h > \boldsymbol{\lambda}_h$. No feedback is available for state-action pairs on the opposite side of the boundary.*

One-sided feedback naturally arises in applications such as inventory control with lost sales. A concrete example is provided in Appendix. Our techniques apply symmetrically to lower and higher one-sided feedback. We focus on lower one-sided feedback without loss of generality.

Objective. Consider a dataset $\mathcal{D} = \left\{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\right\}_{\tau\in[n]}^{h\in[H]}$ collected under an unknown behavior policy μ. By *unknown behavior policy*, we mean that the data were generated by some fixed, possibly stochastic, policy μ, whose form or parameters are not accessible to the learner. This setting is standard in offline RL, where interaction with the environment is prohibited and the only available information is the static dataset $\mathcal{D}$. The goal of offline reinforcement learning is to find a policy π_{alg} such that $v^\star - v^{\pi_{\text{alg}}} \leq \alpha$ for a target accuracy level $\alpha > 0$, while ensuring *differential privacy* (see Section 2.2).

2.2 Differential Privacy

Differential privacy (DP) provides a principled framework for ensuring privacy in learning algorithms. The standard DP definition [Dwork, 2006] is as follows:

Definition 3 ((ϵ,δ)-DP)**.** *A randomized mechanism $\mathcal{M}$ satisfies (ϵ,δ)-DP if for any pair of neighboring datasets $\mathcal{U}$ and $\mathcal{U}'$, which differ by exactly one data point, and for any measurable event E: $\mathbb{P}[\mathcal{M}(\mathcal{U}) \in E] \leq e^\epsilon \cdot \mathbb{P}[\mathcal{M}(\mathcal{U}') \in E] + \delta$.*

In offline RL, each data point corresponds to a trajectory, and our goal is to design privacy-preserving algorithms that

[1] This is a standard assumption in control and offline RL settings, where reward uncertainty is negligible compared to transition uncertainty; see [Qiao and Wang, 2024].

guarantee trajectory-level privacy. We adopt ρ-zCDP (defined in Definition 4) as our primary privacy notion due to its analytical simplicity, following [Qiao and Wang, 2024].

Definition 4 (ρ-zCDP [Bun and Steinke, 2016])**.** *A randomized mechanism $\mathcal{M}$ satisfies ρ-Zero-Concentrated Differential Privacy (ρ-zCDP) if for all pairs of neighboring datasets $\mathcal{U}$,$\mathcal{U}'$ and all $\alpha > 1$: $R_\alpha\left(\mathcal{M}(\mathcal{U}) \,\|\, \mathcal{M}(\mathcal{U}')\right) \leq \rho\alpha$, where R_α is the Rényi divergence of order α of $\mathcal{M}(\mathcal{U})$ from $\mathcal{M}(\mathcal{U}')$ [Van Erven and Harremos, 2014].*

Additional technical properties of ρ-zCDP (e.g., composition) are provided in Section 3 of the Appendix. We adopt ρ-zCDP in the subsequent analysis of privacy guarantees. It is worth noting that ρ-zCDP guarantees can be translated into (ϵ, δ)-DP through an appropriate parameter mapping. We refer interested readers to Lemma 2 in the Appendix for details.

3 Privacy Preserving RL with One-Sided Feedback

In this section, we propose a differentially private algorithm named POOL for multi-dimensional RL with one-sided feedback that ensures ρ-zCDP. The key algorithmic components are outlined in Section 3.2, while detailed analysis of the privacy mechanisms is deferred to Appendix due to space constraints.

3.1 Design Overview

Algorithm 1 provides an overview of POOL, which builds upon the framework introduced by [Qiao and Wang, 2024]. They constructs private value functions for RL in finite state-action spaces with full feedback. To extend this framework to multi-dimensional, continuous environments with one-sided feedback, we introduce a tailored partial discretization technique that addresses the complexity of continuous action spaces and the partial observability characteristic of one-sided feedback. In addition, we propose a multi-dimensional piecewise-linear approximation to represent the value function over discretized action zones, facilitating efficient and accurate estimation under privacy constraints. In summary, our algorithm proceeds iteratively over each stage $h \in [H]$ with the following steps:

- Step 1: Discretization. Apply the partial discretization technique to divide the multi-dimensional action space into M zones based on the ℓ_2 norm (Line 5 in Algorithm 1).
- Step 2: Private estimation. Construct differentially private value function estimates for each discretized region (Line 7 in Algorithm 1).
- Step 3: Piecewise-linear approximation. Interpolate the value function using a multi-dimensional piecewise-linear approximation based on sampled points within each zone (Line 13 in Algorithm 1).

