Preserving Expert-Level Privacy in Offline Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

The offline reinforcement learning (RL) problem aims to learn an optimal policy from historical data collected by one or more behavioural policies (experts) by interacting with an environment. However, the individual experts may be privacy-sensitive in that the learnt policy may retain information about their precise choices. In some domains like personalized retrieval, advertising and healthcare, the expert choices are considered sensitive data. To provably protect the privacy of such experts, we propose a novel consensus-based expert-level differentially private offline RL training approach compatible with any existing offline RL algorithm. We prove rigorous differential privacy guarantees, while maintaining strong empirical performance. Unlike existing work in differentially private RL, we supplement the theory with proof-of-concept experiments on classic RL environments featuring large continuous state spaces, demonstrating substantial improvements over a natural baseline across multiple tasks.

1 INTRODUCTION

Leveraging existing offline datasets to learn high-quality decision policies via offline Reinforcement Learning (RL) is a critical requirement in many settings ranging from robotics (Fu et al., 2020; Levine et al., 2020) and recommendation systems (Bottou et al., 2013; Ie et al., 2019; Cai et al., 2017) to healthcare applications (Oberst & Sontag, 2019; Tang et al., 2022). Correspondingly, there is now a rich literature on effective techniques (Fujimoto et al., 2019b; Kumar et al., 2019; 2020; Cheng et al., 2022) for learning from such offline datasets, collected using one or more *behavioural policies*. An often overlooked aspect of this literature is, however, that in privacy sensitive domains such as personalized retrieval, advertising, and healthcare, the behavioural policies might reveal private information about the preferences / strategies used by the corresponding experts (users, advertisers, health care providers etc.) whose decisions underlie the offline data. In such scenarios, we ideally seek to uncover the broadly beneficial strategies employed in making decisions by a large cohort of experts whose demonstrations the offline data is collected with, while not leaking the private information of any given expert. We call this novel setting *offline RL with expert-level privacy*, and design algorithms with strong privacy guarantees for it.

There has been a growing interest in differentially private RL, both in online (Wang & Hegde, 2019; Qiao & Wang, 2023; Zhou, 2022; Liao et al., 2023) and offline (Qiao & Wang, 2024) settings. The offline RL literature, which is most relevant to this work, primarily focuses on protecting privacy at the *trajectory-level*. However, when each expert contributes multiple trajectories to the demonstration dataset, private information can still be leaked under trajectory-level differential privacy. Protecting the privacy of an expert in this scenario requires significantly more careful techniques. Secondly, the prior works mostly focus on tabular or linear settings (Qiao & Wang, 2023; 2024; Zhou, 2022), where the offline RL problem can be effectively solved using count or linear regression based techniques, and is hence amenable to existing techniques for privatizing counts or linear regression models. In contrast, we make no assumptions on the size of the state space, the parameterization of the learned policy or the policy underlying expert demonstrations. For a more detailed discussion of related work, we refer the reader to Section 2.

Contributions: With this context, our paper makes the following contributions:

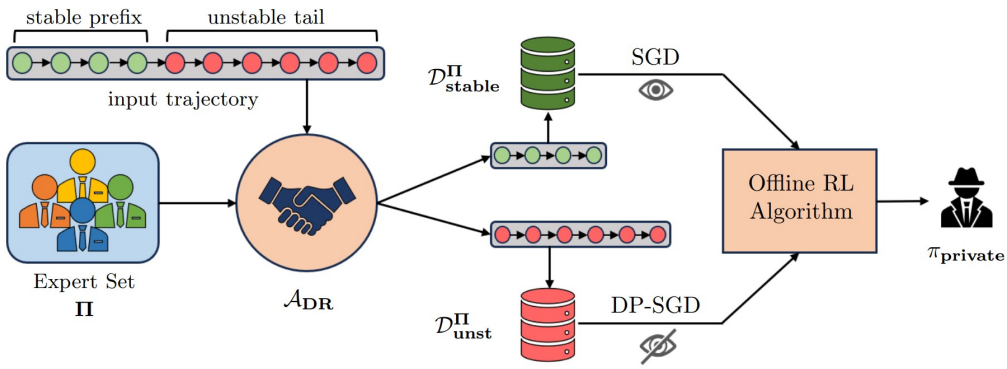


Figure 1: Training pipeline for Expert-level Differentially Private Offline RL given an expert set $\Pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$ and input trajectories logged into an offline dataset D_Π. $\mathcal{A}_{DR}$ (Algorithm 2) splits input trajectories and adds stable prefixes to D^{Π}_{stable} discarding the rest to D^{Π}_{unst}. These are used to train any off-the-shelf offline RL algorithm as described in Algorithm 3 and learn an expert-level differentially policy $\pi_{private}$.

- We identify and formalize the problem of offline RL with expert-level privacy and motivate it with practical examples. We also explain the inadequacy of prior algorithms in this setting.
- We provide *practical algorithms* for offline RL with strong expert-level privacy guarantees. This is in contrast with prior approaches which use noisy statistics of the data in offline RL (Qiao & Wang, 2024) or rely on very strict assumptions like tabular settings or linear function approximators (Wang & Hegde, 2019), limiting real-world applicability, Our approach removes these constraints, supporting any gradient-based offline RL algorithm with general function approximators and continuous state spaces.
- DP-SGD (Abadi et al., 2016) is the go-to method today for privately training gradient-based general function approximators. We adapt DP-SGD for expert-level privacy as a natural baseline in our problem setting. Our algorithm identifies a subset of the data which can be used in learning without any added noise, and significantly improves upon the DP-SGD baseline since it *selectively adds noise to gradients only where it is needed*, unlike DP-SGD which uniformly noises all gradient updates. We evaluate our algorithm on standard RL benchmarks, and discuss the results in Section 6.

Our algorithm: Our algorithm proceeds in two stages; see Figure 1. The first stage is a data filtering stage that produces a set of trajectory prefixes, which we call D^{Π}_{stable} and can be used in training without any further noise addition, in a privacy-preserving manner. The intuition for this set is that it is quite likely to occur across many experts, and hence using it in training does not violate the privacy of an individual expert. The second stage involves selectively running DP-Stochastic Gradient Descent (DP-SGD)[1] (Song et al., 2013; Bassily et al., 2014), but with parameters adapted for expert-level privacy on the remainder of the trajectories, $D^{\Pi} \setminus D^{\Pi}_{stable}$. For the first stage, we use a variant of sparse vector technique for privacy accounting (Dwork & Roth, 2014). Our empirical results strongly show that combining both stages of the algorithm described above, performs better than either of them individually. The only assumption we make on the underlying offline RL algorithm is that it is gradient-based, so that we can use DP-SGD to privatize its updates.

2 RELATED WORK

Comparison to prior work on DP Offline RL: There have been recent work exploring the intersection of differential privacy (DP) and reinforcement learning (RL) (Wang & Hegde, 2019; Qiao & Wang, 2023;

[1]DP-SGD is to an SGD based approach where gradients coming from individual experts are clipped, and appropriate Gaussian noise is added to ensure DP.

2024; Zhou, 2022; Liao et al., 2023). Qiao & Wang (2023; 2024); Zhou (2022) explore the problem either in the tabular setting or in the linear function approximation setting. In either of the cases there is a sufficient statistic (like discrete state-action visit counts or feature covariances to privatize sensitive data) which if made DP, would suffice for the whole algorithm to be DP. In our setting, these algorithms are not applicable because we operate in *continuous state space*, and with general functional approximation for the RL tasks.

Additionally, Wang & Hegde (2019) explores the use of differential privacy (DP) in continuous spaces to only protect the value function approximator. Liao et al. (2023) explores RL with linear function approximation under local differential privacy guarantee, a stronger notion of privacy where users contributing data do not trust the central aggregator. For reasons mentioned earlier, these works are not applicable to our setting.

Furthermore, to the best of our knowledge, expert-level DP has not been dealt with in the context of RL so far. Prior works often limit themselves to exploring trajectory-level privacy, which has finer granularity of privacy protection which makes the setting arguably "easier". In this work, we adopt an expert-level notion of neighbourhood when a large number of behavioural policies (or experts) are used for data collection.

Background on DP-SGD and Sparse-vector: DP-SGD first introduced in Song et al. (2013); Bassily et al. (2014), and adapted for deep learning by Abadi et al. (2016), proposed gradient noise injection to guarantee differentially private model training. Several variants of DP-SGD have been proposed such as McMahan et al. (2017), which allows user-level DP training in federated learning scenarios and Kairouz et al. (2021) a private variant of the Follow-The-Regularized-Leader (DP-FTRL) algorithm that compares favourably with amplified DP-SGD (Abadi et al., 2016; Balle et al., 2018).

Sparse-vector algorithm (Roth & Roughgarden, 2010; Hardt & Rothblum, 2010; Dwork & Roth, 2014), developed over a sequence of works, allows one to carefully track the privacy budget across queries, and often provide tighter privacy guarantee as opposed to vanilla composition Dwork & Roth (2014). At a high-level, the idea is to identify "stable queries", i.e., the queries that do not change their answers by moving to neighboring data set, and not pay the cost of privacy for answering those queries. In Algorithm 2, we use a modification of this idea to output parts of the expert trajectories that are "stable" without incurring the cost of privacy (in ε).

Relationship with the offline RL literature: As mentioned above, our approach is agnostic to the underlying offline RL algorithm used, and can be combined with off-the-shelf model-based (Kidambi et al., 2020; Yu et al., 2020; 2021) or model-free (Kumar et al., 2019; 2020; Cheng et al., 2022) offline RL algorithms. We focus on model-free approaches in our experiments.

3 PRELIMINARIES

We start by introducing the relevant background in offline RL and differential privacy, before formally defining the setting of offline RL with expert-level privacy.

3.1 Offline RL

Offline RL (Levine et al., 2020; Prudencio et al., 2023) is a data-driven approach to RL that aims to circumvent the prohibitive cost of interactive data collection in many real-world scenarios. Under this paradigm, a static dataset collected by some behaviour policy π_β is used to train another policy π without any further interaction with the environment.

