Corruption Robust Offline Reinforcement Learning with Human Feedback

Debmalya Mandal
University of Warwick

Andi Nika
MPI-SWS

Parameswaran Kamalaruban
Featurespace

Adish Singla
MPI-SWS

Goran Radanović
MPI-SWS

Abstract

We study data corruption robustness for reinforcement learning with human feedback (RLHF) in an offline setting. Given an offline dataset of pairs of trajectories along with feedback about human preferences, an ε-fraction of the pairs is corrupted (e.g., feedback flipped or trajectory features manipulated), capturing an adversarial attack or noisy human preferences. We aim to design algorithms that identify a near-optimal policy from the corrupted data, with provable guarantees. Existing theoretical works have separately studied the settings of corruption robust RL (learning from scalar rewards directly under corruption) and offline RLHF (learning from human feedback without corruption); however, they are inapplicable to our problem of dealing with corrupted data in offline RLHF setting. To this end, we design novel corruption robust offline RLHF methods under various assumptions on the coverage of the data-generating distributions. At a high level, our methodology robustifies an offline RLHF framework by first learning a reward model along with confidence sets and then learning a pessimistic optimal policy over the confidence set. Our key insight is that learning optimal policy can be done by leveraging an offline corruption-robust RL oracle in different ways (e.g., zero-order oracle or first-order oracle), depending on the data coverage assumptions. To our knowledge, ours is the first work that provides provable corruption robust offline RLHF methods.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful paradigm for addressing complex tasks across diverse domains, ranging from large language models (LLMs) to robotics and game-playing Christiano et al. (2017); Ziegler et al. (2019); Stiennon et al. (2020); Ouyang et al. (2022); Bai et al. (2022); Shin et al. (2023). At the core of RLHF is its unique ability to model reward functions solely from preference data, making it particularly well-suited for scenarios where explicit reward signals are challenging to define. Following reward model estimation, traditional RLHF approaches employ online reinforcement learning algorithms for subsequent policy optimization. However, the integration of offline RL within the RLHF pipeline holds promise for alleviating limitations inherent to online RL, notably in terms of sample efficiency and safety concerns Levine et al. (2020); Kidambi et al. (2020). By incorporating offline RL algorithms, RLHF becomes more adaptable to scenarios where online data collection proves prohibitive, facilitating the reuse of valuable pre-existing datasets Shin et al. (2023).

The real-world deployment of RLHF faces substantial challenges rooted in the reliability of the preference data, which is integral to its effectiveness. These challenges primarily arise from two sources: adversarial corruption and inherent noise Casper et al. (2023); Xue et al. (2023); Chhan et al. (2024). Adversarial entities, acting with malicious intent, may deliberately manipulate feedback labels or trajectory features, introducing potential biases in the reward model. Simultaneously, inherent human subjectivity within crowd-sourced preference data can contribute substantial noise, impeding accurate reward estimation. In light of these challenges, a pivotal research question emerges: *Can we devise a robust variant of RLHF that efficiently learns from adversarially corrupted or noisy preference data, exhibiting graceful scalability amidst increasing corruption levels?*

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

Type of Coverage	Suboptimality Gap	Robust RL Oracle	# Oracle Calls
Uniform (ξ)	$O\left(\frac{H^3+\sqrt{Hd}}{\xi}\varepsilon^{1-o(1)}\right)$	R-LSVI (Zhang et al., 2022), zero-order access	1
Relative Condition Number (α)	$\widetilde{O}\left(H^2 d\kappa\sqrt{\alpha\varepsilon}\right)$ $+\ \widetilde{O}\left(H^{5/4}d^{3/4}(\alpha\varepsilon)^{1/4}\right)$	R-LSVI (Zhang et al., 2022), zero-order access	$\widetilde{O}\left(\frac{H^{3/2}d^5}{\varepsilon^3}\right)$
Generalized Coverage Ratio (ν)	$O\left(\nu\kappa\sqrt{\varepsilon}H^2 d^{3/2}\right)$	Algorithm 7 (Our method), first-order access	$O\left(\frac{1}{\varepsilon\nu}\right)$

Table 1: We design *corruption robust RLHF* through reduction to *corruption robust offline RL* problem. Here H is the horizon length, d is the dimension of the features, and κ and α are constants . Under uniform coverage and low relative condition number, we use R-LSVI as an oracle, and obtain suboptimality gap of $O(\varepsilon^{1-o(1)})$ and $O(\varepsilon^{1/4})$ respectively, in terms of ε (fraction of corrupted data). Calls to R-LSVI are zero-order i.e. we only obtain a robust policy and an estimate of the value function. Under bounded generalized coverage ratio, we design a new robust offline RL method (algorithm (7)) that also returns an estimate of the sub-gradient (first order access). Using algorithm (7), we can improve the dependence on ε to $O(\sqrt{\varepsilon})$ and also significantly reduce the number of oracle calls.

In this paper, we initiate the study of *corruption-robust offline reinforcement learning from human feedback*. Although there are several works on corruption robust offline reinforcement learning Zhang et al. (2022); Ye et al. (2023b), and provable preference based reinforcement learning Zhan et al. (2023); Zhu et al. (2023), ours is the first work to combine these two threads and provide provable corruption robust offline RLHF methods. We design corruption robust offline RLHF methods through reduction to corruption robust offline RL methods. In particular, we modify the existing RLHF framework through three steps – (1) Robustly learn a reward model by solving a robust logistic regression problem, (2) Construct a confidence set around the learned model, and (3) learn a pessimistic optimal policy over the confidence set through reduction to offline RL. We instantiate this general framework for datasets with various types of coverage assumptions, and as is often the case in offline RL, different coverage assumptions require different algorithms.

In particular, we consider the standard *Huber contamination* model where ε-fraction of the data (human feedback, features of the trajectories or both) are corrupted. Moreover, we consider a linear Markov decision process Jin et al. (2020) with horizon length H, and feature dimension d. Then, we prove the following set of results.

1. When the offline data has *uniform coverage*, we show that it is possible to learn a policy with sub-optimality gap at most $O\big(H^3\sqrt{d}\varepsilon^{1-o(1)}\big)$.[*]

2. When the offline data satisfies the condition of *low relative condition number*, a condition substantially weaker than the uniform coverage, we bound the sub-optimality gap by $\widetilde{O}\left(H^2 d\sqrt{\varepsilon} + H^{5/4}d^{3/4}\varepsilon^{1/4}\right)$. For ε small (i.e. $< 1/d$) it can be checked that the upper bound is $O(H^2 d^{3/4}\varepsilon^{1/4})$. In order to achieve this bound, we reduce our problem to corruption robust offline RL, by using an existing corruption-robust method (Zhang et al., 2022) as a biased, zero-order oracle, and using the technique of *Gaussian approximation* Nesterov and Spokoiny (2017) to construct an approximate sub-gradient. We also develop a method of convex optimization with biased zero-order oracle that might be of independent interest.

3. Finally, we show that we can improve the sub-optimality gap to $\widetilde{O}(H^2 d^{3/2}\sqrt{\varepsilon})$ if the offline data satisfies the assumption of *bounded generalized coverage ratio*, an assumption recently considered by Gabbianelli et al. (2024) (see also Jin et al. (2021) for a similar coverage ratio). In this case, we construct a new corruption robust offline RL that is *first-order* i.e. not only returns an approximately optimal policy but also an approximate sub-gradient of the optimal value function.

1.1 Related Work

Preference-based RL: Our work is related to preference-based reinforcement learning (PbRL) (Wirth et al., 2017; Lee et al., 2021). Although the field of PbRL is not new, there have been significant recent interests in designing provably optimal RL methods from preferences (Zhan et al., 2023; Zhu et al., 2023; Wang et al., 2023). In particular, Zhu et al. (2023) proposed a pessimistic maximum likelihood estimation for provable PbRL

[*] As $\varepsilon \to 0$, $\varepsilon^{1-o(1)}$ approaches ε. We actually show a dependence of $\varepsilon \cdot \exp(\sqrt{\log(1/\varepsilon)})$ which is $\varepsilon^{1-o(1)}$.

under clean data. Our algorithm, in particular the reward confidence set construction, is related to the method proposed by Zhan et al. (2023). However, unlike Zhan et al. (2023) we don't build a confidence set around the probability transpition function, but rather use reduction to offline RL. Finally, there are several works on PbRL in online setting (Pacchiano et al., 2021; Chatterji et al., 2021; Chen et al., 2022) which are complementary to the offline setting.