3.2 Key Ingredients

In this section, we introduce three key ingredients of our proposed algorithm: the *Gaussian mechanism*, a *discretization strategy*, and a *multi-dimensional piecewise-linear approximation*. These components are crucial for efficiently addressing the challenges posed by multi-dimensional RL problems with one-sided feedback.

Gaussian mechanism. The Gaussian mechanism is one of the most widely used techniques in the differential privacy literature for preserving privacy [Dwork *et al.*, 2014]. A straightforward approach in multi-dimensional RL is to add Gaussian noise directly to sensitive data, but this often leads to severe performance degradation [Chen *et al.*, 2022]. To address this, we adopt a more principled approach by applying the Gaussian mechanism to privatize estimates of visit counts, transition kernels, and value functions. Specifically, following [Qiao and Wang, 2024], we add Gaussian noise to the counts of state-action pairs, with the variance of this noise calibrated according to the sensitivity and privacy parameters. To achieve ρ-zCDP, we build on the privatized counts from the Gaussian mechanism to construct estimators, such as private transition probabilities denoted by $\widetilde{P}_h$ and private value functions denoted by $\widetilde{Q}_h$ and $\widetilde{V}_h$, respectively (for further details, see Section 2 in Appendix). This approach ensures differential privacy throughout the learning process while maintaining estimation accuracy.

Discretization strategy. A natural approach to solving continuous RL problems is to discretize the state-action space and apply tabular RL methods [Qin *et al.*, 2023; Gong and Simchi-Levi, 2023]. However, this approach scales poorly in multi-dimensional settings due to the exponential growth in discrete zones when each dimension is discretized independently. To illustrate, consider a state-action space with $w + d$ dimensions, each discretized into q bins. This results in q^{w+d} discrete zones, which grows exponentially with the number of dimensions. For example, with 10 dimensions and 5 bins per dimension, the total reaches 5^{10}, rendering the method impractical for multi-dimensional problems.

Another challenge in RL with one-sided feedback is the partial observability of state-action pairs and their corresponding transition kernels and value functions. Existing approaches often use partial discretization of the action space [Qin *et al.*, 2023; Gong and Simchi-Levi, 2023], where the observable region is divided into M equal-length segments, and unobserved regions are assigned the estimated Q-values for the boundary. However, this method is tailored to one-dimensional settings and does not scale effectively to the multi-dimensional scenarios common in practice.

To address these challenges, we propose a scalable discretization strategy that partitions the $(w + d)$-dimensional state-action space into M zones based on the ℓ_p norm, constrained to the observable region, i.e., $(s_h, a_h) < \boldsymbol{\lambda}_h$. For simplicity and geometric interpretability, we adopt the ℓ_2 norm.A vector $b \in \mathbb{R}^{w+d}$ is assigned to zone $m \in [M]$ if and only if $\frac{(w+d)(m-1)}{M} \leq \|b\|_2 < \frac{(w+d)m}{M}$, where M is the discretization level, which influences the approximation error of the piecewise-linear model in bounding the performance loss.

This construction leads to intuitive geometric partitions, especially in low-dimensional cases. For example, in one dimension, the space is divided into M equal-length in-

tervals —an approach that aligns with standard discretization technique used in the literature [Qin *et al.*, 2023; Gong and Simchi-Levi, 2023].

This norm-based discretization preserves rotational symmetry and enables structured, zone-wise approximation. Compared to traditional axis-aligned discretization, it reduces the number of discrete regions, improving scalability for multi-dimensional RL tasks.

Multi-dimensional piecewise-linear approximation. Most existing research applies a piecewise-linear approximation to estimate value functions in one-dimensional RL. However, to the best of our knowledge, this approach has not been extended to multi-dimensional settings. To enable scalability in multi-dimensional setting, we iteratively construct a set of $w+d$ linearly independent vectors $\{b_m^1, \ldots, b_m^{w+d}\}$ to span each discretized zone. Any finite-dimensional vector space admits a basis, and since each zone lies in the finite-dimensional space $\mathbb{R}^{w+d}$, such a basis always exists (Theorem 1 in Appendix). We begin by randomly selecting an initial vector b_m^1 in zone m. Subsequently, we iteratively construct additional vectors within zone m, such as b_m^2, ensuring each new vector is linearly independent of the previously selected ones. The process continues until a complete set of $w+d$ linearly independent vectors $\{b_m^1, \ldots, b_m^{w+d}\}$ is obtained.