Let the environment be represented by a Markov Decision Process (MDP) $\mathcal{M} = \langle S, A, r, P, \rho_0, \gamma \rangle$, where S is the state space, A is the action space, $r : S \times A \to \mathbb{R}$ is the (deterministic) reward function, $P : S \times A \to \Delta(S)$ is the transition kernel ($\Delta(\mathcal{X})$ is the set of all probability distributions over $\mathcal{X}$), $\rho_0 \in \Delta(S)$ is the initial state distribution and $\gamma \in [0, 1]$ is the discount factor.

Definition 3.1 (Offline RL)**.** Given an environment represented by the MDP $\mathcal{M} = \langle S, A, r, P, \rho_0, \gamma \rangle$ defined above and a dataset $D = \{(s_i, a_i, r_i, s'_i)_i\}_{i=1}^n$ generated with $(s_i, a_i) \sim \mu$ for some state-action distribution

μ, $r_i = r(s_i, a_i)$ and $s'_i \sim P(\cdot|s_i, a_i)$, the objective is to find an optimal policy $\pi^* : S \to \Delta(A)$, defined as

$$\pi^* = \arg\max_{\pi} \mathbb{E}^{\pi}_{s_0 \sim \rho_0}[R_\tau]$$

, where $R_\tau = \sum_{t=0}^{\infty} \gamma^t r_t$ is the discounted return of trajectory τ, generated from π and $\mathcal{M}$ as: $s_0 \sim \rho_0$, $a_t \sim \pi(.|s_t)$, $r_t = r(s_t, a_t)$ and $s_{t+1} \sim P(.|s_t, a_t)$ for $t \geq 0$.

A key concern in offline RL is that the state-action distribution encountered in the offline data to learn a policy π might significantly differ from that encountered upon actually executing π in the MDP. There is a host of existing offline RL methods that address this challenge in different ways (Kumar et al., 2019; Fujimoto et al., 2019b; Kumar et al., 2020; Cheng et al., 2022). We use BCQ (Fujimoto et al., 2019b) and CQL (Kumar et al., 2020) as our offline RL methods due to their strong performance across a variety of tasks in prior evaluations and popularity in existing offline RL literature..

We consider a scenario where the dataset contains demonstrations collected from a number of different behavioural policies or experts with varying degrees of optimality. We denote the set of behaviour policies by $\Pi = \{\pi_1, \ldots, \pi_m\}$. Note that imposing the constrain of expert-level privacy, which we aim to address next, can conflict with the rewards attainable by the learned policy. For example, if most experts are similar but a single, high-reward expert diverges from them, an inherent tradeoff arises between expert-level privacy and utility.

3.2 Differential Privacy

Differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014) quantifies the loss of privacy associated with data release from any statistical database. Informally, a procedure satisfies DP if the distribution of its outputs on two datasets which differ only in one record is very similar. We use approximate-DP (or (ε, δ)-DP) throughout our work. This notion is formally defined below.

Definition 3.2 ((ε, δ)-DP)**.** For positive real numbers $\varepsilon > 0$ and $0 \leq \delta \leq 1$, an algorithm $\mathcal{A}$ is (ε, δ)-DP iff, for any two neighbouring datasets $D, D' \in \mathcal{D}^*$ ($\mathcal{D}$ is the space of all data records), and $S \subset Range(\mathcal{A})$:

$$Pr(\mathcal{A}(D) \in S) \leq e^{\varepsilon} Pr(\mathcal{A}(D') \in S) + \delta \tag{1}$$

where $Range(\mathcal{A})$ is the set of all possible outputs of $\mathcal{A}$.

Definition 3.2 guarantees that the probability of seeing a specific output on any two neighbouring datasets can differ at most by a factor of e^{ε}, with an additional relaxation of δ. In most ML applications reasonable privacy guarantees are provided by $\varepsilon \leq 10$ and $\delta \leq \frac{1}{n}$ (Ponomareva et al., 2023), where n is the cardinality of the dataset.

One can instantiate Definition 3.2 with different of neighbourhood relations. Next, we describe the neighbourhood relations underlying our setting of *expert-level privacy* and the more commonly studied *trajectory-level privacy.*

Definition 3.3 (Expert- and Trajectory-level privacy)**.** Let $\mathcal{D}$ be the domain of all valid trajectories possible in the underlying MDP. Let $D^{\Pi} \in \mathcal{D}^*$ be a data set of trajectories generated by a set of experts Π. We say $D^{\Pi}, D^{\Pi'} \in \mathcal{D}^*$ to be *expert-level neighbours* if the set of experts $|\Pi \Delta \Pi'| = 1$, where Δ is the set difference. Alternatively, if $D, D' \in \mathcal{D}^*$ are sets of trajectories and $|D \Delta D'| = 1$, then we say D and D' are *trajectory-level neighbours.*

We note that the key difference between expert and trajectory level privacy notions, as defined above is that removing an expert π can remove up to $|D^{\pi}|$ trajectories from the overall dataset, where D^{π} is the set of trajectories contributed by π. This is a much larger change than allowed in trajectory-level privacy, and a naïve treatment would result in paying an additional cost due to group privacy Dwork & Roth (2014), scaling linearly in the number of trajectories each expert contributes. At the same time, as we argue below, via a set of settings, expert-level privacy is often a critically required notion of privacy protection. We also note that we do not include the MDP dynamics and rewards, P and r, in the private information we protect, since that would require incorporating changes to the MDP in our neighbourhood relation too.

Example (*Personalized healthcare interventions*): Many works (Saria, 2018; Gottesman et al., 2019; Tang et al., 2022) study the promise of offline RL for learning individual treatment recommendations, using existing clinical data, since real-world data interventions in the form of clinical trials are extremely expensive and limited in their power for learning personalized treatments. Here, the actions comprise of the units of medication administered to an intensive-care patient, now a well-studied setting in the case of sepsis care (Oberst & Sontag, 2019). States are outcomes of tests as well as relevant information from the patient's health record including demographic information, recent measaurements of relevant variables such as heart rate, blood pressure etc. The experts correspond to doctors or clinics who choose the observed treatment regimes in the offline samples, and we would like to learn high-performing policies without violating their privacy. Rewards correspond to long-term health outcomes such as recovery time.

As evident in the example above, expert-level privacy can reduce the barrier for the decision makers to make the decision trajectories available for learning, without worrying about leakage of proprietary information or regulatory/litigation concerns. We note that expert-level privacy differs from the privacy of a patient in the healthcare domains, or broader privacy notions also including the transitions and rewards of the MDP, since we focus on the feasibility of learning from the expert choices via offline RL.

3.3 Problem Setup

Dataset. Let Π denote a set of m privacy-sensitive policies, corresponding to m experts, $\Pi = \{\pi_1, \ldots, \pi_m\}$. $D^{\Pi} = D^{\pi_1} \cup \cdots \cup D^{\pi_m}$ is the aggregated dataset generated by the interaction of the m experts with the environment. Each D^{π_i} is a set of N trajectories $\tau = (s_1, a_2, s_2, a_2, \ldots, s_L, a_L, s_{L+1})$ of length L obtained by logging the interactions of policy π_i with the MDP $\mathcal{M} = \langle S, A, r, P, \rho_0, \gamma \rangle$, where $s_1 \sim \rho_0$, $a_h \sim \pi_i(\cdot|s_h)$ and $s_{h+1} \sim P(\cdot|s_h, a_h)$. For notational simplicity, we omit rewards throughout the paper as they do not play a role in our privatization approach and we assume they can be computed or are available as needed for a given state-action pair.

We denote by $|\tau|$ the length of the trajectory as the number of full state-action pairs it consists of. Further, we use $\tau_{h:h'} = (s_h, a_h, \ldots s_{h'}, a_{h'}, s_{h'+1})$, where $h \leq h'$, to denote the sub-trajectory of τ starting from the h-th timestep until the h'-th, including the trailing next state $s_{h'+1}$.

Goal. The goal is to learn a policy using an offline RL algorithm with expert-level DP guarantees on D^{Π}, thus ensuring that the policy learnt in this manner is not much different from a policy learnt using $D^{\Pi'}$.

We make the following mild assumptions in the design of our solution.

Assumption 3.4. The action space A is discrete.

This is a fairly mild assumption satisfied in many natural settings. The assumption arises in the first-phase of identifying stable trajectory prefixes for our algorithm, and extending this to continuous action spaces is an interesting direction for future work.

Assumption 3.5. We can query the sensitive expert policies, i.e., evaluate $\pi_i(a|s)$ for a given i, s and a.

Similar assumptions are reasonable to make in setups involving learning a student model from an ensemble of privacy-sensitive teacher models, where access to the teacher models are assumed (e.g. Papernot et al. (2022)). The privacy-sensitive experts in our setting are analogous to these privacy-sensitive teacher models. Again, query access only comes up in computing the stability of the trajectory prefix, and can likely be relaxed to using behavior-cloned versions of the experts with a slightly worse utility bound.

Assumption 3.6. Let Π_s be the class of expert policies considered by our algorithm. Hence, $\Pi \subseteq \Pi_s$. We assume access to a quantity we term *minimum action probability (p_{min})*, defined as follows,

$$p_{\min} \leq \min_{\pi \in \Pi_s} \inf_{s \in S} \min_{a \in A} \pi(a|s)$$

This value is used in the data filtering stage (first stage) of our algorithm. Note that we can always take $p_{min} = 0$. However, our algorithm is more meaningful when $p_{min} > 0$. Examples of common policy classes with $p_{min} > 0$ include ε-greedy policies and softmax policies with bounded parameters.

Most existing work in DP-RL relies on very strict assumptions like tabular settings or linear function approximation, limiting real-world applicability. Our approach removes these constraints, allowing the use of any sophisticated, gradient-based offline RL algorithm with general function approximators and continuous state spaces.