Corruption robust RL: Our work is closely aligned with the research on corruption robust RL, where the challenge lies in designing agents that can effectively learn in the presence of adversarial corruption on both rewards and transitions Rakhsha et al. (2020). Zhang et al. (2022) has considered linear MDP, and have designed corruption robust offline RL by robustifying the least squares value iteration method. On the other hand, Ye et al. (2023b) has considered corruption robustness in general MDPs by adopting uncertainty weighting to nonlinear function approximation Ye et al. (2023a). In the online RL setting, Lykouris et al. (2021); Chen et al. (2021) proposed robust RL methods capable of accommodating up to $\epsilon \leq O(1/\sqrt{T})$ fraction of corruptions. Zhang et al. (2021) developed an online policy gradient method that is resilient against a constant fraction of adaptive corruption. Furthermore, there are other approaches for robustness in offline RL, including model selection Wei et al. (2022), hybrid RL Panaganti et al. (2022), and others Wu et al. (2022); Yang et al. (2023).

2 PRELIMINARIES

Markov Decision Process: Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P^\star, r^\star, H, \rho)$ be an episodic Markov Decision Process (MDP) where $\mathcal{S}$ denotes the state space and $\mathcal{A}$ denotes the action space. The initial state is sampled from the distribution ρ. $P^\star = (P_1^\star, \ldots, P_H^\star)$ denote the transition kernels where for each $h \in [H]$, $P_h^\star(\cdot|s,a) \in \Delta(\mathcal{S})$ denotes the distribution over states given that the system is in state s at step h and action a is taken. Let $r^\star : \mathcal{S} \times \mathcal{A}$ be the reward function where $r_h^\star(s,a)$ is the reward obtained from taking action a from state s at time-step h. We can also extend the reward function to reward over trajectories by taking the sum of the rewards over the H steps. Specifically, given a trajectory $\tau = (s_1, a_1, s_2, \ldots, s_{H+1})$, we define $r^\star(\tau) = \sum_{h=1}^H r_h^\star(s_h, a_h)$.

Policy: Policies denote mappings from histories of traversed state-action pairs to distributions over actions. Formally, a non-stationary history-dependent policy $\pi = (\pi_1, \ldots, \pi_H)$ is a sequence of mappings where, for each $h \in [H]$, $\pi_h : \mathcal{H}_h \to \Delta(\mathcal{A})$, with $\mathcal{H}_h = \mathcal{S} \times (\mathcal{S} \times \mathcal{A})^{H-1} \times \mathcal{S}$ denoting the history space up to time-step h. The space of such policies is denoted by Π_{his}. We further denote by $q_h^\pi(s,a) = \mathbb{P}(s_h = s, a_h = a | \pi, P^\star)$ the state-action occupancy measure for every time-step $h \in [H]$.The expected performance of a given policy π with respect to the true transitions $P^\star$ and true reward $r^\star$ is denoted by

$$V^\pi(P^\star, r^\star) = \mathbb{E}\left[\sum_{h=1}^H r^\star(s_h, a_h) \middle| s_h \sim P_h^\star, a_h \sim \pi_h \, \forall h\right].$$

2.1 Offline RLHF

We have an offline dataset $\mathcal{D} = \left\{(\tau^{n,0}, \tau^{n,1}, o^n\right\}_{n=1}^N$ of N pairs of trajectories, where each pair $(\tau^{n,0}, \tau^{n,1})$ is associated with feedback $o^n \in \{+1, -1\}$ representing the human preference, coming from a latent model assumed to satisfy the following assumption.

Assumption 1 (Preference-based model)**.** *Given a pair of trajectories (τ^0, τ^1), and a preference $o \in \{+1, -1\}$, the probability that τ^1 is preferred over τ^0 satisfies*

$$\mathbb{P}\left(o = 1 | \tau^0, \tau^1\right) = \sigma\left(r^\star(\tau^1) - r^\star(\tau^0)\right) ,$$

where $\sigma : \mathbb{R} \to [0,1]$ is a monotonically increasing link function.

In this paper, we will utilize the sigmoid link function $\sigma(x) = 1/(1 + \exp(-x))$, commonly used in the literature on RLHF Christiano et al. (2017). For our setting, the rewards are bounded and the range of the function is bounded away from 0 and 1. This also implies that there exists a constant κ such that $\sup_{p \in [0,1]} \left|\frac{d\sigma^{-1}(p)}{dp}\right| \leq \kappa$. The performance of a given policy π is measured by the notion of suboptimality gap with respect to a target policy $\pi^\star$. Formally, we want to minimize

$$\text{SubOpt}(\pi, \pi^\star) = V^{\pi^\star}(r^\star, P^\star) - V^\pi(r^\star, P^\star) .$$

We will assume that the pairs of trajectories (τ^0, τ^1) in our offline dataset $\mathcal{D}$ are drawn from pairs of behaviour policies (μ_0, μ_1), and we will denote it as $\tau^0 \sim \mu_0, \tau^1 \sim \mu_1$. Additionally, we will write $\Sigma_{\mu_0,\mu_1}^{\text{diff}}$ to denote the difference feature covariance matrix, which is defined as

$$\Sigma_{\mu_0,\mu_1}^{\text{diff}} = \mathop{\mathbb{E}}_{\tau^0 \sim \mu_0, \tau^1 \sim \mu_1}\left[\left(\phi(\tau^0) - \phi(\tau^1)\right)\left(\phi(\tau^0) - \phi(\tau^1)\right)^\top\right]. \tag{1}$$

Similarly, we will write $\Sigma_{\mu_0,\mu_1}^{\text{avg}}$ to denote the average feature covariance matrix, which is defined as

$$\Sigma_{\mu_0,\mu_1}^{\text{avg}} = \mathop{\mathbb{E}}_{\tau^0 \sim \mu_0, \tau^1 \sim \mu_1}\left[\left(\phi(\tau^0) + \phi(\tau^1)\right)\left(\phi(\tau^0) + \phi(\tau^1)\right)^\top\right]. \tag{2}$$

2.2 Contamination Model

In this paper, we study the problem of corruption robustness in offline RLHF. We assume that the collected data contains an ϵ-fraction of contaminated samples, i.e. an attacker, who is given access to the data beforehand, is allowed to arbitrarily modify up to an ϵ-fraction of the data samples (both the the trajectory features, and the human feedback). Formally, *Huber contamination model of human preferences* is defined as follows.

Assumption 2 (ε-corruption in Offline RLHF)**.** *Let $\varepsilon \in [0,1]$ denote the contamination parameter and $\widetilde{\mathcal{D}} = \{(\widetilde{\tau^{n,0}}, \widetilde{\tau^{n,1}}, \widetilde{o}^n\}_{n=1}^N$ be a dataset of N pairs of trajectories and human preferences. An attacker inspects $\widetilde{\mathcal{D}}$ and arbitrarily modifies up to ϵN tuples from $\widetilde{\mathcal{D}}$. We denote the corrupted dataset by $\mathcal{D} = \{(\tau^{n,0}, \tau^{n,1}, o^n)\}_{n=1}^N$. In other words, there are at most ϵN indices n, for which we have $\widetilde{o}^n \neq o^n$, or $\widetilde{\tau^{n,1}} \neq \tau^{n,1}$, or $\widetilde{\tau^{n,0}} \neq \tau^{n,0}$.*

2.3 Parametric Markov Decision Processes

It is generally impossible to design provable offline RL algorithms without making any parametric assumptions on the underlying MDP. Therefore, throughout this paper, we will assume that the MDP is linear i.e. the reward and the transition are linear functions of by d-dimensional features.

Definition 1 (Linear MDP Jin et al. (2020))**.** *We assume access to known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, and that there exist $\{\theta_h\}_{h\in[H]}$ and signed measures $\boldsymbol{\mu}_h = (\mu_h^1, \ldots, \mu_h^d)$ over the state space so that*

$$r_h(s,a) = \phi(s,a)^\top \theta_h \text{ and } P_h(s' \mid s,a) = \phi(s,a)^\top \boldsymbol{\mu}_h(s').$$

We also assume $\left\|\phi(s,a)\right\|_2 \leq 1$ for any s, a, $\max\{\left\|\theta_h\right\|_2, \left\|\boldsymbol{\mu}_h(\mathcal{S})\right\|_2\} \leq \sqrt{d}$ for any $h \in [H]$.

Given a trajectory $\tau = (s_1, a_1, s_2, \ldots, s_{H+1})$ we will write $\phi(\tau) = [\phi(s_1,a_1); \ldots; \phi(s_H, a_H)]$ to denote the feature of the trajectory τ. Note that $\phi(\tau) \in \mathbb{R}^{Hd}$ and $\left\|\phi(\tau)\right\|_2 \leq \sqrt{H}$.