To improve computational efficiency, we employ orthogonal basis vectors, as their independence enables direct computation of projection coefficients without solving linear systems. This significantly accelerates interpolation. Once the $w+d$ linearly independent vectors are identified, we apply the Gram-Schmidt orthogonalization process [Strang, 2022] to construct orthogonal basis vectors.

For any state-action pair (s_h, a_h) within a zone, we approximate its Q-value as a linear combination of the Q-values at the basis vectors. Specifically, we represent (s_h, a_h) as $(s_h, a_h) = \sum_{j=1}^{w+d} m_j b_m^j$. Then its Q-value is approximated by $\underline{Q}_h(s_h, a_h) = \sum_{j=1}^{w+d} m_j \overline{Q}_h(b_m^j)$, where $\underline{Q}_h(s_h, a_h)$ denotes the approximated Q-value at any continuous state-action pair, and $\overline{Q}$ refers to the Q-value at discrete base points. The corresponding value function is then given by: $\widetilde{V}_h(s_h) = \max_{a_h} \underline{Q}_h(s_h, a_h)$. For state-action pairs that fall outside the observable boundary, i.e., $(s_h, a_h) \geq \boldsymbol{\lambda}_h$, the estimated Q-value is assigned a constant value equal to that at the boundary.

4 Theoretical Analysis

We now present a formal analysis of Algorithm 1, establishing its differential privacy guarantee and analyzing its sample complexity, i.e., the number of samples required to ensure both privacy preservation and near-optimal performance guarantee (Theorem 2).

We begin by introducing a mild regularity condition on the reward function.

Assumption 1. *For all $h \in [H]$, the reward function $\widetilde{r}_h$ satisfies the following generalized Lipschitz condition:*

$$|\widetilde{r}_h(b_h) - \widetilde{r}_h(b_h')| \leq l_h \left| \|b_h\|_2 - \|b_h'\|_2 \right|,$$

Algorithm 1 Privacy-Oriented One-Sided Learning (POOL)

Input: Offline dataset $\mathcal{D}$, reward r, constants $C_1 = \sqrt{2}, C_2 = 16, C > 1$, failure probability δ, zCDP budget ρ, E_ρ and ι
Initialize: Estimate $\widetilde{P}_h$, set $\widetilde{V}_{H+1} = 0$
for $h = H$ to 1 **do**
  if $(s_h, a_h) < \boldsymbol{\lambda}_h$ **then**
    Partition the state-action space into K zones according to ℓ_2 norm
    for each zone $m \in [M]$ and basis vector b_m^j, $j \in [w+d]$ **do**
      $\widetilde{Q}_h(b_m^j) \leftarrow r_h(b_m^j) + \widetilde{P}_h \cdot \widetilde{V}_{h+1}(b_m^j)$
      Compute $\Gamma_h(b_m^j)$ (or CH if $n \leq E_\rho$)
      $\widehat{Q}_h^p(b_m^j) \leftarrow \widetilde{Q}_h(b_m^j) - \Gamma_h(b_m^j)$
      $\overline{Q}_h = \min\left\{\widehat{Q}_h^p, H-h+1\right\}^+$
    end for
    Represent (s_h, a_h) in zone m as: $(s_h, a_h) = \sum_{j=1}^{w+d} m_j b_m^j$
    $\underline{Q}_h(s_h, a_h) = \sum_{j=1}^{w+d} m_j \overline{Q}_h(b_m^j)$
  end if
  $\underline{Q}_h(s_h, a_h) = \overline{Q}_h(\boldsymbol{\lambda}_h)$
  $\widetilde{\pi}_h(\cdot \mid s_h) = \arg\max_\pi \mathbb{E}_{a_h \sim \pi}[\underline{Q}_h(s_h, a_h)]$
  $\widetilde{V}_h(s_h) = \mathbb{E}_{a_h \sim \widetilde{\pi}_h(\cdot|s_h)}\left[\underline{Q}_h(s_h, a_h)\right]$
end for
Output: Policy $\{\widetilde{\pi}_h\}_{h=1}^H$

where l_h denotes a generalized Lipschitz constant.