4 ALGORITHM

For a given trajectory prefix τ[2], we define the total probability of τ being generated by all the experts as: $count_\tau(\Pi) = \sum_{i=1}^{m} \left[\prod_{j=1}^{|\tau|} \pi_i(a_j|s_j)\right]$. Algorithm 1 deems τ stable if $count_\tau(\pi)$ exceeds a carefully chosen threshold, intuitively implying that enough experts are likely to generate τ. These stable prefixes may be used in training an offline RL algorithm without any noise addition. We note that this count-based filtering step is where we need Assumption 3.4.

Algorithm 1 PrefixQuery ($\mathcal{A}_{PQ}$): Tests if count of experts expected to execute a trajectory is large enough

Require: Expert policies Π, trajectory $\tau = (s_1, a_1, s_2, a_2, \dots)$, stability threshold $\hat{\theta}$, ε

Compute expected experts: $count_\tau(\Pi) \leftarrow \sum_{\pi \in \Pi} \prod_{j=1}^{|\tau|} \pi(a_j|s_j)$
if $count_\tau(\Pi) + Lap(\frac{4}{\varepsilon}) > \hat{\theta}$ **then**
    return $\top$
else
    return $\perp$

Algorithm 2 DataRelease ($\mathcal{A}_{DR}$): Releasing public dataset after privatisation

Require: Expert dataset D^Π, expert policies Π, ε_1, δ_1, min action-selection probability p_{min}, unstable query cutoff $T \leq N \cdot m$, trajectory length L

$\varepsilon' \leftarrow \frac{\varepsilon_1}{\sqrt{32T\log(2/\delta_1)}}, \quad \delta' \leftarrow \frac{\delta_1}{(2TL)}$
$c_{min} \leftarrow \frac{e^{\varepsilon'}}{(e^{\varepsilon'}-1)}, \quad \theta \leftarrow \frac{c_{min}}{p_{min}}, \quad D^\Pi_{stable} \leftarrow \{\}$
Reshuffle D^Π
for $c = 1, \dots, T$ **do**
    $\tau \leftarrow$ next trajectory from D^Π
    $\hat{\theta} \leftarrow \theta + \frac{4}{\varepsilon'}\log(1/\delta') + Lap(\frac{2}{\varepsilon'})$
    for $i = 1$ to L **do**
        $r \leftarrow \mathcal{A}_{PQ}(\Pi, \tau_{1:i}, \hat{\theta}, \varepsilon')$
        if $r = \top$ **then**
            if $i = L$ **then**
                $D^\Pi_{stable} \leftarrow D^\Pi_{stable} \cup \{\tau_{1:i}\}$
        else
            if $i \neq 1$ **then**
                $D^\Pi_{stable} \leftarrow D^\Pi_{stable} \cup \{\tau_{1:i-1}\}$
            break
return D^Π_{stable}

Algorithm 2 uses Algorithm 1 as a subroutine, traversing up to T trajectories as long as the prefix remains stable. The dataset D^Π_{stable} returned by Algorithm 2 is composed of these stable trajectory prefixes. Any trajectories after the first T for a suitably chosen parameter T are not included in D^Π_{stable} to respect the privacy budget. The discarded tails of each trajectory (including several full trajectories that were discarded) are collected in another dataset:

$$D^\Pi_{unst} = \{\tau_{k+1:L} \text{ if } \tau_{1:k} \in D^\Pi_{stable} \text{ for } k \geq 1 \text{ or } \tau, \quad \tau : D^\Pi\}$$

[2] τ hereafter can also refer to some trajectory prefix of length $h \in [1, L]$.

Consider any parameterized gradient-based offline RL algorithm $\mathcal{H}_\phi$, where ϕ denotes the parameters. Training this algorithm consists of using (s, a, r, s') tuples collected in the offline dataset to update the parameters (by splitting the trajectories into into individual transitions). Having split the dataset into stable and unstable trajectories, we can now leverage the stability of D^{Π}_{stable} to accumulate the gradient updates of $\mathcal{H}_\phi$ on transitions drawn from this dataset without any noise addition. The full algorithm, detailing the training using a mix of noisy gradient updates for transitions from D^{Π}_{unst} and noiseless gradient updates for transitions from D^{Π}_{stable} is given in Algorithm 3.

Algorithm 3 Selective DP-SGD for Offline RL

Require: Stable dataset D^{Π}_{stable}, Unstable dataset D^{Π}_{unst}, Sampling probability p, Gradient-based offline RL algorithm $\mathcal{H}_\phi$, ε_2, δ_2, N training steps, B batch size

```
Initialize ϕ_0
for n = 1 to N do
    b ~ Bernoulli(p)
    if b = 1 then
        Sample batch B ~ D^Π_unst
        ϕ_{n+1} ← DP-SGD(ϕ_n, B, ε_2, δ_2, N)
    else
        Sample batch B ~ D^Π_stable
        ϕ_{n+1} ← SGD(ϕ_n, B)
return H_{ϕ_N}
```

The sampling probability p controls the expected fraction of batches sampled from the unstable dataset D^{Π}_{unst}. The DP-SGD updates of the form DP-SGD$(\phi_n, \mathcal{B}, \varepsilon_2, \delta_2, N)$ in Algorithm 3 refer to noisy gradient updates with additive noise scaled as to ensure $(\varepsilon_2, \delta_2)$ expert-level DP. To do so we add the modifications to the standard DP-SGD algorithm (Abadi et al., 2016):

- During gradient updates, we ensure that an expert contributes *at most one transition* to a batch.
- To account for privacy amplification while sampling batches from the dataset, we use the use the fraction of experts from which we sample. Hence, the subsampling coefficient q for DP-SGD analysis is $q = p \cdot \frac{|B|}{m}$, where m is the number of experts.

A detailed description of the modified expert-level DP-SGD, which we also employ as a baseline in our empirical work in Section 6 is provided in Appendix B. The privacy analysis for the overall pipeline combining Algorithms 2 and 3 is discussed in Section 5.

Sub-optimality guarantees for the learned policy: While we show in the following section that the Algorithm 3 provides expert-level privacy, we do not provide explicit bounds on the suboptimality of the returned policy. Typical offline RL literature shows that using pessimistic approaches, the learned policy is competitive with any other policy whose state-action distribution is well-covered by the offline data distribution (Xie et al., 2021; Cheng et al., 2022). However, our training distribution undergoes two changes from the apriori offline data distribution. First is that the stable prefixes might be distributionally rather different from the entire state-action distribution, and potentially emphasize the states and actions near the start of a trajectory, as shorter prefixes are more likely to be stable. Second, the gradient clipping in DP-SGD adds an additional bias. Understanding the effect of latter with arbitrary function approximation is beyond the scope of this paper. We do, however, study the effects of distributional biases from adding varying amounts of the stable prefixes in our experiments, and find that they are typically very beneficial, particularly in tasks where the optimal behaviour does not change significantly over time.

5 PRIVACY ANALYSIS

We assume for the benefit of the analysis, that all trajectories are of the same length L (number of actions taken before the episode ends). Practically this can be replaced by the length of the longest trajectory for a worst-case privacy analysis.

Algorithm 2 can be thought of as T iterations of a combination of two algorithms $\mathcal{A}_1, \mathcal{A}_2$:

- $\mathcal{A}_1$ samples a random expert $\pi \sim \text{Unif}(\Pi)$ and a trajectory $\tau \sim \pi$ and converts it into L queries, $Q^\tau = \{count_{\tau_1}(.), ..., count_{\tau_L}(.)\}$.
- $\mathcal{A}_2$ is invoked on Π and the query set Q^τ, running an iteration of the Sparse Vector Technique (Dwork & Roth (2014)). $\mathcal{A}_2$ treats Π as the privacy-sensitive database, Q^τ as the query sequence, and $\hat{\theta}$ as the threshold for the Sparse Vector Algorithm. $\mathcal{A}_2(\Pi, Q^\tau)$ outputs $r \in \{\top, \bot\}^{k+1}$ such that $r[i] = \top \ \forall \ i \leq k$ and $r[k+1] = \bot$, where $k \leq L$ is the length of the stable prefix. Note that $r[i] = \mathcal{A}_{PQ}(\Pi, \tau_{1:i}, \hat{\theta}, \varepsilon')$.
- $\mathcal{A}_1$, upon receiving r, releases the prefix $\tau_{1:k} = (s_1, a_1, \ldots, s_k, a_k, s_{k+1})$.

For a fixed stream of queries Q^τ, Algorithm $\mathcal{A}_2$ is $(\varepsilon', 0)$-DP since it just invokes one round of the Sparse Vector Algorithm (Dwork & Roth (2014)). Note that the trajectories are just used for creating the queries for SVT, and the set of expert policies Π is the privacy database. For a given trajectory τ, $\mathcal{A}_2$ identifies the longest prefix which is stable, or in other words, has enough experts which are likely to sample this prefix. Note that, the privacy cost ε' for identifying this prefix is independent of the length of the trajectory.

However, Q^τ itself is a function of the set of experts Π. Despite this, in the following analysis, we show that $\mathcal{A} = \mathcal{A}_1 \circ \mathcal{A}_2$ is also differentially private (due to the carefully chosen threshold in Algorithm 2 Line 6). We then compose the privacy guarantees over T trajectories, since $\mathcal{A}_{DR}$ is simply $\mathcal{A}$ applied T times.

We now state two key lemmas for proving that $\mathcal{A}$ is differentially private. For a given trajectory prefix τ_k of length k, let E_τ denote the event that $\mathcal{A}_1$ samples any trajectory τ' containing τ_k as a prefix; that is, $\tau'_{1:k} = \tau_k$.

Lemma 5.1. *For a trajectory prefix τ_k, and neighbouring expert sets Π, Π', if $count_{\tau_k}(\Pi) \geq c_{min}$ then,*

$$\Pr(E_{\tau_k}|\Pi) \leq e^{\varepsilon'} \Pr(E_{\tau_k}|\Pi') \tag{2}$$

where c_{min} and ε' are as defined in Algorithm 2.