3 ROBUST RLHF WITH UNIFORM COVERAGE

We now provide our first algorithm for corruption robust reinforcement learning from human feedback (RLHF). Standard RLHF framework estimates the reward parameter by solving a maximum likelihood estimation problem. We essentially robustify this step and replace it with a robust version of logistic regression. Let $\mathbb{P}_\theta(o \mid \phi(\tau^1) - \phi(\tau^0))$ be the probability of observing feedback o from a comparison of trajectory τ^1 and τ^0. Then algorithm 1 solves a trimmed maximum likelihood estimation problem,

$$\widehat{\theta} \leftarrow \operatorname{argmax}_\theta \max_{S \subseteq \mathcal{D}: |S| = (1-\varepsilon)N} \sum_{n \in S} \log \mathbb{P}_\theta(o^n \mid \mathrm{x}^n) \quad (3)$$

where $\mathrm{x}^n = \phi(\tau^{n,1}) - \phi(\tau^{n,0})$. Therefore, the estimate $\widehat{\theta}$ is chosen to maximize the likelihood over best subset containing $(1-\varepsilon)$-fraction of the points. With an estimate $\widehat{\theta}$ of the reward parameter, algorithm 1 uses a robust offline RL method (input RobRL) to compute an approximately optimal policy $\widetilde{\pi}$. For this step, algorithm 1 uses the features from the dataset $\widehat{\mathcal{D}}$ but the reward is defined according to the estimated model $\widehat{\theta}$ i.e. $r_h(s,a) = \widehat{\theta}_h^\top \phi(s,a)$. Finally, note that the dataset $\mathcal{D}$ is split uniformly at random into two datasets of equal sizes at the beginning, and one partition is used for reward estimation, and the other for policy optimization.

```
ALGORITHM 1: Robust RLHF (with Uniform Coverage)
Input: (a) Dataset $\mathcal{D}$, (b) corruption parameter $\epsilon$,
       (c) corruption robust offline RL algorithm RobRL.
Partition dataset $\mathcal{D}$ uniformly at random into two
  datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ of equal size.
/* Estimate reward parameter of linear MDP
   $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_H)$.   */
Let $\mathrm{x}^n = \phi(\tau^{n,1}) - \phi(\tau^{n,0})$. Using algorithm 2, find an
  approximate stationary point of the following
  trimmed maximum likelihood estimation problem.
    $\widehat{\theta} \leftarrow \arg\max_\theta \max_{\substack{\widehat{S} \subseteq \mathcal{D}_1 \\ |\widehat{S}| = (1-\varepsilon)N/2}} \sum_{n \in \widehat{S}} \log \mathbb{P}_\theta(o^n \mid \mathrm{x}^n)$    (4)
Let $\widetilde{\pi}$ be the policy returned by RobRL with reward
  function $r_h(s,a) = \phi(s,a)^\top \widehat{\theta}_h$ and dataset $\mathcal{D}_2$.
return $\widetilde{\pi}$.
```

Solving Trimmed MLE: Before providing the performance guarantees of algorithm 1, we discuss how to solve the trimmed maximum likelihood estimation (MLE) problem (3). In general, the optimization problem is hard to solve, but it is known that for the setting of generalized linear models, alternating optimization methods converge to a stationary point under certain assumptions on the link function Awasthi et al. (2022). Therefore, algorithm 1 calls an alternating optimization method (algorithm 2) to obtain an approximate stationary point of the trimmed MLE problem.

Let us now define the notion of an approximate stationary point. Given an estimate $\hat{\theta}$, let the set $\widehat{S}$ contain $(1-\varepsilon)N$ data-points with the largest log-likelihood under $\hat{\theta}$. Then we say $\hat{\theta}$ is a γ-stationary point if the following condition is satisfied.

$$\frac{1}{N} \sum_{n \in \widehat{S}} \nabla_\theta \log \mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)^\top \frac{(\theta^\star - \hat{\theta})}{\|\theta^\star - \hat{\theta}\|_2} \leq \gamma \qquad (5)$$

ALGORITHM 2: Alternating Optimization

```
Input: Corrupted dataset $\mathcal{D}$, corruption parameter
       $\epsilon$, slackness parameter $\eta$, transition model $P^\star$.
Let $\mathrm{x}^n = \phi(\tau^{n,1}) - \phi(\tau^{n,0})$.
Set $\widehat{\theta}_1 = 0$.
for $t = 1, 2, \dots$ do
    $\widehat{S}_t \leftarrow \arg\max_{S \subseteq [N], |S| = (1-\varepsilon)N} \sum_{n \in \widehat{S}} \log \mathbb{P}_{\widehat{\theta}_t}(o^n \mid \mathrm{x}^n)$.
    $\widehat{\theta}_{t+1} \leftarrow \arg\max_{\theta : \|\theta\|_2 \leq \sqrt{Hd}} \sum_{n \in \widehat{S}_t} \log \mathbb{P}_\theta(o^n \mid \mathrm{x}^n)$.
    /* Check Progress */
    if
     $\sum_{n \in \widehat{S}} \log \mathbb{P}_{\widehat{\theta}_{t+1}}(o^n | \mathrm{x}^n) \leq \sum_{n \in \widehat{S}} \log \mathbb{P}_{\widehat{\theta}_t}(o^n | \mathrm{x}^n) + \eta$
     then
        Return $\widehat{\theta}_t$.
```

We propose an alternating optimization based method (Algorithm 2) to obtain such an approximate stationary point of the trimmed MLE problem. At each iteration t, the alternating optimization updates $\widehat{S}$ and $\widehat{\theta}$ as follows.

1. $\widehat{S}_t \leftarrow \arg\max_{S \subseteq [N], |S| = (1-\varepsilon)N} \sum_{n \in S} \log \mathbb{P}_{\widehat{\theta}_t}(o^n \mid \mathrm{x}^n)$.
2. $\widehat{\theta}_{t+1} \leftarrow \arg\max_{\theta : \|\theta\|_2 \leq \sqrt{Hd}} \sum_{n \in \widehat{S}_t} \log \mathbb{P}_\theta(o^n \mid \mathrm{x}^n)$,

where $\mathrm{x}^n = \phi(\tau^{1,n}) - \phi(\tau^{0,n})$. The method stops when the improvement in the likelihood is less than a threshold η. The next lemma shows (proof in Appendix A) that algorithm 2 obtains a γ-stationary point of the trimmed maximum likelihood estimation problem (3) for $\gamma = \max\left\{2L\varepsilon, \frac{\varepsilon^2}{\|\theta^\star - \widehat{\theta}\|_2}\right\}$.

Lemma 1. *Suppose $\|\mathrm{x}_n\|_2 \leq L$ for all n and we set $\eta = \varepsilon^2$. Then algorithm 2 computes a $\max\left\{2L\varepsilon, \frac{\varepsilon^2}{\|\theta^\star - \widehat{\theta}\|_2}\right\}$-stationary point of the trimmed maximum likelihood estimation problem (3).*

Note that algorithm 2 stops when the improvement in the objective falls below ε^2. Furthermore, the value of the log-likelihood can be bounded by $O(L)$ which implies that algorithm 2 runs for at most $O(L/\varepsilon^2)$ iterations. We next show that an approximate sationary solution of trimmed MLE is enough to provide a bound on the sub-optimality of the policy $\widetilde{\pi}$ returned by algorithm 1. We will make the following assumption regarding the pair of policies μ_0, μ_1 that generate the offline data.

Assumption 3 (Uniform Coverage)**.** *Suppose $\|\phi(\tau)\|_2 \leq L$ for any trajectory τ. Then there exists a constant $\xi > 0$ such that*

$$\Sigma^{diff}_{\mu_0,\mu_1} \succcurlyeq \xi L \cdot \mathrm{Id}_d.$$

Theorem 3.1. *Suppose* `RobRL` *returns a $f(\varepsilon)$-robust estimate of the optimal value function, and assumption 3 holds with $\xi \geq 5\varepsilon$ and $N \geq \Omega\left(\frac{H^{3/2}}{\varepsilon^2}(d + \log(1/\delta))\right)$. Then with probability at least $1 - \delta$, the policy $\widetilde{\pi}$ returned by algorithm 1 satisfies,*

$$V^\star(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \leq f(\varepsilon) + 2\sqrt{Hd} C_1 \frac{\varepsilon}{\xi} \cdot \exp(\sqrt{\log(1/2\delta\varepsilon)})^\dagger$$

The proof of the theorem is provided in the Appendix, but the main idea is to show that the estimate $\widehat{\theta}$ obtained from solving trimmed MLE satisfies $\left\|\widehat{\theta} - \theta^\star\right\|_2 = O(\varepsilon^{1-o(1)})$. We now instantiate Theorem 3.1 for the setting of linear MDP. For corruption robust offline RL, we use algorithm `R-LSVI` from Zhang et al. (2022) as an oracle. It requires a coverage assumption similar to assumption (3).