This condition is relatively mild, as it is satisfied by a broad class of distributions (see Appendix for details), and frequently arises in practical domains such as inventory control and capital management. For example, [Qin *et al.*, 2023] adopt a similar condition in a one-dimensional RL setting.

Notably, this assumption modifies the traditional Lipschitz condition to depend on differences of norms rather than norms of differences. This adjustment reflects that the reward primarily depends on the magnitude of the state–action vector rather than its direction, yielding a less restrictive yet sufficient smoothness condition for multi-dimensional RL analysis. Furthermore, it facilitates bounded discretization error, enabling efficient learning in high-dimensional settings.

We first derive the following bound on the piecewise-linear approximation error:

Theorem 1 (Bound on Piecewise-Linear Approximation Error)**.** *For any $b_h = (s_h, a_h) \in [0,1]^{w+d}$, the piecewise-linear approximation error is bounded as follows:* $\left|\underline{Q}_h(b_h) - \overline{Q}_h(b_h)\right| \leq \frac{L_h\sqrt{w+d}}{M} + L_h \left|w + d - \|\boldsymbol{\lambda}_h\|_2\right|$, *where $L_h = (H-h+1)L$, and $L = \max_h l_h$.*

Theorem 1 highlights key factors influencing the accuracy of piecewise-linear approximation. Specifically, the approximation error increases when feedback is highly censored (i.e., small $\|\boldsymbol{\lambda}_h\|_2$) or when the state-action space is multi-

dimensional, underscoring the challenges of one-sided feedback in multi-dimensional RL. Conversely, increasing the number of discretized zones M can help reduce this error.

Next, we state our main theoretical result characterizing both the privacy and performance guarantees.

Theorem 2. *(1) Algorithm 1 satisfies ρ-zCDP. (2) For any target accuracy $\alpha > 0$, let $M = \Theta\left(\frac{w+d}{\alpha}\right)$ denote the number of discretization zones. If the number of samples per stage satisfies $n = \tilde{O}\left(\frac{H^3(1+E_\rho)}{\alpha}\right)$, then the learned policy $\widetilde{\pi}$ satisfies $v^\star - \alpha - \tilde{O}\left(L_1 \left|w + d - \|\boldsymbol{\lambda}_h\|_2\right|\right) \leq v^{\widetilde{\pi}} \leq v^\star$ with probability at least $1-\delta$, where $E_\rho = 4\sqrt{\frac{H \log\left(\frac{4HM^2(w+d)w}{\delta}\right)}{\rho}}$, $L_h = (H-h+1)L$, and $L = \max_h l_h$.*

Theorem 2 provides an upper bound, across problem instances, on the minimal number of samples needed to ensure that the learned policy is both privacy-preserving and near-optimal. Such an upper bound is crucial for understanding the theoretical and practical efficiency of RL algorithms. It provides performance guarantees, ensuring that the learned policy is privacy-preserving and approximates the optimal policy within a desired accuracy level given sufficient data. It also offers insights into the sample efficiency of the method, particularly important in offline settings where data collection is constrained. Moreover, this upper bound can serve as a benchmark for comparing algorithms and guide the design of more efficient learning strategies by identifying critical factors—such as dimensionality and feedback sparsity, that impact learning complexity. According to the theorem, this bound scales polynomially with the episode length H, the discretization granularity M, the combined dimension $(w+d)$, and inversely with the privacy budget ρ. As expected, stronger privacy (i.e., smaller ρ) requires more samples to maintain the same level of optimality.

Quality of the bound. The sample complexity bound of POOL is tight up to logarithmic factors in the full-feedback regime, which corresponds to the case where $|w + d - \|\boldsymbol{\lambda}_h\|_2| = 0$. In this case, our upper bound $\tilde{O}\left(H^3\alpha^{-2}\right)$ matches the information-theoretic lower bounds $\tilde{\Omega}\left(H^3\alpha^{-2}\right)$ established by [Sidford *et al.*, 2018]. Furthermore, in the special case where the problem is one-dimensional ($d = 1$), our sample complexity bound aligns with the established result under one-sided feedback from [Qin *et al.*, 2023], validating our generalization to higher-dimensional settings as a significant advancement in the theory of RL under partial feedback. Finally, compared to the upper bound $\tilde{O}((1+E_\rho)H^3\alpha^{-2})$ established in [Qiao and Wang, 2024] for private RL in one-dimensional tabular settings, our result achieves the same order of complexity, even in multi-dimensional MDPs with continuous state-action spaces. This highlights the dimensional scalability and robustness of our bound across diverse and more complex RL environments.