This lemma states that if the probability of seeing a trajectory prefix τ_k across all experts is large enough, then the probability that $\mathcal{A}_1$ samples any trajectory with prefix τ_k is stable under changing one of the experts. The full proof is available in Appendix A.

Lemma 5.2. *For any trajectory prefix τ_k, and expert set Π such that $count_{\tau_k}(\Pi) < \theta$ the following holds:*

$$\Pr(\mathcal{A}_{PQ}(\Pi, \tau_k, \hat{\theta}, \varepsilon') = \top) < \delta' \tag{3}$$

where $\varepsilon', \delta', \theta, \hat{\theta}$ are as defined in Algorithm 2.

Informally, this lemma states that if the total probability of seeing a trajectory prefix τ_k across all experts is *not* large enough, then the probability that it is labelled as stable by Algorithm 1 is less than δ'. The full proof is described in Appendix A.

Theorem 5.3. *Under Assumption 3.4, $\mathcal{A}$ is $(2\varepsilon', \frac{\delta_1}{2T})$-DP, where ε', δ_1 are defined in Algorithm 2.*

Proof Sketch: The main result used to prove this theorem is the following:

$$\Pr(\mathcal{A}(\Pi) = \tau) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau) + \Pr(E_\tau|\Pi) \cdot \delta' \tag{4}$$

which is obtained from Lemma 5.4 that we give below.

Let $\mathcal{T}$ denote the set of all possible trajectory prefixes using S, A. To prove that $\mathcal{A}$ is $(2\varepsilon', \delta_1/(2T))$-differentially private, we need to prove the following for any pair of neighbouring $\Pi, \Pi' \subseteq \Pi_{p_{min}}$:

$$\sum_{\tau \in \mathcal{T}} \max\{\Pr(\mathcal{A}(\Pi) = \tau) - e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau), 0\} \leq \frac{\delta_1}{2T}.$$

The above can be proved using equation 4, and we defer the full proof to Appendix A.

The following lemma gives the key result used to prove Theorem 5.3.

Lemma 5.4. *Fix a trajectory prefix τ_k and a pair of neighboring expert sets $\Pi, \Pi' \subseteq \Pi_{p_{min}}$. If $count_{\tau_k}(\Pi) < c_{min}$ then,*

$$\Pr(\mathcal{A}(\Pi) = \tau_k) < \Pr(E_{\tau_k}|\Pi) \cdot \delta'. \tag{5}$$

If $count_{\tau_k}(\Pi) \geq c_{min}$ then,

$$\Pr(\mathcal{A}(\Pi) = \tau_k) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau_k) + \Pr(E_{\tau_k}|\Pi) \cdot \delta' \tag{6}$$

Proof Sketch: The proof of this lemma follows from Lemmas 5.1 and 5.2. Let $r = \top \cdots \top \bot \in \{\top, \bot\}^{k+1}$ and $\gamma^{(L)} = \top^L$. When $k < L$ we have

$$\Pr(\mathcal{A}(\Pi) = \tau_k) = \sum_{a \in A} [\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a}|\Pi)]$$

where $\tau_k \cdot a$ denotes the trajectory prefix τ_k followed by action a. And when $k = L$,

$$\Pr(\mathcal{A}(\Pi) = \tau_k) = \Pr(\mathcal{A}_2(\Pi, Q_k^{\tau}) = \gamma^{(L)}) \Pr(E_{\tau_k}|\Pi)$$

Consider the case $k = L$. The main idea behind this proof is that if $count_{\tau_k}(\Pi)$ is small enough, then $\Pr(\mathcal{A}_2(\Pi, Q_k^{\tau}) = \gamma^{(L)})$ is less than δ' (Lemma 5.2) and we get equation 5. On the other hand, if $count_{\tau_k}(\Pi)$ is greater than c_{min}, then $\Pr(E_{\tau_k}|\Pi)$ is bounded by Lemma 5.1. Using that $\mathcal{A}_2$ is always ε'-DP when the queries are fixed we get equation 6. The case of $k < L$ is less straightforward, but can also be analyzed by considering two cases for each $a \in A$, $count_{\tau_k \cdot a}(\Pi) \geq c_{min}$ or $count_{\tau_k \cdot a}(\Pi) < c_{min}$. The full proof is present in Appendix A. □

Theorem 5.5. *Under Assumption 3.4, Algorithm 2 is $(\varepsilon_1, \delta_1)$-DP.*

Proof. Algorithm 2 simply executes Algorithm $\mathcal{A}$, T times. Since $\mathcal{A}$ is $(\varepsilon_1/\sqrt{8T\log(2/\delta_1)}, \delta_1/(2T))$-DP as shown in Theorem 5.3, using advanced composition (Corollary 3.21 in Dwork & Roth (2014)), Algorithm 2 can be shown to be $(\varepsilon_1, \delta_1)$-DP. □

Using all the above, we derive the following expert-level privacy guarantee for the overall training pipeline.

Theorem 5.6. *Algorithm 3 is (ε, δ)-DP, where:*

$$\varepsilon = \varepsilon_1 + \varepsilon_2 \quad \textit{and} \quad \delta = \delta_1 + \delta_2 \tag{7}$$

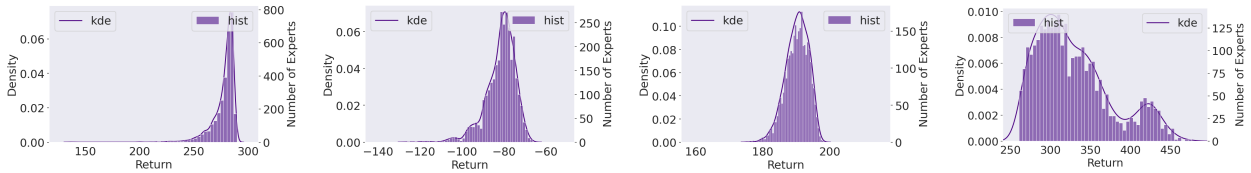
Proof. Shown trivially by sequential composition of privacy guarantees for the release of D^{Π}_{stable} and DP-SGD training on samples from D^{Π}_{unst}. □

Note on novelty: We have presented a novel, useful and non-trivial extension of the Sparse Vector Technique to preserve expert-level DP in Offline RL. The stream of queries in vanilla SVT, (Algorithm 1 Section 3.6, Dwork & Roth (2014)), is not dependent on the private database. On the other hand, in our setup, the analogous "stream of queries" is a function of the set of the experts (which is the private database to be protected), and so directly applying SVT does not ensure privacy. However, we show in our analysis that despite the query stream being a function of the private database, SVT can work if the threshold (line 6 in Algorithm 2) is chosen in a certain way. We use the probability distribution of the query stream derived from the experts' policies to prove this.

6 EXPERIMENTS

Ensuring reasonable levels of privacy requires significant amounts of hyperparameter tuning and a large number of data points, and typically comes with a non-trivial performance impact[3]. Intuitively, with the demand for expert-level privacy in our setting, we risk severe degradation of utility since each "record" (each expert's data) forms a much larger fraction of the overall dataset. In our experiments, we show that our method is able to recover much of this performance loss, beating DP-SGD by significant margins.

6.1 Dataset Generation



(a) LunarLander (b) Acrobot (c) CartPole (d) HIV Treatment

Figure 2: Expert return histograms, with KDE (Kernel Density Estimation), for experts trained on heterogeneous LunarLander, Acrobot, CartPole and HIV Treatment environments. We use a mix of experts that shows a wide range of performance on the test environment. For HIV Treatment, the return values are normalised by 10^6.

We train 3000 experts each on the Cartpole, Acrobot, LunarLander and HIV Treatment environments. Cartpole, Acrobot and LunarLander are from the Gymnasium package (Towers et al., 2023). The HIV Treatment simulator is based on the implementation by Geramifard et al. (2015) of the model described in Ernst et al. (2006). Experts are trained on variations of the default environment, created by modifying the values of key parameters (e.g. gravity in LunarLander environment). The trained experts are then used to generate demonstrations on a default setting of the environment, collectively forming the aggregated dataset for offline RL. This training methodology ensures that the experts show varied (often sub-optimal) behaviours on the default environment. As shown in Figure 2, the experts' return-histograms indicate that the mix of experts we create shows significant diversity in the policies learnt. Further details regarding the environment settings used to train the expert policies are available in Appendix B.

6.2 Training Setup

Our algorithm allows us to use any off-the-shelf gradient-based offline RL algorithm, and we demonstrate this by experimenting with Conservative Q-Learning (CQL) (Kumar et al., 2020) and Batch Constrained Deep Q-Learning (BCQ) (Fujimoto et al., 2019b). More specifically, we use the DQN (Deep Q-Network) version of CQL, and the discrete-action version of BCQ (Fujimoto et al., 2019a), since we operate under the assumption of discrete action spaces. To represent each of the Q-value functions in both the algorithms and the generative model in BCQ, we use a neural network with 2 hidden layers of 256 units each.

During training, we perform a grid search to find the optimal set of hyper-parameters (learning rate, batch size, sampling probability p, DP-SGD noise) for each environment; see Appendix B.

The maximum trajectory length is fixed to 200 for Acrobot and CartPole in the collected demonstrations. While evaluating the final learned policy, the maximum episode length is again set as 200 for Acrobot, we increase this to 1000 for Cartpole for increased difficulty in an otherwise "simple" environment. We allow all DP-SGD training runs to progress until the privacy budget is used up completely.

[3]In supervised learning, for example, the accuracy on ImageNet training can go from $> 75\%$ to $\sim 47.9\%$ in the private case (Kurakin et al., 2022), under example-level privacy.