Assumption 4 (Uniform Coverage: V2)**.** *Suppose, $\|\phi(\tau)\|_2 \leq L$ for any trajectory τ. Then there exists a constant $\xi > 0$ such that*

$$\Sigma^{avg}_{\mu_0,\mu_1} \succcurlyeq \xi L \cdot \mathrm{Id}_d.$$

Under assumption (4) `R-LSVI` returns a policy $\widetilde{\pi}$ so that $V^{\widetilde{\pi}}(s_0) \geq V^\star(s_0) - f(\varepsilon)$ where $f(\varepsilon) = \widetilde{O}\left(\frac{H^{5/2}}{\xi\sqrt{N}}\mathrm{poly}(d) + \frac{H^3}{\xi}\varepsilon\right)$. Note that if $N \geq \widetilde{\Omega}\left(\frac{\mathrm{poly}(d)}{\varepsilon^2}\right)$, we have $f(\varepsilon) = \frac{H^3}{\xi}\varepsilon$. Substituting this value of $f(\varepsilon)$ in the bound of theorem (3.1) gives us the following bound on the suboptimality gap.

Proposition 1. *Suppose assumptions 3 and 4 hold. Then for the setting of linear MDP and $N \geq \widetilde{\Omega}\left(\frac{poly(d)\log(1/\delta)}{\varepsilon^2}\right)$, Algorithm 1 returns a policy $\widetilde{\pi}$ so that with probability at least $1 - \delta$,*

$$V^\star(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \leq O\left(\frac{H^3 + \sqrt{Hd}}{\xi}\varepsilon \cdot \exp(\sqrt{\log(1/2\delta\varepsilon)})\right).$$

4 LOW RELATIVE CONDITION NUMBER

Although the assumption of uniform coverage allows us to design $O(\varepsilon^{1-o(1)})$-optimal policy, it is a strong assumption since the features generated by the offline policy might not cover the entire d-dimensional space.

[†] For most of our results, the sub-optimality can be shown to be of the form $g(\varepsilon) + c/\sqrt{N}$. We assume that the number of samples N is large so that the term involving ε is the dominating term. This is to simplify the results and follows existing literature Diakonikolas and Kane (2019).

In this section, we relax this assumption to a new assumption named *Low Relative Condition number*, which is significantly weaker.

Assumption 5 (Relative Condition Number)**.** *Let $\Sigma^{diff}_{\mu_0,\mu_1}$ (resp. $\Sigma^{diff}_{\pi_0,\pi_1}$) be the difference feature covariance matrix under pair of policies μ_0 and μ_1 (resp. π_0 and π_1), as defined in eq. (1). Then there exists a constant $\alpha > 0$ such that the following holds.*

$$\sup_w \frac{w^\top \Sigma^{diff}_{\pi_0,\pi_1} w}{w^\top \Sigma^{diff}_{\mu_0,\mu_1} w} = \alpha < \infty$$

Although the above assumption is stated for a pair of policies π_0, π_1, given a target policy $\pi^\star$ one can choose $\pi_0 = \pi^\star$ and $\pi_1 = \mu_1$, and the assumption needs to hold only for this pair of policies.

Algorithm 3 provides our new corruption robust RLHF method under the assumption of low relative condition number. The algorithm begins similarly to algorithm 1 by solving the trimmed maximum likelihood estimation to obtain a robust estimate $\widehat{\theta}$ of the reward parameter $\theta^\star$. However, in the absence of uniform coverage, $\widehat{\theta}$ might not be close to $\theta^\star$ in terms of L_2 distance. So the following lemma provides a bound in terms of the likelihood at $\widehat{\theta}$ and $\theta^\star$.

Lemma 2. *Let $\mathbb{P}_\theta(y \mid x) = 1/(1 + e^{-y\cdot\theta^\top x})$ and for any $n \in [N]$ define $\mathrm{x}_n = \phi(\tau^{1,n}) - \phi(\tau^{0,n})$. Then with probability at least $1 - \delta$, we have*

$$\frac{1}{N}\sum_{n=1}^{N} \log\left(\frac{\mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)}\right) \le 6\varepsilon H\sqrt{d} + c\cdot\frac{d}{N}\log\left(\frac{HN}{\delta}\right)$$

The above result is a generalization of Lemma 1 from Zhan et al. (2023), and allows us to build a confidence set around the estimate $\widehat{\theta}$ even when a ε-fraction of the data has been corrupted (line 3 of algorithm (3)). Now we leverage two important observations.

First, the above approach requires the set $\Theta(\mathcal{D}_1)$ is a convex set. It can be easily checked this is true if the function $\log \mathbb{P}_\theta(\cdot)$ is concave. Moreover, for the case of sigmoid link function $\nabla^2_\theta \log \sigma(\theta^\top x) \preccurlyeq 0$ i.e. $\Theta(\mathcal{D}_1)$ is a convex set. Second, for the setting of linear MDP, the optimal value function $V^\star(\theta) = \max_\pi V^\pi(\theta)$ is a convex function in the reward parameter θ. This follows from the occupancy measure based representation of MDP. Indeed, $V^\star(\theta) = \max_{q^1,\ldots,q^H \in \mathcal{C}} \sum_{h=1}^{H} q_h^\top \Phi\theta$, where $\mathcal{C}$ is the set of valid occupancy measures, and Φ is the feature matrix. Since for any θ, $V^\star(\theta)$ is a maximum over linear functions, $V^\star(\cdot)$ is convex.

```
ALGORITHM 3: Robust RLHF with Condition Number
Input: (a) Corrupted dataset $\mathcal{D}$, (b) corruption parameter $\epsilon$, (c) corruption robust offline RL
       algorithm RobRL, (d) reference distribution $\mu_{\mathrm{ref}}$.
1 Partition dataset $\mathcal{D}$ uniformly at random into $\mathcal{D}_1$ and $\mathcal{D}_2$ of equal size.
  /* Estimate an estimate $\widehat{\theta}$ of the reward parameter, as in Algorithm (1).          */
2 Set $\zeta = 6\varepsilon H\sqrt{d} + 2\frac{d}{N}\log\left(\frac{HN}{\delta}\right)$ and $\Theta(\mathcal{D}_1) = \left\{\theta : \|\theta\|_2 \le \sqrt{Hd} \wedge \frac{2}{N}\sum_{n=1}^{N/2} \log\frac{\sigma\left(\theta^\top \mathrm{x}^n\right)}{\sigma\left(\widehat{\theta}^\top \mathrm{x}^n\right)} \ge -\zeta\right\}$
  /* Run Projected Sub-gradient Descent with Biased Oracle                                       */
3 Initialize $\theta_0 \in \Theta(\mathcal{D}_1)$.
4 for $t = 0, \ldots, T-1$ do
      /* Sub-Gradient Construction                                                               */
5     Generate $u_1, \ldots, u_K$ uniformly at random from the standard normal distribution.
6     Let $g_t = \frac{1}{K}\sum_{k=1}^{K} \frac{\widehat{V}(\theta_t + \mu u_k) - \widehat{V}(\theta) - \mu\cdot\mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}\left[\phi(\tau)^\top u_k\right]}{\mu} u_k$ be
      the approximate sub-gradient, where $\widehat{V}(\theta)$ is the value estimate returned by RobRL with reward
      function $r_h(s,a) = \phi(s,a)^\top \theta_{t,h}$ and dataset $\mathcal{D}_2$.
7     $\theta_{t+1} = \mathrm{Proj}_{\Theta(\mathcal{D}_1)}\left(\theta_t - \eta g_t\right)$
8 Set $\overline{\theta} = \frac{1}{T}\sum_{k=1}^{T} \theta_k$ and let $\widetilde{\pi}$ be the policy returned by running RobRL with reward function
  $r_h(s,a) = \phi(s,a)^\top \overline{\theta}_h$.
9 return $\widetilde{\pi}$.
```