Remark. The sample complexity bound for the one-sided feedback case remains valid in the full-feedback setting, corresponding to $|w + d - \|\boldsymbol{\lambda}_h\|_2| = 0$ for all $h \in [H]$. This demonstrates the robustness of POOL across both feedback regimes, highlighting its versatility and suitability for large-scale, continuous RL tasks with diverse feedback structures.

5 Experiments

We conduct a comprehensive evaluation of POOL to address the following research questions:

- RQ1: How does POOL perform compared to baseline methods?
- RQ2: How do key hyperparameters of POOL affect its performance?
- RQ3: How effective and efficient is the discretization strategy employed in POOL?

5.1 Experimental Setup

We evaluate POOL across three aspects: comparison with baseline methods, sensitivity to key hyperparameters, and discretization efficiency. Experiments are conducted on lost-sales inventory control problems using both synthetic data and real-world data from the Rossmann Sales dataset [Qin *et al.*, 2023; Gong and Simchi-Levi, 2023]. Detailed descriptions of the problem settings and synthetic data generation are provided in the Appendix.

Evaluation Metrics

Performance is quantified using the *Relative Optimality Gap*. Let c_i denote the cumulative expected cost of policy i, and c_i^* denote the benchmark cost obtained via sample average approximation with $n = 10{,}000$ samples and zero initial inventory. The relative gap is defined as:

$$\text{Relative Gap} = \frac{c_i - c_i^*}{c_i^*} \times 100\%.$$

All results report the mean and standard deviation over 10 independent runs.

Baseline Methods

We compare POOL with the following baselines:

1. NonPrivate: A non-private algorithm without added noise.
2. Input perturbation (IP): Adds Gaussian noise to the input data.
3. Output perturbation (OP): Adds Gaussian noise to the output data.

5.2 Performance Comparison (RQ1)

We evaluate POOL against IP and OP under varying privacy budgets $\rho \in \{0.1, 1, 5, 10, 20, 40\}$. Each experiment is repeated 10 times. Figure 1 reports the relative optimality gaps for both synthetic and real-world datasets. Across all settings, POOL consistently outperforms standard private baselines, which suffer from substantial performance degradation, and closely approaches the performance of the non-private algorithm. This demonstrates that our private method is significantly more effective than conventional input/output perturbation techniques while still preserving privacy.

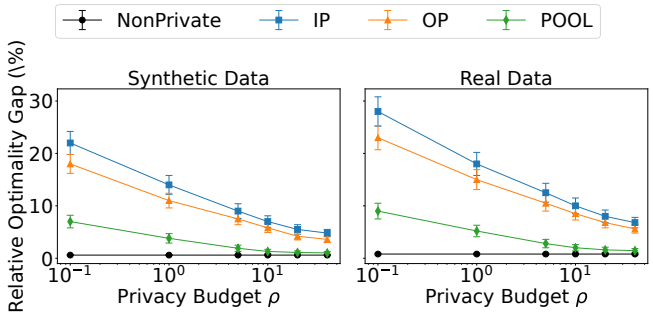

Figure 1: Relative optimality gap of POOL and baseline methods under varying privacy budgets ρ. Left: synthetic data; Right: real-world data. The x-axis is logarithmically scaled.

5.3 Hyperparameter Analysis (RQ2)

We study the effect of key hyperparameters on POOL using synthetic data: horizon length $H \in \{5, 10, 15, 20, 40\}$, feedback parameter $|\lambda| \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, state-action dimensionality $w+d \in \{2, 4, 6, 8, 16\}$, and discretization granularity $M \in \{50, 100, 200, 400, 800\}$. Each configuration is evaluated over multiple independent runs, reporting mean relative gaps and standard deviations.

Horizon length (H). Longer horizons generally lead to higher relative gaps due to increased sample complexity (Figure 2, top-left).

Feedback parameter ($|\lambda|$). One-sided feedback increases learning difficulty, whereas larger $|\lambda|$ improves performance (top-right).

State-action dimensionality ($w + d$). Higher-dimensional state-action spaces result in larger gaps, reflecting the increased complexity of multi-dimensional reinforcement learning problems (bottom-left).