6.3 Baselines

We compare our method against DP-SGD adapted for expert-level privacy. A brief discussion of the changes we made are presented in Section 4. A full discussion, along with analysis, is presented in the appendix for reference. We keep the underlying RL algorithm same for fair comparison. In the absence of prior methods for the expert-level privacy setting, we limit our empirical studies to this naive baseline.

We also train a non-private policy, which ignores privacy concerns and trains over the whole dataset using vanilla CQL/BCQ. This may be taken as an empirical upper-bound on the performance achievable in the presence of privacy concerns. In our experiments, we report episodic returns, normalized to lie between 0 (random policy) and 1 (optimal policy) averaged over 10 evaluation runs at the end of training, as a fraction of the return of the non-private baseline.

6.4 Results

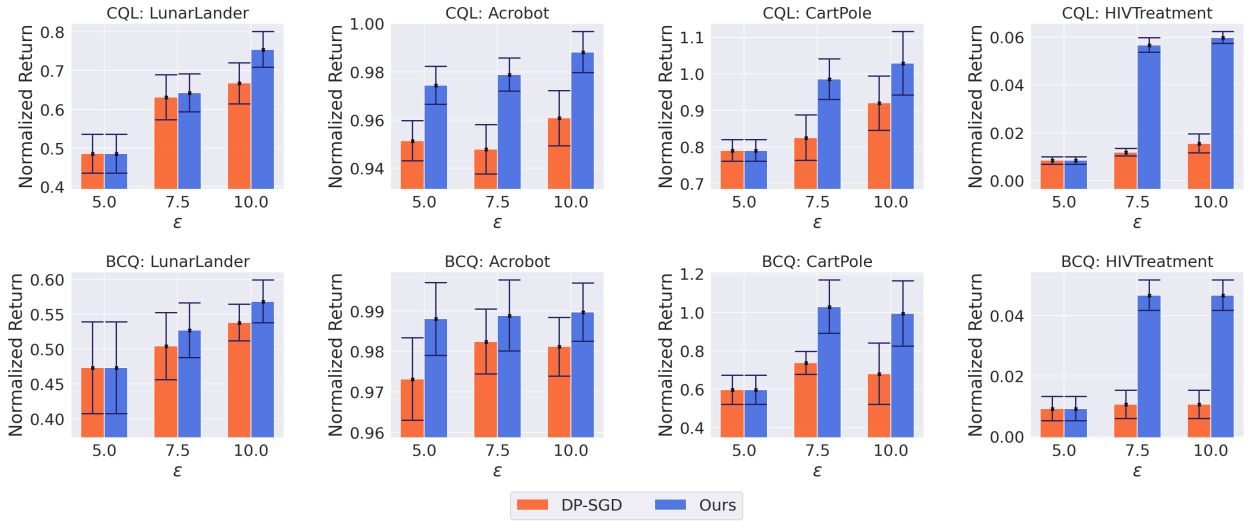


Figure 3: Performance of our method and DP-SGD for different values of ε with $\delta = 1/m$, where m is the number of experts. We report the episodic return, normalized to lie between 0 (random policy) and 1 (optimal policy) averaged over 10 evaluation runs at the end of training, as a fraction of the non-private baseline, along with 95% confidence intervals of the mean. Our method consistently outperforms DP-SGD, especially in the high ε regions.

Our experiments show that our algorithm, which leverages expert consensus to select stable trajectory prefixes prior to offline RL training, shows superior performance to offline RL training by merely using an adapted expert-level DP-SGD algorithm. Figure 3 compares the performance of our method with DP-SGD across different values of ε. All the results are averaged over 3 runs and 95% confidence intervals are also reported. Our method is able to recover much of the performance loss due to DP-SGD, especially in the high ε-regions.

We provide the values of sampling probability that led to the best performance for each environment in Table 1. Interestingly, for Acrobot (with CQL and BCQ) and HIV Treatment (with CQL) environments, just D^{Π}_{stable} was enough to get the optimal performance in high ε-region and we did not need to pay any privacy cost of DP-SGD during model training.

We also evaluate the effect of varying the number of experts on the performance of our algorithm relative to DP-SGD, in figure 4, on the Acrobot environment. The phenomenon of increasing performance with number of experts can be explained as follows. As the number of experts increases, the likelihood that the value of $count_{\tau}(\Pi)$ exceeds the noisy threshold $\hat{\theta}$ also increases, and with it the number and length of trajectory prefixes in the stable dataset D^{Π}_{stable}. Noiseless training on the stable datasets boosts our

Environment	CQL			BCQ		
	$\varepsilon = 5$	$\varepsilon = 7.5$	$\varepsilon = 10$	$\varepsilon = 5$	$\varepsilon = 7.5$	$\varepsilon = 10$
LunarLander	1.0	0.8	0.8	1.0	0.5	0.5
Acrobot	0.0	0.0	0.0	0.0	0.0	0.0
CartPole	1.0	0.9	0.8	1.0	0.5	0.5
HIV Treatment	1.0	0.0	0.0	1.0	0.9	0.9

Table 1: Best choices for probability of sampling from $\mathcal{D}^{\Pi}_{unstable}$ ie. p for Algorithm 3 on test environments

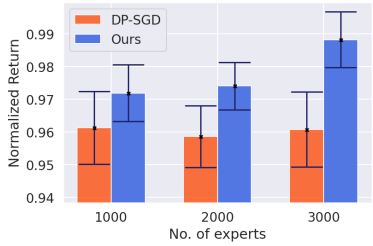


Figure 4: Performance of our method and DP-SGD for $m = 1000, 2000$ and 3000 experts, on the Acrobot environment with $\varepsilon = 10.0$ and CQL as the underlying offline RL algorithm. We observe increasing improvements of our method over DP-SGD as m increases.

algorithm's performance, and the difference scales with the number of experts. For very low values of m, our algorithm will yield an empty dataset and will be functionally identical to expert-level DP-SGD.

Note that our method degenerates to DP-SGD if we take $\varepsilon_1 = 0, \varepsilon_2 = \varepsilon$ for any given ε, which is what we do in the lower ε-region ($\varepsilon = 5$ for Lunar Lander, CartPole and HIVTreatment) where the privacy budget is not enough to get sufficient stable data for training. For higher values of ε, we take $\varepsilon_1 = \frac{3\varepsilon}{4}, \varepsilon_2 = \frac{\varepsilon}{4}$ for LunarLander and CartPole. For Acrobot and HIVTreatment, we set $\varepsilon_1 = \varepsilon, \varepsilon_2 = 0$, that is, we only train on D^{Π}_{stable}. Similarly, we split δ as $\delta_1 = \frac{9\delta}{10}, \delta_2 = \frac{\delta}{10}$.

7 CONCLUSION AND FUTURE WORK

In this work, we initiate the study of offline RL with expert-level privacy. We describe the novel challenges of the setting, and provide theoretically sound algorithms which are evaluated in proof-of-concept evaluation. In future work, we would like to relax Assumptions 3.4-3.6, which limit the applicability of our results. We expect replacing the queries to the experts with behaviour cloned versions to be less challenging than eliminating the other assumptions. Understanding the utility implications of Algorithm 3, despite the influence of our privatization scheme on the training distribution is another important avenue for future work. Finally, while our empirical evaluation convincingly establishes the promise of this approach, we leave a larger evaluation across diverse and larger-scale RL benchmarks for future studies.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL `http://dx.doi.org/10.1145/2976749.2978318`.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Cur-

ran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3b5020bb891119b9f5130f1fea9bd773-Paper.pdf.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pp. 464–473, USA, 2014. IEEE Computer Society. ISBN 9781479965175. doi: 10.1109/FOCS.2014.56. URL https://doi.org/10.1109/FOCS.2014.56.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the tenth ACM international conference on web search and data mining*, pp. 661–670, 2017.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 667–672. IEEE, 2006.

Justin Fu, Aviral Kumar, Ofir Nachum, G. Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *ArXiv*, abs/2004.07219, 2020. URL https://api.semanticscholar.org/CorpusID:215827910.

Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708*, 2019a.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Alborz Geramifard, Christoph Dann, Robert H Klein, William Dabney, and Jonathan P How. Rlpy: a value-function-based reinforcement learning framework for education and research. *J. Mach. Learn. Res.*, 16(1):1573–1578, 2015.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st annual symposium on foundations of computer science*, pp. 61–70. IEEE, 2010.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, et al. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. *arXiv preprint arXiv:1905.12767*, 2019.

Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5213–5225. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kairouz21b.html.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c2073ffa77b5357a498057413bb09d3a-Paper.pdf.

Aviral Kumar, Aurick Zhou, G. Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *ArXiv*, abs/2006.04779, 2020. URL https://api.semanticscholar.org/CorpusID:219530894.

Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, A. Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *ArXiv*, abs/2201.12328, 2022. URL https://api.semanticscholar.org/CorpusID:246411162.

Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020. URL https://api.semanticscholar.org/CorpusID:218486979.

Chonghua Liao, Jiafan He, and Quanquan Gu. Locally differentially private reinforcement learning for linear mixture markov decision processes. In Emtiyaz Khan and Mehmet Gonen (eds.), *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pp. 627–642. PMLR, 12–14 Dec 2023. URL https://proceedings.mlr.press/v189/liao23a.html.

H. B. McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:3461939.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890. PMLR, 2019.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2022.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, July 2023. ISSN 1076-9757. doi: 10.1613/jair.1.14649. URL http://dx.doi.org/10.1613/jair.1.14649.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–0, 2023. ISSN 2162-2388. doi: 10.1109/tnnls.2023.3250269. URL http://dx.doi.org/10.1109/TNNLS.2023.3250269.

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

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

Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 765–774, 2010.

Suchi Saria. Individualized sepsis treatment using reinforcement learning. *Nature medicine*, 24(11):1641–1642, 2018.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings*, pp. 245–248, 2013. ISBN 9781479902484. doi: 10.1109/GlobalSIP.2013.6736861.