Therefore, we run a projected subgradient descent over the set $\Theta(\mathcal{D}_1)$. At each iteration t, algorithm (3) selects a reward parameter θ_t. Although the corruption robust offline RL method RobRL can return an approximately optimal policy with reward parameter θ, we need a subgradient i.e. $g_t \in \delta_\theta V^\star(\theta_t) = \left\{v : V^\star(\theta') \ge V^\star(\theta_t) + v^\top(\theta' - \theta_t)\ \forall \theta'\right\}$. So we treat RobRL as a biased, zero-order oracle and explicitly build an estimator of a subgradient (lines 8-9) Nesterov and Spokoiny (2017); Duchi et al. (2015); Flaxman et al. (2004). In particular, we use the *gaussian approximation* technique introduced by Nesterov and Spokoiny (2017). Given a convex function $f : E \to \mathbb{R}^d$, let f_μ be its smoothed Gaussian approximation, defined as

$$f_\mu(\theta) = \frac{1}{\kappa}\int_E f(\theta + \mu\cdot u)e^{-1/2\|u\|_2^2}du,$$

where $\kappa = \int_E e^{-1/2\|u\|_2^2}du$. The Gaussian approximation method performs a gradient descent of the smoothed function f_μ, with the gradient

$$\nabla f_\mu(\theta) = \frac{1}{\kappa}\int_E \frac{f(\theta + \mu\cdot u) - f(\theta)}{\mu}e^{-1/2\|u\|_2^2}du.$$

Algorithm 3 constructs an estimator of $\nabla f_\mu(\theta)$ for $f(\theta) = V^\star(\theta) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}\left[\phi(\tau)^\top\theta\right]$[‡]. The algorithm

[‡] We subtract rewards according to a reference policy μ_{ref} since we only have preference data over rewards.

finally computes the average reward parameter $\overline{\theta} = 1/T \cdot \sum_{k=1}^{T} \theta_k$ and returns a robust policy $\widetilde{\pi}$ with respect to the parameter $\overline{\theta}$. Algorithm 3 provides our full implementation of the reduction to corruption robust RL. The next theorem provides a bound on the sub-optimality gap of algorithm 3, assuming access to a $f(\varepsilon)$-robust offline RL method.

Theorem 4.1. *Suppose assumption 5 holds, $\sup_{p\in[0,1]} \left|\frac{d\sigma^{-1}(p)}{dp}\right| \leq \kappa$, and RobRL returns a $f(\varepsilon)$-robust estimate of the optimal value function. If $N \geq \widetilde{\Omega}\left(\frac{H^{3/2}d^5}{\varepsilon^3}\right)$, then for a target policy $\pi^\dagger$, the policy $\widetilde{\pi}$ output by algorithm 3 satisfies the following w.p. at least $1-\delta$.*

$$V^{\pi^\dagger}(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \leq f(\varepsilon) + 8\sqrt{f(\varepsilon)}(Hd)^{1/4} + c\kappa\sqrt{\alpha}\left(\sqrt{\varepsilon H}d^{1/4} + \sqrt{d/N \cdot \log\left(HdN/\delta\right)}\right)$$

We now instantiate Theorem 4.1 for the setting of linear MDP. For corruption robust offline RL, we use algorithm (`R-LSVI` from Zhang et al. (2022)) as an oracle, which requires a coverage assumption.

Assumption 6 (Relative Condition Number: V2). *Let $\Sigma^{avg}_{\mu_0,\mu_1}$ be the average feature covariance matrix under pair of distributions μ_0 and μ_1, as defined in eq. (2). Then there exists a constant $\alpha > 0$ such that the following condition holds.*

$$\sup_w \frac{w^\top \Sigma_{\pi^\star} w}{w^\top \Sigma^{avg}_{\mu_0,\mu_1} w} = \alpha < \infty$$

Under assumption 6, `R-LSVI` returns a policy $\widetilde{\pi}$ so that $V^{\widetilde{\pi}}(s_0) \geq V^\star(s_0) - f(\varepsilon)$ where $f(\varepsilon) = \widetilde{O}\left(\frac{H^{5/2}}{\sqrt{N}}\sqrt{\alpha}\mathrm{poly}(d) + H^2 d\sqrt{\alpha\varepsilon}\right)$.

Proposition 2. *Suppose assumptions (5) and (6) hold. Moreover, suppose $\sup_{p\in[0,1]}\left|\frac{d\sigma^{-1}(p)}{dp}\right| \leq \kappa$. Then for the setting of linear MDP and $N \geq \widetilde{\Omega}\left(\frac{H^{3/2}}{\varepsilon^3}\cdot poly(d, 1/\delta)\right)$, algorithm 3 returns a policy $\widetilde{\pi}$ so that with probability at least $1-\delta$,*

$$V^\star(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \leq \widetilde{O}(H^2 d\kappa\sqrt{\alpha\varepsilon}) + \widetilde{O}\left(H^{5/4}d^{3/4}(\alpha\varepsilon)^{1/4}\right)$$

Proposition 2 provides an upper bound of $O(\varepsilon^{1/4})$ when other parameters are constant. The reason we obtain sub-optimal dependence on ε is because we assume a zero-order access to the offline robust RL oracle. We now show that we can improve the dependence on ε with access to a first-order oracle.

```
ALGORITHM 4: Robust FreeHand with First-Order Oracle
Input: (a) Corrupted dataset $\mathcal{D}$, (b) corruption parameter $\epsilon$, (c) corruption robust offline RL
       algorithm RobRL, (d) reference distribution $\mu_{\mathrm{ref}}$.
/* Estimate $\widehat{\theta}$ and build confidence interval $\Theta(\mathcal{D}_1)$ as in algorithm (3).  */
Initialize $\theta_0 \in \Theta(\mathcal{D}_1)$.
for $t = 0, \ldots, T-1$ do
    Let $g_t$ be the sub-gradient returned by running RobRL with reward parameter
    $r_h(s,a) = \phi(s,a)^\top \theta_{t,h}$ and dataset $\mathcal{D}_2$.
    $\theta_{t+1} = \mathrm{Proj}_{\Theta(\mathcal{D}_1)}\left(\theta_t - \eta\left(g_t + \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[\phi(\tau)]\right)\right)$
Set $\overline{\theta} = \frac{1}{T}\sum_{k=1}^{T}\theta_k$ and let $\widetilde{\pi}$ be the policy returned
  by running RobRL with reward function
  $r_h(s,a) = \phi(s,a)^\top \overline{\theta}_h$ and dataset $\mathcal{D}_2$.
return $\widetilde{\pi}$.
```

5 BOUNDED GENERALIZED COVERAGE RATIO

Algorithm 4 assumes access to a robust offline RL oracle `RobRL`, that given any reward parameter θ, returns an approximate sub-gradient of the optimal value function $V^\star(\theta) = \max_\pi V^\pi(\theta)$. Given such a first order oracle, it essentially performs a projected subgradient descent to determine an approximately optimal reward parameter $\overline{\theta}$, and the corresponding policy $\widetilde{\pi}$.

Theorem 5.1. *Suppose assumption 5 holds, $\sup_{p\in[0,1]}\left|\frac{d\sigma^{-1}(p)}{dp}\right| \leq \kappa$, and RobRL returns a $f(\varepsilon)$-robust estimate of the optimal value function, and $f(\varepsilon)$-approximate subgradient with norm at most G. If $N \geq \Omega\left(\frac{H^{3/2}dG}{f(\varepsilon)^2}\right)$, then with probability at least $1-\delta$, the following holds for any policy $\pi^\dagger$.*

$$V^{\pi^\dagger}(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \leq 2f(\varepsilon) + c\kappa\sqrt{\alpha}\left(\sqrt{\varepsilon H}d^{1/4} + \sqrt{d/N \cdot \log\left(HdN/\delta\right)}\right)$$

We now construct a corruption robust sub-gradient estimator of the function $V^\star(\theta) = \max_\pi V^\pi(\theta)$. Given a reward parameter $\theta = (\theta_1, \ldots, \theta_H)$, the optimal value function can be expressed as follows.

$$V^\star(\theta) = \max_{q=(q_1,\ldots,q_H)\in\mathcal{C}} \sum_{h=1}^{H} q_h^\top \Phi \theta_h.$$

Here q_h is the state, action occupancy measure at time step h, and the constraint set $\mathcal{C}$ ensures the Bellman flow constraints. Now from the definition of subgradient of a convex function which is expressed as a maximum of affine function (Nesterov (2018), chapter 3) we can write down the following expression of the sub-differential.

$$\delta_\theta V^\star(\theta) = \mathrm{co}\Big\{(\Phi^\top q_1, \dots, \Phi^\top q_H) : \\ (q_1, \dots, q_H) \in \underset{q=(q_1,\dots,q_H)\in\mathcal{C}}{\arg\max} \sum_{h=1}^{H} q_h^\top \Phi \theta_h \Big\}$$

Here $\mathrm{co}(S)$ is the convex-hull of a set S. Since q_h is the state, action occupancy measure at time-step h, $\Phi^\top q_h$ is the average feature observed at time-step h, and the result states that the subdifferential set is the convex hull of reward-maximizing average features. Therefore, we will construct a corruption robust offline RL method, that not only returns an approximately optimal policy but also the average feature under that policy. We make the following assumption.