Discretization granularity (M). Finer discretization reduces the relative gap, mitigating the curse of dimensionality and improving learning efficiency (bottom-right).

5.4 Effectiveness and Efficiency of Discretization (RQ3)

We compare POOL's discretization strategy against standard grid-based methods in terms of relative optimality gap and computational time. Experiments are conducted on synthetic and real-world datasets using discretization granularities $M \in \{50, 100, 200, 400, 800\}$, and running times are measured in minutes. Each experiment is repeated multiple times to account for variance.

Relative optimality gap. POOL achieves consistently lower gaps than grid-based methods for all M, indicating it effectively captures the critical structure of the state-action space and maintains near-optimal performance even with coarser discretization.

Computational efficiency. POOL substantially reduces computational cost. While grid-based approaches exhibit superlinear increases in running time with M, POOL scales more gracefully, making it suitable for large-scale inventory control problems.

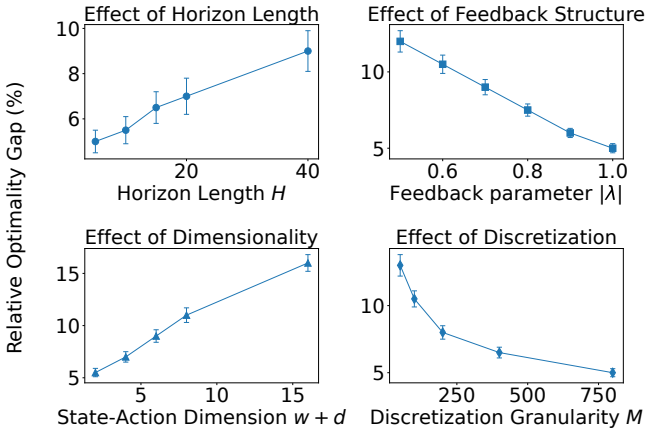

Figure 2: Impact of key hyperparameters on relative optimality gap. Top-left: Horizon length H. Top-right: Feedback parameter $|\lambda|$. Bottom-left: State-action dimensionality $w + d$. Bottom-right: Discretization granularity M. Error bars indicate standard deviations.

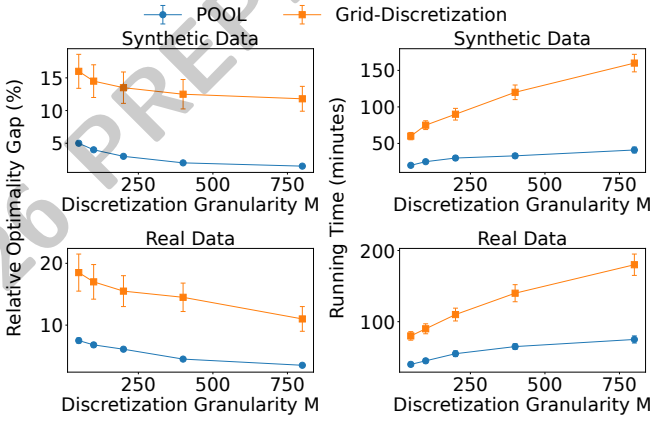

Figure 3: Comparison of POOL and grid-based discretization. Left: relative optimality gap; Right: running time. POOL consistently achieves lower gaps and faster computation across datasets.

6 Conclusion

This paper initiates the study of privacy-preserving RL in multi-dimensional decision making with one-sided feedback—an important yet underexplored problem with practical relevance in areas like business optimization and personalized recommendations. We propose POOL, a novel algorithm that ensures differential privacy while achieving strong sample efficiency. Our theoretical analysis and empirical results show that POOL effectively balances privacy and learning performance, offering a solid foundation for safe and scalable decision-making in sensitive environments. Future work includes extending the framework to online learning with evolving, partial feedback and developing contextual privacy-preserving RL methods that adapt to dynamic, feature-rich environments.

Acknowledgements

This work was supported by the DTC Singapore Innovation Grant (DTC-IGC-07) under the project Multi-Party Evaluation: Blockchain-Enabled AI Safety and Verification Framework.

References

[Antos *et al.*, 2007] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

[Bun and Steinke, 2016] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[Carrara *et al.*, 2019] Nicolas Carrara, Edouard Leurent, Romain Laroche, Tanguy Urvoy, Odalric-Ambrym Maillard, and Olivier Pietquin. Budgeted reinforcement learning in continuous state space. *Advances in neural information processing systems*, 32, 2019.