Shengpu Tang, Maggie Makar, Michael Sjoding, Finale Doshi-Velez, and Jenna Wiens. Leveraging factored action spaces for efficient offline reinforcement learning in healthcare. *Advances in Neural Information Processing Systems*, 35:34272–34286, 2022.

Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023. URL https://zenodo.org/record/8127025.

Baoxiang Wang and Nidhi Hegde. Privacy-preserving q-learning with functional noise in continuous spaces. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6646b06b90bd13dabc11ddba01270d23-Paper.pdf.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

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

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.

Xingyu Zhou. Differentially private reinforcement learning with linear function approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–27, 2022.

A Theorem Proofs

A.1 Proof of Lemma 5.1

Lemma 5.1. *For a trajectory prefix τ_k, and neighbouring expert sets Π, Π', if $count_{\tau_k}(\Pi) \geq c_{min}$ then,*

$$\Pr(E_{\tau_k}|\Pi) \leq e^{\varepsilon'} \Pr(E_{\tau_k}|\Pi') \tag{2}$$

where c_{min} and ε' are as defined in Algorithm 2.

Proof. Recall E_{τ_k} is the event a trajectory with prefix τ_k is generated. Assuming $\tau_k = s_1, a_1, s_2, \ldots, a_k, s_{k+1}$, the probability that a fixed expert π generates such a prefix when interacting with the MDP is $\rho_0(s_1)\left(\Pi_{j=1}^{k}\pi(a_j|s_j)P(s_{j+1}|s_j,a_j)\right)$. Since the expert is chosen uniformly at random, we have:

$$\Pr(E_{\tau_k}|\Pi) = \frac{1}{m}\sum_{i=1}^{m}\left[\rho_0(s_1)\left(\Pi_{j=1}^{k}\pi_i(a_j|s_j)P(s_{j+1}|s_j,a_j)\right)\right] \tag{8}$$

$$= \frac{1}{m}\rho_0(s_1)\left(\Pi_{j=1}^{k}P(s_{j+1}|s_j,a_j)\right)\sum_{i=1}^{m}\left(\Pi_{j=1}^{k}\pi_i(a_j|s_j)\right) \tag{9}$$

$$= \frac{1}{m}\rho_0(s_1)\left(\Pi_{j=1}^{k}P(s_{j+1}|s_j,a_j)\right) count_{\tau_k}(\Pi) \ . \tag{10}$$

Similarly,

$$\Pr(E_{\tau_k}|\Pi') = \frac{1}{m'}\rho_0(s_1)\left(\Pi_{j=1}^{k}P(s_{j+1}|s_j,a_j)\right) count_{\tau_k}(\Pi') \ . \tag{11}$$

Where $m' = |\Pi'|$. Since $|\Pi \Delta \Pi'| = 1$,

$$|count_{\tau_k}(\Pi) - count_{\tau_k}(\Pi')| \leq 1 \quad \text{and} \quad |m - m'| = 1 \ . \tag{12}$$

This gives us the following,

$$\frac{\Pr(E_{\tau_k}|\Pi)}{\Pr(E_{\tau_k}|\Pi')} = \frac{count_{\tau_k}(\Pi) \times m'}{count_{\tau_k}(\Pi') \times m} \tag{13}$$

$$\leq \max\left(\frac{count_{\tau_k}(\Pi)(m+1)}{(count_{\tau_k}(\Pi)+1)m}, \frac{count_{\tau_k}(\Pi)(m-1)}{(count_{\tau_k}(\Pi)-1)m}, \frac{m+1}{m}, \frac{m-1}{m}\right) \tag{14}$$

$$= \max\left(\frac{count_{\tau_k}(\Pi)(m-1)}{(count_{\tau_k}(\Pi)-1)m}, \frac{m+1}{m}\right) \ . \tag{15}$$

Note that, since $m + 1 > count_{\tau_k}(\Pi)$,

$$\frac{m+1}{m} < \frac{count_{\tau_k}(\Pi)}{count_{\tau_k}(\Pi) - 1} \tag{16}$$

$$\max\left(\frac{count_{\tau_k}(\Pi)(m-1)}{(count_{\tau_k}(\Pi)-1)m}, \frac{m+1}{m}\right) < \frac{count_{\tau_k}(\Pi)}{count_{\tau_k}(\Pi) - 1} \ . \tag{17}$$

This allows us to bound the ratio of probabilities of E_{τ_k} for Π and Π' as

$$\frac{\Pr(E_{\tau_k}|\Pi)}{\Pr(E_{\tau_k}|\Pi')} < \frac{count_{\tau_k}(\Pi)}{count_{\tau_k}(\Pi) - 1} \leq e^{\varepsilon'} \ , \tag{18}$$

where the last inequality uses $count_{\tau}(\Pi) \geq c_{min} = e^{\varepsilon'}/(e^{\varepsilon'} - 1)$. □

A.2 Proof of Lemma 5.2

Lemma 5.2. *For any trajectory prefix τ_k, and expert set Π such that $count_{\tau_k}(\Pi) < \theta$ the following holds:*

$$\Pr(\mathcal{A}_{PQ}(\Pi, \tau_k, \hat{\theta}, \varepsilon') = \top) < \delta' \tag{3}$$

where $\varepsilon', \delta', \theta, \hat{\theta}$ are as defined in Algorithm 2.

Proof.

$$\begin{aligned}
\Pr(\mathcal{A}_{PQ}(\Pi, \tau, \hat{\theta}, \varepsilon') = \top) &= \Pr\left[count_{\tau_k}(\Pi) + Lap(4/\varepsilon') > \theta + (4/\varepsilon')\log(1/\delta') + Lap(2/\varepsilon')\right] && (19)\\
&< \Pr\left[\theta + Lap(4/\varepsilon') > \theta + (4/\varepsilon')\log(1/\delta') + Lap(2/\varepsilon')\right] && (20)\\
&= \Pr\left[Lap(4/\varepsilon') > (4/\varepsilon')\log(1/\delta') + Lap(2/\varepsilon')\right] && (21)
\end{aligned}$$

Consider $u \sim Lap(4/\varepsilon')$ and $v \sim Lap(2/\varepsilon')$. For a fixed v, we have,

$$\Pr\left[u \geq (4/\varepsilon')\log(1/\delta') + v\right] \leq \frac{\delta'}{2} e^{-v(\varepsilon'/4)} \tag{22}$$

[Using Laplace Tail Bound]

Accounting for the randomness of v, from equation 21, equation 22 and the pdf of the Laplace distribution, we have,

$$\begin{aligned}
\Pr\left[Lap(4/\varepsilon') > (4/\varepsilon')\log(1/\delta') + Lap(2/\varepsilon')\right] &\leq \int_{-\infty}^{\infty} \left(\frac{\delta'}{2} e^{-v(\varepsilon'/4)} \times \frac{\varepsilon'}{4} e^{-|v|(\varepsilon'/2)}\right) dv && (23)\\
&< \delta' && (24)
\end{aligned}$$

This gives the statement of the lemma.

□

A.3 Proof of Lemma 5.4

Lemma 5.4. *Fix a trajectory prefix τ_k and a pair of neighboring expert sets $\Pi, \Pi' \subseteq \Pi_{p_{min}}$. If $count_{\tau_k}(\Pi) < c_{min}$ then,*

$$\Pr(\mathcal{A}(\Pi) = \tau_k) < \Pr(E_{\tau_k}|\Pi) \cdot \delta'. \tag{5}$$

If $count_{\tau_k}(\Pi) \geq c_{min}$ then,

$$\Pr(\mathcal{A}(\Pi) = \tau_k) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau_k) + \Pr(E_{\tau_k}|\Pi) \cdot \delta' \tag{6}$$

Proof. We split this proof across 3 lemmas. Lemma A.1 proves this bound for the case $count_{\tau_k}(\Pi) < c_{min}$ and $k < L$. Lemma A.2 handles the case $count_{\tau_k}(\Pi) \geq c_{min}$ and $k < L$. Finally, Lemma A.3 proves this for $k = L$. Combining these 3 gives the statement of Lemma 5.4. □

Lemma A.1. *For any trajectory prefix $\tau_k, k < L$, and neighboring expert sets $\Pi, \Pi' \subseteq \Pi_{p_{min}}$, if $count_{\tau_k}(\Pi) < c_{min}$ then,*

$$\Pr(\mathcal{A}(\Pi) = \tau_k) < \Pr(E_{\tau_k}|\Pi)\delta'$$

where δ' is as defined in Algorithm 2.

Proof. Let $r = \top \cdots \top \bot \in \{\top, \bot\}^{k+1}$

$$\Pr(\mathcal{A}(\Pi) = \tau_k) = \sum_{a \in A} \left[\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a}|\Pi)\right] \tag{25}$$

where $\tau_k \cdot a$ denotes the trajectory prefix τ_k followed by action a.

$\mathcal{A}_2$ runs the Sparse Vector Algorithm using the the query sequence $Q^{\tau_k \cdot a}$ to output a sequence of the form $\top \cdots \top \bot$. Here, each $\top$ or $\bot$ is an output of $\mathcal{A}_{PQ}$ (Algorithm 1). The probability that $\mathcal{A}_2$ outputs the sequence r can be upper-bounded by the probability of getting $\top$ on the kth call to $\mathcal{A}_{PQ}$. Formally,

$$\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \leq \Pr(\mathcal{A}_{PQ}(\Pi, \tau_k, \hat{\theta}, \varepsilon') = \top) \tag{26}$$

$\forall a$, Using $count_\tau(\Pi) < c_{min} < \theta$, (since $\theta = c_{min}/p_{min}$), Lemma 5.2, and equation 26,

$$\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \leq \delta' \tag{27}$$

Combining equation 25 and equation 27, we get,

$$\Pr(\mathcal{A}(\Pi) = \tau_k) < \delta' \sum_{a \in A} \Pr(E_{\tau_k \cdot a}|\Pi) = \delta' \Pr(E_{\tau_k}|\Pi) \tag{28}$$

□

Lemma A.2. *For any trajectory prefix $\tau_k = (s_1, a_1, ..., s_k, a_k, s_{k+1}), k < L$, and neighboring expert sets $\Pi, \Pi' \subseteq \Pi_{p_{min}}$, if $count_{\tau_k}(\Pi) \geq c_{min}$ then,*

$$\Pr(\mathcal{A}(\Pi) = \tau_k) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau_k) + \Pr(E_{\tau_k}|\Pi)\delta'$$

where ε', δ' are as defined in Algorithm 2.