Assumption 7 (Bounded Generalized Coverage Ratio)**.** *For a target policy $\pi^\star$, there exists $\nu > 0$ so that*

$$\mathbb{E}_{\tau\sim\pi^\star}[\phi(\tau)]^\top \left(\Sigma^{avg}_{\mu_0,\mu_1}\right)^{-2} \mathbb{E}_{\tau\sim\pi^\star}[\phi(\tau)] < \nu$$

We have stated the above assumption assuming $\Sigma^{\text{avg}}_{\mu_0,\mu_1}$ is invertible, but this is only for simplicity and consistency with prior literature. An alternate way to state this assumption would be that there exists a vector $y \in \mathbb{R}^d$ such that $\mathbb{E}_{(s,a)\sim\pi^\star}[\phi(s,a)] = \Sigma^{\text{avg}}_{\mu_0,\mu_1} y$ and $\|y\|_2^2 < \nu$.

Our method is based on the primal-dual framework of linear MDP and builds upon the recent work by Gabbianelli et al. (2024), who considered a similar assumption for discounted MDP. The standard linear program for a finite horizon linear MDP is the following optimization problem.

$$\begin{aligned} \max_q \ & \sum_{h=1}^{H} q_h^\top \Phi \theta_h \\ \text{s.t.} \ & \sum_a q_1(s,a) = \rho(s) \ \forall s \\ & E q_{h+1} = \boldsymbol{\mu}_h \Phi^\top q_h \ \forall h \in \{1, 2, \dots, H-1\} \\ & q_h \geq 0 \ \forall h \in [H] \end{aligned}$$

Here $\Phi \in \mathbb{R}^{SA\times d}$ is the feature matrix and the matrix $E \in \mathbb{R}^{S\times SA}$ is defined as $E[s,(s',a')] = \mathbb{1}\{s = s'\}$. The constraints specify Bellman-flow constraints at each time step h. We now substitute $\lambda_h = \Phi^\top q_h$ to the above LP formulation, with the interpretation that λ_h denotes the expected feature at time step h.

$$\begin{aligned} \max_{\substack{\{q_h : q_h \geq 0\}_{h=1}^H, \\ \{\lambda_h\}_{h=1}^H}} \ & \sum_{h=1}^{H} \lambda_h^\top \theta_h \\ \text{s.t.} \ & E q_1 = \rho, \ E q_{h+1} = \boldsymbol{\mu}_h \lambda_h \ \forall h \in [H-1] \\ & \lambda_h = \Phi^\top q_h \ \forall h \in [H] \end{aligned} \tag{6}$$

Note that this substitution doesn't change the optimal value of the LP and we aim to solve the optimization problem 6 instead of the original LP. The dual problem of eq. (6) is given as follows.

$$\begin{aligned} \min_{\{v_h\}_{h=1}^H, \{w_h\}_{h=1}^H} \ & \rho^\top v_1 \\ \text{s.t.} \ & E^\top v_h \geq \Phi w_h \ \forall h \in [H] \\ & w_h \geq \theta_h + \boldsymbol{\mu}_h^\top v_{h+1} \ \forall h \in [H-1] \\ & w_H \geq \theta_H \end{aligned} \tag{7}$$

Suppose $\mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w})$ is the Lagrangian corresponding to the optimization problem above. Then the main idea is to solve a saddle point of the Lagrangian i.e. $\max_{\boldsymbol{q},\boldsymbol{\lambda}} \min_{\boldsymbol{v},\boldsymbol{w}} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w})$ through a gradient descent-ascent based algorithm. However, $\boldsymbol{q}$ and $\boldsymbol{v}$ are infinite dimensional parameters. So we represent them symbolically in terms of $\boldsymbol{\lambda}$ and $\boldsymbol{w}$, and perform gradient descent-ascent steps over the $H \cdot d$ dimensional parameters $\boldsymbol{\lambda}$ and $\boldsymbol{w}$.

Note that, we don't exactly know the Lagrangian, and hence can only estimate the gradients through samples collected from the offline behavioral policy. However, recall that a ε-fraction of the data is corrupted, and hence we use robust mean to estimate the gradient from corrupted data. Additionally, as noted by Gabbianelli et al. (2024), computing estimates of the gradients require explicit knowledge of the feature covariance matrix $\Lambda_h = \mathbb{E}_{(s,a)\sim\mu^h_{\text{ref}}}\left[\phi(s,a)\phi(s,a)^\top\right]$. It turns out that a substitution $\lambda_h = \Lambda_h \beta_h$ lets us compute an estimate of the gradient without knowledge of the covariance matrix Λ_h. Hence we compute the saddle point of the Lagrangian $\mathcal{L}_R(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w}) \mid_{\{\lambda_h = \Lambda_h \beta_h\}_{h\in[H]}}$ through *robust* gradient descent-ascent method.

Once we obtain a solution $(\overline{\boldsymbol{\beta}}, \overline{\boldsymbol{w}})$, we choose policy $\overline{\pi}_h(a \mid s) \propto \exp\left(\phi(s,a)^\top \overline{w}_h\right)$ and set the primal solution λ_h as $\widehat{\Lambda}_h \overline{\beta}_h$. Here, $\widehat{\Lambda}_h$ is an estimate of the feature covariance matrix at step h. Since ε-fraction of our data is corrupted, we use robust covariance estimation to build $\widehat{\Lambda}_h$, and thereby obtain an approximate average features under $\overline{\pi}$. The full details of the algorithm is provided in the appendix (algorithm 7), and the next theorem provides the guarantees.

Theorem 5.2. *Suppose assumption (7) holds, and $N \geq \Omega\left(\frac{H^2 d^4 \nu^4}{\varepsilon^2}(\log^2 d + \log^2 A)\right)$. Then there is an algorithm that runs in time poly(H, d) and returns policy $\overline{\pi}$ and a vector $\widehat{v} = (\widehat{v}_1, \dots, \widehat{v}_H)$ s.t.*

$$\max_\pi V^\pi(\theta) - \mathbb{E}\left[V^{\overline{\pi}}(\theta)\right] \leq O\left(\nu\sqrt{\varepsilon} H^2 d^{3/2}\right), \textit{and}$$

$$V^\star(\theta') \geq V^\star(\theta) + \sum_{h=1}^{H} \langle \widehat{v}_h, \theta_h \rangle - O\left(\nu\sqrt{\varepsilon} H^2 d^{3/2}\right) \ \forall \theta'.$$

With such a first-order oracle, the next result states the improved guarantees given by algorithm 4.

Proposition 3. *Suppose assumptions (5) and (7) hold, and* $\sup_{p\in[0,1]}\left|\frac{d\sigma^{-1}(p)}{dp}\right| \le \kappa$*. If* $N \ge \widetilde{\Omega}\left(\frac{H^2 d^4 \nu^4}{\varepsilon^2}\right)$*, algorithm 4 returns a policy* $\widetilde{\pi}$ *so that the following holds.*

$$V^\star(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \le O\left(\nu\sqrt{\varepsilon}H^2 d^{3/2}\right)$$

6 SIMULATION RESULTS

We consider a 400 × 400 grid with five actions (`top`, `down`, `left`, `bottom`, and `random`). Selecting any one of the first four actions changes the current state in the intended direction with probability 0.75, and picks the other three directions with probability 0.25. The fifth action `random` selects one of the four directions uniformly at random. Since the state-space is huge, we generated $D(=40)$ dimensional features based for each (state, action) pair. This was done using the Random Fourier Features Rahimi and Recht (2007). Then we defined reward as $r_h(s,a) = \langle \phi(s,a), \theta_h\rangle$ where θ_h were randomly drawn from a multivariate Gaussian distribution, but fixed for the data generation. The horizon length was $H = 30$.