[Chen *et al.*, 2022] Xi Chen, David Simchi-Levi, and Yining Wang. Privacy-preserving dynamic personalized pricing with demand learning. *Management Science*, 68(7):4878–4898, 2022.

[Cohen *et al.*, 2020] Maxime C Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

[Cormen *et al.*, 2022] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

[Dwork *et al.*, 2014] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[Dwork, 2006] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[Feng *et al.*, 2018] Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. Learning to bid without knowing your value. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 505–522, 2018.

[Garcelon *et al.*, 2021] Evrard Garcelon, Vianney Perchet, Ciara Pike-Burke, and Matteo Pirotta. Local differential privacy for regret minimization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10561–10573, 2021.

[Gong and Simchi-Levi, 2020] Xiao-Yue Gong and David Simchi-Levi. Provably more efficient q-learning in the one-sided-feedback/full-feedback settings. *arXiv preprint arXiv:2007.00080*, 2020.

[Gong and Simchi-Levi, 2023] Xiao-Yue Gong and David Simchi-Levi. Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Management Science*, 2023.

[Guha Thakurta and Smith, 2013] Abhradeep Guha Thakurta and Adam Smith. (nearly) optimal algorithms for private online learning in full-information and bandit settings. *Advances in Neural Information Processing Systems*, 26, 2013.

[Haoyu and Wei, 2020] Zhao Haoyu and Chen Wei. Online second price auction with semi-bandit feedback under the non-stationary setting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6893–6900, 2020.

[Hossain and Lee, 2023] Md Tamjid Hossain and John WT Lee. Hiding in plain sight: Differential privacy noise exploitation for evasion-resilient localized poisoning attacks in multiagent reinforcement learning. In *2023 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 209–216. IEEE, 2023.

[Jia and Zhou, 2022] Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50, 2022.

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

[Lobel *et al.*, 2018] Ilan Lobel, Renato Paes Leme, and Adrian Vladu. Multidimensional binary search for contextual decision-making. *Operations Research*, 66(5):1346–1361, 2018.

[Munos and Bourgine, 1997] Rémi Munos and Paul Bourgine. Reinforcement learning for continuous stochastic control problems. *Advances in neural information processing systems*, 10, 1997.

[Nie, 2025] Allen Nie. *Toward Automatic Offline Reinforcement Learning With Applications in Education*. Stanford University, 2025.

[Qiao and Wang, 2023] Dan Qiao and Yu-Xiang Wang. Near-optimal differentially private reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9914–9940. PMLR, 2023.

[Qiao and Wang, 2024] Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. *Advances in Neural Information Processing Systems*, 36, 2024.

[Qin *et al.*, 2023] Hanzhang Qin, David Simchi-Levi, and Ruihao Zhu. Sailing through the dark: Provably sample-efficient inventory control. *Available at SSRN 4652347*, 2023.

[Raghu *et al.*, 2017] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163. PMLR, 2017.

[Rengarajan *et al.*, 2024] Desik Rengarajan, Nitin Ragothaman, Dileep Kalathil, and Srinivas Shakkottai. Federated ensemble-directed offline reinforcement learning. *Advances in Neural Information Processing Systems*, 37:6154–6179, 2024.

[Shariff and Sheffet, 2018] Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

[Sidford *et al.*, 2018] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

[Strang, 2022] Gilbert Strang. *Introduction to linear algebra*. SIAM, 2022.

[Van Erven and Harremos, 2014] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[Vietri *et al.*, 2020] Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, and Steven Wu. Private reinforcement learning with pac and regret guarantees. In *International Conference on Machine Learning*, pages 9754–9764. PMLR, 2020.

[Wang *et al.*, 2020] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

[Yuan *et al.*, 2021] Hao Yuan, Qi Luo, and Cong Shi. Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science*, 67(10):6089–6115, 2021.

[Zhao and Chen, 2019] Haoyu Zhao and Wei Chen. Stochastic one-sided full-information bandit. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 150–166. Springer, 2019.

[Zhao *et al.*, 2024] Hanyang Zhao, Wenpin Tang, and David Yao. Policy optimization for continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhou *et al.*, 2024] Doudou Zhou, Yufeng Zhang, Aaron Sonabend-W, Zhaoran Wang, Junwei Lu, and Tianxi Cai. Federated offline reinforcement learning. *Journal of the American Statistical Association*, 119(548):3152–3163, 2024.