Proof. Let $r = \top \cdots \top \bot \in \{\top, \bot\}^{k+1}$.

$$\Pr(\mathcal{A}(\Pi) = \tau) = \sum_{a \in A} \left[\Pr(\mathcal{A}_2(\Pi, Q^{\tau \cdot a}) = r^{k+1}) \Pr(E_{\tau \cdot a}|\Pi)\right] \tag{29}$$

Let $A' = \{a : a \in A, count_{\tau_k \cdot a}(\Pi) \geq c_{min}\}, A'' = A \setminus A'$.
$\forall a \in A''$,

$$count_{\tau_k \cdot a}(\Pi) = \sum_{i=1}^{m} [(\Pi_{j=1}^{k} \pi_i(a_j|s_j)) \times \pi_i(a|s_{k+1})] \geq p_{min} \times \sum_{i=1}^{m} [\Pi_{j=1}^{k} \pi_i(a_j|s_j)] = p_{min} \times count_{\tau_k}(\Pi) \tag{30}$$

$$p_{min} \times count_{\tau_k}(\Pi) \leq count_{\tau_k \cdot a}(\Pi) < c_{min} \tag{31}$$

$$\Rightarrow count_{\tau_k}(\Pi) < \frac{c_{min}}{p_{min}} = \theta \tag{32}$$

The output of $\mathcal{A}_2$ is the output of a sequence of calls to $\mathcal{A}_{PQ}$ (Algorithm 1). The probability that $\mathcal{A}_2$ outputs the sequence r can be upper-bounded by the probability of getting $\top$ on the kth call to $\mathcal{A}_{PQ}$. Using this fact along with equation 32 and Lemma 5.2, we can say,

$$\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \leq \Pr(\mathcal{A}_{PQ}(\Pi, \tau_k, \hat{\theta}, \varepsilon') = \top) < \delta' \tag{33}$$

$$\Pr(\mathcal{A}(\Pi) = \tau_k) = \sum_{a \in A} [\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a} | \Pi)] \tag{34}$$

$$= \sum_{a \in A'} [\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a} | \Pi)] + \sum_{a \in A''} [\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a} | \Pi)] \tag{35}$$

$$= \sum_{a \in A'} [\Pr(\mathcal{A}_2(\Pi, Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a} | \Pi)] + \delta' \sum_{a \in A''} \Pr(E_{\tau_k \cdot a} | \Pi) \qquad \text{[Using equation 33]} \tag{36}$$

$$\leq e^{2\varepsilon'} \sum_{a \in A'} [\Pr(\mathcal{A}_2(\Pi', Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a} | \Pi')] + \delta' \Pr(E_{\tau_k} | \Pi) \qquad \text{[Using: } \mathcal{A}_2 \text{ is } (\varepsilon', 0)\text{-DP and Lemma 5.1]} \tag{37}$$

$$\leq e^{2\varepsilon'} \sum_{a \in A} [\Pr(\mathcal{A}_2(\Pi', Q^{\tau_k \cdot a}) = r) \Pr(E_{\tau_k \cdot a} | \Pi')] + \delta' \Pr(E_{\tau_k} | \Pi) \tag{38}$$

$$\Pr(\mathcal{A}(\Pi) = \tau_k) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau_k) + \Pr(E_{\tau_k} | \Pi) \delta' \tag{39}$$

□

Lemma A.3. *For any trajectory prefix τ_L, where L is the maximum length of a trajectory, and a pair of neighboring expert sets $\Pi, \Pi' \subseteq \Pi_{p_{min}}$, if $count_{\tau_L}(\Pi) < c_{min}$ then,*

$$\Pr(\mathcal{A}(\Pi) = \tau_L) < \Pr(E_{\tau_L} | \Pi) \cdot \delta'. \tag{40}$$

If $count_{\tau_L}(\Pi) \geq c_{min}$ then,

$$\Pr(\mathcal{A}(\Pi) = \tau_L) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau_L) \tag{41}$$

Proof. Let $\gamma^{(L)} = \top^L$.

$$\Pr(\mathcal{A}(\Pi) = \tau_L) = \Pr(\mathcal{A}_2(\Pi, Q_L^{\tau}) = \gamma^{(L)}) \Pr(E_{\tau_L} | \Pi) \tag{42}$$

If $count_{\tau_L}(\Pi) \geq c_{min}$, then using the fact that $\mathcal{A}_2$ is $(\varepsilon', 0)$-DP and Lemma 5.1, we can say that,

$$\Pr(\mathcal{A}(\Pi) = \tau_L) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau_L) \tag{43}$$

Recall that the output of $\mathcal{A}_2$ is the output of a sequence of calls to $\mathcal{A}_{PQ}$ (Algorithm 1). The probability that $\mathcal{A}_2$ outputs the sequence r can be upper-bounded by the probability of getting $\top$ on the Lth call to $\mathcal{A}_{PQ}$. Formally,

$$\Pr(\mathcal{A}_2(\Pi, Q^{\tau_L}) = \gamma^{(L)}) \leq \Pr(\mathcal{A}_{PQ}(\Pi, \tau_L, \hat{\theta}, \varepsilon') = \top) \tag{44}$$

If $count_{\tau_L}(\Pi) < c_{min}$, then using $count_{\tau}(\Pi) < c_{min} < c_{min}/p_{min} = \theta$, equation 44 and Lemma 5.2, we get,

$$\Pr(\mathcal{A}_2(\Pi, Q^{\tau_L}) = \gamma^{(L)}) \leq \Pr(\mathcal{A}_{PQ}(\Pi, \tau_L, \hat{\theta}, \varepsilon') = \top) < \delta' \tag{45}$$

$$\Pr(\mathcal{A}(\Pi) = \tau_L) \leq \Pr(E_{\tau_L} | \Pi) \delta' \qquad \text{[From } equation \text{ 42]} \tag{46}$$

Combining 43 and 46, we get the statement of the lemma. □

A.4 Proof of Theorem 5.3

Theorem 5.3. *Under Assumption 3.4, $\mathcal{A}$ is $(2\varepsilon', \frac{\delta_1}{2T})$-DP, where ε', δ_1 are defined in Algorithm 2.*

Proof. Let $\mathcal{T}$ denote the set of all possible trajectory prefixes using S, A. From Lemma 5.4, we have, $\forall \tau \in \mathcal{T}$, and all pairs of neighboring sets of experts $\Pi, \Pi' \subseteq \Pi_{p_{min}}$,

$$\Pr(\mathcal{A}(\Pi) = \tau) \leq e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau) + \Pr(E_\tau|\Pi) \cdot \delta' \tag{47}$$

where $\delta' = \delta_1/(2LT)$.
To prove that $\mathcal{A}$ is $(2\varepsilon', \delta_1/(2T))$-differentially private, we need to prove the following for any pair of neighbouring $\Pi, \Pi' \subseteq \Pi_{p_{min}}$:

$$\sum_{\tau \in \mathcal{T}} \max\{\Pr(\mathcal{A}(\Pi) = \tau) - e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau), 0\} \leq \frac{\delta_1}{2T} \tag{48}$$

Note that, for any fixed k,

$$\sum_{\tau \in \{\tau : \tau \in \mathcal{T}, |\tau| = k\}} \Pr(E_\tau|\Pi) = 1 \tag{49}$$

where $|\tau|$ denotes the length of the trajectory prefix τ.

$$\sum_{\tau \in \mathcal{T}} \Pr(E_\tau|\Pi) = \sum_{k=1}^{L} \left[\sum_{\tau \in \{\tau : \tau \in \mathcal{T}, |\tau| = k\}} \Pr(E_\tau|\Pi) \right] = L \tag{50}$$

where L is the length of a full trajectory. From equation 47, we have,

$$\sum_{\tau \in \mathcal{T}} \max\{\Pr(\mathcal{A}(\Pi) = \tau) - e^{2\varepsilon'} \Pr(\mathcal{A}(\Pi') = \tau), 0\} \leq \sum_{\tau \in \mathcal{T}} \Pr(E_\tau|\Pi)\delta' = L\delta' = \frac{\delta_1}{2T} \tag{51}$$

This proves Eqn 48, and hence the statement of the theorem.

□

B EXPERIMENTAL DETAILS

B.1 Dataset Generation

We set $p_{min} = 0.02$ for all the environments.

Lunar Lander: Among the various variable which govern the dynamics of this environment, we chose to modify gravity (default = 9.8), wind power (default = 0.0) and turbulence (default = 0.0) to train different experts. We modified gravity over 10 equally spaced values from 9.0 to 11.0, wind power over 10 equally spaced values between 0.0 to 5.0, and similarly turbulence over 10 equally spaced values from 0.0 to 0.5, to generate 1000 environment settings. For each of these environment settings, we trained agents using the PPO algorithm for $1e6$ updates along with hyperparameter tuning of learning rate and batch size. Each agent is a neural network with hidden-layers of sizes (512, 256, 64). For each environment setting, we chose the top 3 agents that performed the best on *the corresponding environment*, to form our 3000 experts.