For data generation, we chose a pair of policies π_0, π_1, where π_0 is approximately optimal, and π_1 is sub-optimal. The policies were obtained through value iteration, and π_1 was obtained by stopping the value iteration method way before convergence. We generated $N = 20000$ pairs of tuples from the two policies, and generated pairwise preferences according to the logit model. We corrupted both the preference and the trajectory for an ϵ fraction of the data where $\epsilon \in \{0.05, 0.075, 0.1, 0.125, 0.15\}$.

Figure 1 shows the normalized sub-optimality gap $(V^\star - V^\pi)/V^\star$ under three different algorithms, and after averaging over 20 iterations. Here algorithms 1, 3, and 4 respectively correspond to the versions assuming full coverage, low relative condition number, and bounded coverage ratio. Algorithm 1 is supposed to perform better since the data satisfies full coverage. However, the other two algorithms also perform comparatively well, and we believe their performance can be further improved through hyper-parameter search.

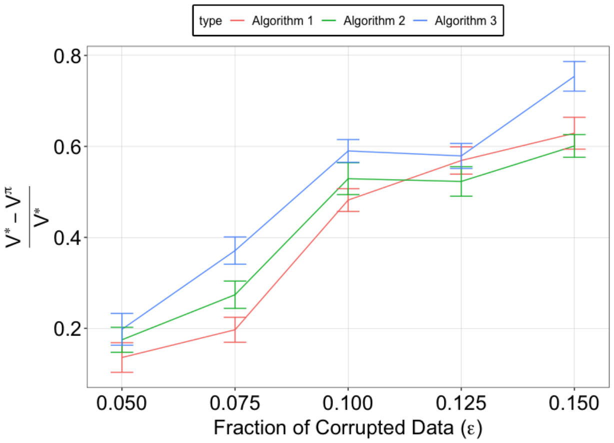

Figure 1: Performance of various corruption robust RLHF algorithms. For each value of ε (fraction of corrupted data) we show the normalized sub-optimality gap $(V^\star - V^\pi)/V^\star$.

7 CONCLUSION

We have designed corruption robust offline RLHF algorithms under different types of coverage assumptions. When uniform coverage holds, we can recover almost optimal dependency on the parameter ε. It is also possible to obtain an upper bound of $O(\sqrt{\varepsilon})$ under a substantially weaker assumption of bounded general coverage ratio. In the standard offline RL, the assumption of a low relative condition number is sufficient to obtain a dependence of $O(\sqrt{\varepsilon})$. As pointed out by Gabbianelli et al. (2024), these two assumptions are not directly comparable, and there is scope to explore the design of robust RLHF further.

In terms of future work, we have considered linear MDP in this work, and it would be interesting to consider non-convex reward functions or RLHF with general function approximation. However, such an extension is quite challenging. Algorithm 1 can be generalized by utilizing recent corruption robust RL under general function approximation Ye et al. (2023b), but we are not aware of similar results with weaker coverage assumptions. Furthermore, algorithms 3, and 4 crucially depend on the fact that $V^\star(\theta) = \max_\pi V^\pi(\theta)$ is convex in θ for linear MDPs, and in the presence of non-convex reward functions, we will require new proof techniques for gradient based methods.

Another interesting direction is to consider trajectory based rewards Zhan et al. (2023), which requires non-Markovian RL policies. In this case, the computation of optimal policy itself is a hard problem, and the design of corruption robust RLHF will require different approaches. Finally, we have provided preliminary simulation results considering a large grid-world and linear parametrization. It would be quite interesting but challenging to see the effects of data corruption in practical RLHF setting e.g. fine-tuning large language models.

Acknowledgements

The work of Andi Nika and Goran Radanovic was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Awasthi, P., Das, A., Kong, W., and Sen, R. (2022). Trimmed maximum likelihood estimation for robust generalized linear model. *Advances in Neural Information Processing Systems*, 35:862–873.

Bai, Y., Jones, A., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*.

Chatterji, N., Pacchiano, A., Bartlett, P., and Jordan, M. (2021). On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34:3401–3412.

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. (2022). Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR.

Chen, Y., Du, S., and Jamieson, K. (2021). Improved Corruption Robust Algorithms for Episodic Reinforcement Learning. In *International Conference on Machine Learning*, pages 1561–1570. PMLR.

Chhan, D., Novoseller, E., and Lawhern, V. J. (2024). Crowd-PrefRL: Preference-Based Reward Learning from Crowds. *arXiv preprint arXiv:2401.10941*.

Christiano, P. F., Leike, J., et al. (2017). Deep Reinforcement Learning from Human Preferences. In *NeurIPS*.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017). Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2004). Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*.

Gabbianelli, G., Neu, G., Papini, M., and Okolo, N. M. (2024). Offline primal-dual reinforcement learning for linear mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 3169–3177. PMLR.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

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

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823.

Lee, K., Smith, L., Dragan, A., and Abbeel, P. (2021). B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*.

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

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. (2022). When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR.

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

Manne, A. S. (1960). Linear programming and sequential decisions. *Management Science*, 6(3):259–267.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566.

Ouyang, L., Wu, J., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*.

Pacchiano, A., Saha, A., and Lee, J. (2021). Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. (2022). Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Rakhsha, A., Radanovic, G., Devidze, R., Zhu, X., and Singla, A. (2020). Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *International Conference on Machine Learning*, pages 7974–7984. PMLR.

Shin, D., Dragan, A. D., and Brown, D. S. (2023). Benchmarks and Algorithms for Offline Preference-Based Reward Learning. *Trans. Mach. Learn. Res.*, 2023.

Stiennon, N., Ouyang, L., et al. (2020). Learning to Summarize with Human Feedback. In *NeurIPS*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, Y., Liu, Q., and Jin, C. (2023). Is rlhf more difficult than standard rl? a theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wei, C.-Y., Dann, C., and Zimmert, J. (2022). A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR.

Wirth, C., Akrour, R., Neumann, G., Fürnkranz, J., et al. (2017). A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46.

Wu, F., Li, L., Xu, C., Zhang, H., Kailkhura, B., Kenthapadi, K., Zhao, D., and Li, B. (2022). Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *10th International Conference on Learning Representations, ICLR 2022*.

Xue, W., An, B., Yan, S., and Xu, Z. (2023). Reinforcement Learning from Diverse Human Preferences. *arXiv preprint arXiv:2301.11774*.

Yang, R., Zhong, H., Xu, J., Zhang, A., Zhang, C., Han, L., and Zhang, T. (2023). Towards robust offline reinforcement learning under diverse data corruption. *arXiv preprint arXiv:2310.12955*.

Ye, C., Xiong, W., Gu, Q., and Zhang, T. (2023a). Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pages 39834–39863. PMLR.

Ye, C., Yang, R., Gu, Q., and Zhang, T. (2023b). Corruption-robust offline reinforcement learning with general function approximation. *arXiv preprint arXiv:2310.14550*.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. (2023). Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2021). Robust Policy Gradient against Strong Data Corruption. In *International Conference on Machine Learning*, pages 12391–12401. PMLR.

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

Zhu, B., Jiao, J., and Jordan, M. I. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*.

Zhu, B., Jiao, J., and Steinhardt, J. (2022). Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256–2283.

Ziegler, D. M. et al. (2019). Fine-tuning Language Models from Human Preferences. *CoRR*, abs/1909.08593.

CHECKLIST

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

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator if your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Table of Contents

A MISSING PROOFS FROM SECTION 3

A.1 Convergence of Alternating Optimization

Proof. Let us write $H = \frac{1}{N\left\|\theta^\star - \hat{\theta}\right\|_2^2}(\theta^\star - \hat{\theta})^\top \sum_{n \in \widehat{S}} \nabla_\theta^2 \log \mathbb{P}_\theta(o^n \mid \mathrm{x}^n)(\theta^\star - \hat{\theta})$ be the second order derivative in the direction of $\theta^\star - \hat{\theta}$.

$$\begin{aligned}
|H| &\leq \frac{1}{N\left\|\theta^\star - \hat{\theta}\right\|_2^2} \sum_{n \in \widehat{S}} \left| (\theta^\star - \hat{\theta})^\top \frac{\exp(-o^n \cdot \theta^\top \mathrm{x}_n)}{(1 + \exp(-o^n \cdot \theta^\top \mathrm{x}_n))^2} \right. \\
&\qquad \left. \mathrm{x}_n \mathrm{x}_n^\top (\theta^\star - \hat{\theta}) \right| \\
&\leq \frac{1}{N\left\|\theta^\star - \hat{\theta}\right\|_2^2} \sum_{n \in \widehat{S}} \left| \mathrm{x}_n^\top (\theta^\star - \hat{\theta}) \right| \\
&\leq \frac{1}{N} \sum_{n \in \widehat{S}} \left\|\mathrm{x}_n\right\|_2^2 \leq L^2
\end{aligned}$$