CartPole: For CartPole, we chose to vary gravity, the magnitude of the force exerted when the cart is pushed and the mass of the cart. The default values of gravity, force magnitude and mass of the cart are 9.8, 10.0 and 1.0 respectively. We varied each control over 10 equally spaced values over a fixed range. For gravity, we fixed this range as (8.75, 11.0), for force magnitude as (9.0, 11.25) and for mass of the cart as (0.8, 1.25). Similar to above, for each of these 1000 environment settings, we trained agents using the PPO

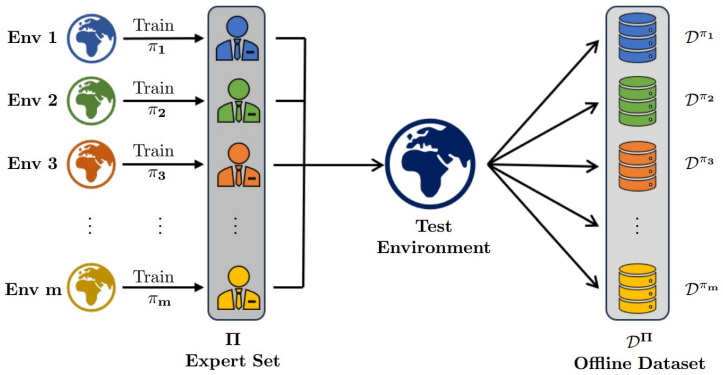


Figure 5: Outline of Data Generation scheme; We train experts on multiple environments as described below.

algorithm for $1e6$ updates while tuning the learning rate and batch size. Each agent is a neural network with hidden-layers of sizes (512, 256, 64). We chose the top 3 agents from each setting to form our 3000 experts.

Acrobot: For this environment, we varied the lengths of the 2 connected links and their masses. The default values of all these parameters is 1.0. We assign the same length to both the links in each setting. We vary this length, the mass of link 1 and the mass of link 2 over 10 equally spaced values across fixed ranges for each. We fix these ranges as (0.8, 1.2) for the length of the links, (0.9, 1.1) for the mass of link 1 and (0.9, 1.1) for the mass of link 2. The combination of these gives us 1000 environment settings. Again, we trained agents using PPO for $1e6$ updates over multiple learning rates and batch sizes and picked the top 3 agents from each setting to form our 3000 experts. Each agent is a neural network with hidden-layers of sizes (512, 256, 64).

HIV Treatment: In this domain, the 6-dimensional continuous state represents the concentrations of various cells and viruses. The actions correspond to the dosages of two drugs. The reward is a function of the health outcomes of the patient, and it also penalises the side-effects due to the quantity of drugs prescribed. Hence, the default reward function has 2 penalty terms, a penalty of 20000 on the dosage of drug 1, and a penalty of 2000 on the dosage of drug 2 (Ernst et al., 2006). In the default setting of the environment, we assume that the treatment at each steps is applied for 20 days and that there are 50 time steps in an episode. To train experts, we vary the penalty on drug 1 over 50 equally spaced values from 18750 to 21250, and the penalty on drug 2 over 20 equally spaced values from 1850 to 2150. Another variation that we introduce is whether the efficacy of the administered drug is noisy or not. So we either add no noise to the drug efficacy, or we add gaussian noise with standard deviation 0.01. All these modifications give us 2000 different environment settings. We trained agents using PPO for $2e6$ updates on each of the environment settings and over multiple learning rates and batch sizes. Each agent is represented by a neural network with a single hidden layer of size 512. We then picked the top agents from each of the 2000 environment settings to create experts. We picked 1000 more top-performing agents from the 2nd best agents of all the environment settings to get 3000 experts in total.

Once we have chosen 3000 experts for any of the above environments, we modify each expert's policy so that for each state, the action with the highest probability gets $(1 - (|A| - 1) \cdot p_{min})$ probability (where $|A|$ denotes the total number of actions), whereas all the other actions get p_{min} probability of being executed. We then use these modified experts to sample 20 trajectories each from the default environment to form our offline dataset.

B.2 Expert-Level DP-SGD

We adapted DP-SGD to the expert-level privacy setting to form the baseline that we compare against. To achieve this, during gradient updates, we ensure that an expert contributes at most one transition to a batch. To construct a batch of size b, we randomly sample b experts from the total m experts, and then pick a

transition from each selected expert. This is analogous to DP-SGD training on a dataset of size m. For privacy accounting, we use PLD accountant from the `tensorflow_privacy` library.

Algorithm 4 Expert-Level DP-SGD (outline)

Require: Model Parameters ϕ, Expert Datasets D^{Π}, Expert Policies Π, Loss function $\mathcal{L}(\phi) = \frac{1}{b}\sum_{i=1}^{b}\mathcal{L}(\phi, x_i)$, Learning Rate η, Batch size b, Noise parameter σ, Clipping threshold C, Training steps N

1: Initialize ϕ_0 randomly
2: **for** $t = 1, \ldots, N-1$ **do**
3: Sample subset of experts B from Π with subsampling probability $\frac{b}{m}$
4: For each $\pi_i \in B$ sample transition $x = (s, a, r, s')$ from D^{π_i} to form a batch of transitions B_{trans}
5: For each $x_i \in B_{trans}$ compute gradients $g_i(\phi_t) \leftarrow \nabla_{\phi_t}\mathcal{L}(\phi_t, x_i)$
6: Scale gradients to fit clipping threshold $\bar{g}_i(\phi_t) \leftarrow g_i(\phi_t)/\max\left(1, \frac{\|g_i(\phi_t)\|_2}{C}\right)$
7: Add gaussian noise to gradients $\tilde{g}_t \leftarrow \frac{1}{b}\left(\sum_{i=1}^{b}\bar{g}_i(\phi_t) + \mathcal{N}(0, \sigma^2 C^2 \mathrm{I})\right)$
8: Update model parameters $\phi_{t+1} \leftarrow \phi_t - \eta \cdot \tilde{g}_t$
9: **return** ϕ_T

Algorithm 4 describes a general outline for expert-level DP-SGD which we use as a baseline to compare our method against. The loss function here corresponds to the offline RL algorithm.

In our method, Algorithm 4 is incorporated in the training scheme in Algorithm 3 (lines 5 and 6) in the main paper. At each training step, with probability p, we execute lines 3-8 in Algorithm 4 to train on a batch of transitions. Otherwise, we sample a batch of transitions from $\mathcal{D}^{\Pi}_{stable}$ and perform a normal SGD update (this has no privacy cost). The privacy analysis for Algorithm 3 proceeds in a manner similar to that in standard DP-SGD (Abadi et al., 2016) with the exception of the subsampling probability which must be modified to be $q = p \cdot \frac{b}{m}$ (in the analysis only) to account for the randomness in picking the unstable dataset for sampling.

B.3 Hyperparameter Tuning

For all the environments, we assume that $p_{min} = 0.02$. For LunarLander, CartPole and HIVTreatment, for $\varepsilon = 5$, we set $\varepsilon_1 = 0, \varepsilon_2 = \varepsilon, \delta_1 = 0, \delta_2 = \delta$ and resort to pure expert-level DP-SGD training since the privacy budget is not enough to get meaningful stable trajectories for training. For Acrobot, we set $\varepsilon_1 = \varepsilon, \varepsilon_2 = 0, \delta_1 = \delta, \delta_2 = 0$. For all other cases, we set $\varepsilon_1 = \frac{3\varepsilon}{4}, \varepsilon_2 = \frac{\varepsilon}{4}, \delta_1 = \frac{9\delta}{10}, \delta_2 = \frac{\delta}{10}$.

We use Conservative Q-Learning (CQL) (Kumar et al., 2020) and Discrete Batch-Constrained Deep Q-Learning (BCQ) (Fujimoto et al., 2019a) for training on the offline data. For Lunar Lander, Acrobot and CartPole, we fix all neural network hidden layer sizes to (256, 256). For HIV Treatment, we use a neural network of one hidden layer of size 512 to represent the Q-function. For DP-SGD updates, we clip all gradient norms to 1.0. For Acrobot, for $\varepsilon = 10$, 7.5, 5, we set T (defined in Algorithm 2) equal to 25, 20, 20 respectively. For LunarLander, for $\varepsilon = 10$, 7.5, we set T equal to 100 and 62 respectively. For CartPole, for $\varepsilon = 10$, 7.5, we set T equal to 25 and 20 respectively. We use $T = 50$ for HIVTreatment.

We tune the following hyperparameters for our method: learning rate η, batch size b, probability of sampling from $\mathcal{D}^{\Pi}_{unst}$ during training (p), and DP-SGD noise multiplier (s) which is the standard deviation of the gaussian noise applied to the gradients divided by the clipping threshold. We perform a grid search over all hyper-parameters. The search spaces for different hyperparameters are as follows: η - [0.0001, 0.0005, 0.001, 0.005, 0.01], b - [64, 128, 256], p - [0.9, 0.8, 0.5, 0.0], and s - [10.0, 20.0, 30.0, 40.0, 50.0, 80.0]. For each configuration of hyperparameters, we let the model train for as long as possible with the given value of ε_2. We report the average episodic return obtained over 10 evaluation runs spaced over the last 10000 steps of training. The best hyper-parameter configurations for each setting are given in Table 2.

Table 2: Best hyper-parameter settings across different environments, offline-RL algorithms and ε values.

Environment	ε	**CQL**				**BCQ**			
		η	b	p	s	η	b	p	s
Acrobot	5	0.001	128	0.0	-	0.01	256	0.0	-
	7.5	0.001	128	0.0	-	0.01	256	0.0	-
	10	0.0005	128	0.0	-	0.01	256	0.0	-
LunarLander	5	0.0005	256	1.0	80	0.005	64	1.0	10
	7.5	0.005	64	0.8	20	0.001	128	0.5	50
	10	0.0005	64	0.8	40	0.005	128	0.5	20
CartPole	5	0.005	256	1.0	20	0.005	128	1.0	10
	7.5	0.001	128	0.9	50	0.01	128	0.5	20
	10	0.0001	128	0.8	80	0.001	128	0.5	40
HIVTreatment	5.0	0.01	256	1.0	10	0.005	128	1.0	20
	7.5	0.0001	64	0.0	-	0.005	64	0.9	20
	10	0.0001	64	0.0	-	0.005	64	0.9	20