The rest of the proof is very similar to the proof of Lemma A.12 of Awasthi et al. (2022). Let $\Delta = \frac{1}{N} \sum_{n \in \widehat{S}} \nabla_\theta \log \mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)^\top \frac{(\theta^\star - \hat{\theta})}{\left\|\theta^\star - \hat{\theta}\right\|_2}$. Writing $F(\theta) = \frac{1}{N} \sum_{n \in \widehat{S}} \log \mathbb{P}_\theta(o^n \mid \mathrm{x}_n)$, we get that there exists θ' such that

$$F(\theta') \leq F(\hat{\theta}) - \frac{\Delta^2}{4L^2}$$

Suppose $\|\theta'\|$ is feasible. Then it must be that $\eta \geq \frac{\Delta^2}{4L^2}$ as it is impossible to make improvement more than η. This implies that $\Delta \leq 2L\sqrt{\eta}$. On the other hand, if θ' is not a feasible solution, then we use the fact that $F(\cdot)$ is a concave function and obtain the following bound.

$$F(\theta^\star) \leq F(\hat{\theta}) + \nabla_\theta F(\hat{\theta})^\top (\theta^\star - \hat{\theta}) = F(\hat{\theta}) + \Delta \left\|\theta^\star - \hat{\theta}\right\|_2$$

Then it must be that $\eta \geq \Delta \left\|\theta^\star - \hat{\theta}\right\|_2$ or $\Delta \leq \frac{\eta}{\left\|\theta^\star - \hat{\theta}\right\|_2}$. Combining the two results and after substituting $\eta = \varepsilon^2$ we get $\Delta \leq \max\left\{2L\varepsilon, \frac{\varepsilon^2}{\left\|\theta^\star - \hat{\theta}\right\|_2}\right\}$. □

A.2 Proof of Theorem 3.1

Proof. By Lemma 3 the reward estimate $\widehat{\theta}$ is $C_1 \frac{\varepsilon}{\zeta} e^{\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}}$ close to the true parameter $\theta^\star$. Since algorithm `RobRL` returns at least $f(\varepsilon)$ optimal policy in terms of value function we are guaranteed that $V^\star(\widehat{\theta}) \geq V^{\widetilde{\pi}}(\widehat{\theta}) \geq V^\star(\widehat{\theta}) - f(\varepsilon)$ for any θ. Using this result we can lower bound $V^{\widetilde{\pi}}(\theta^\star)$.

$$\begin{aligned} V^\star(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) &= V^\star(\theta^\star) - V^{\widetilde{\pi}}(\overline{\theta}) + V^{\widetilde{\pi}}(\overline{\theta}) - V^{\widetilde{\pi}}(\theta^\star) \\ &\leq f(\varepsilon) + V^\star(\theta^\star) - V^\star(\widehat{\theta}) + V^{\widetilde{\pi}}(\overline{\theta}) - V^{\widetilde{\pi}}(\theta^\star) \end{aligned}$$

For the first difference, we use the fact that the optimal value function $V^\star(\cdot)$ is $\sqrt{Hd}$-Lipschitz in the reward parameter (lemma (6)) and obtain the following bound.

$$V^\star(\theta^\star) - V^\star(\widehat{\theta}) \leq \sqrt{Hd}\left\|\theta^\star - \widehat{\theta}\right\|_2 \leq \sqrt{Hd} C_1 \frac{\varepsilon}{\zeta} \exp\left(\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\right)$$

Using lemma (3) the second difference can be bounded as follows.

$$\begin{aligned} V^{\widetilde{\pi}}(\overline{\theta}) - V^{\widetilde{\pi}}(\theta^\star) &= \sum_{h=1}^{H} \sum_{s,a} \mathbb{P}_{\widetilde{\pi}}(s_h = s, a_h = a) \phi(s,a)^\top \left(\overline{\theta}_h - \theta^\star_h\right) \\ &\leq \sum_{h=1}^{H} \sum_{s,a} \mathbb{P}_{\widetilde{\pi}}(s_h = s, a_h = a) \left\|\phi(s,a)\right\|_2 \left\|\overline{\theta}_h - \theta^\star_h\right\|_2 \\ &\leq \sum_{h=1}^{H} \left\|\overline{\theta}_h - \theta^\star_h\right\|_2 \\ &\leq \sqrt{H} \sqrt{\sum_{h=1}^{H} \left\|\overline{\theta}_h - \theta^\star_h\right\|_2^2} \\ &= \sqrt{H} \left\|\overline{\theta} - \theta^\star\right\|_2^2 \\ &\leq \sqrt{H} C_1 \frac{\varepsilon}{\xi} \cdot \exp\left(\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\right) \end{aligned}$$

□

Lemma 3. *Suppose assumption (3) holds with $\xi \geq 5\varepsilon$ and $N \geq \Omega\left(\frac{H^{3/2}}{\varepsilon^2}\left(d + \log(1/\delta)\right)\right)$. Then algorithm (2) returns $\widehat{\theta}$, so that with probability at least $1-\delta$, we have*

$$\|\widehat{\theta} - \theta^\star\|_2 \leq C_1 \frac{\varepsilon}{\xi} \exp\left(\sqrt{\log\left(1/2\delta\varepsilon\right)}\right)$$

Proof. From Lemma 1 we know that algorithm (2) computes a $\gamma = \max\left\{2L\varepsilon, \frac{\varepsilon^2}{\left\|\theta^\star - \widehat{\theta}\right\|_2}\right\}$ stationary point. We can assume that $2L\varepsilon \geq \frac{\varepsilon^2}{\left\|\theta^\star - \widehat{\theta}\right\|_2}$. Otherwise, $\left\|\theta^\star - \widehat{\theta}\right\|_2 \leq \varepsilon/2L$ and we are done.

Let T be the set of uncorrupted samples and E be the set of corrupted samples. Then we can write down the stationarity condition (5) as follows.

$$\frac{1}{N} \sum_{i \in \widehat{S} \cap E} \nabla_\theta \log \mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)^\top \left(\widehat{\theta} - \theta^\star\right) \leq 2L\varepsilon \cdot \left\|\widehat{\theta} - \theta^\star\right\|_2 - \frac{1}{N} \sum_{i \in \widehat{S} \cap T} \nabla_\theta \log \mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)^\top \left(\widehat{\theta} - \theta^\star\right) \qquad (8)$$

We first upper bound the term on the right.

$$-\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\nabla_\theta \log \mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)^\top \left(\widehat{\theta}-\theta^\star\right) = \underbrace{-\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\nabla_\theta \log \mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)^\top \left(\widehat{\theta}-\theta^\star\right)}_{:=T_1}$$
$$+ \underbrace{\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\left(\nabla_\theta \log \mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n) - \nabla_\theta \log \mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)\right)^\top \left(\widehat{\theta}-\theta^\star\right)}_{:=T_2} \tag{9}$$

Using the functional form of sigmoid link function i.e. $\mathbb{P}_\theta(o \mid \mathrm{x}) = \frac{1}{1+\exp(-o\cdot\theta^\top \mathrm{x})}$, we get the following expression for the term T_1.

$$T_1 = -\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\frac{o^n}{1+\exp(o^n\cdot\langle\theta^\star,\mathrm{x}_n\rangle)}\mathrm{x}_n^\top(\widehat{\theta}-\theta^\star)$$

In order to provide a high probability bound on T_1, we first provide a bound on the k-th moment of the random vector $X = \frac{o}{1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle)}\mathrm{x}$. For any unit vector $v\in\mathbb{R}^d$ with $\|v\|_2 = 1$ we have,

$$\mathbb{E}\left[\left(\frac{o}{1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle)}\right)^k \left(\mathrm{x}^\top v\right)^k\right] \le \sqrt{\mathbb{E}\left[\frac{o^{2k}}{(1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle))^{2k}}\right]}\sqrt{\mathbb{E}\left[(\mathrm{x}^\top v)^{2k}\right]}$$
$$\le \sqrt{\mathbb{E}\left[\frac{1}{(1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle))^{2k}}\right]}L^k \le L^k$$

The second inequality uses the fact that $o\in\{-1,1\}$ and $\|\mathrm{x}\|_2 \le L$. Since $\widehat{S}\cap T$ contains uncorrupted samples, and $\left|\widehat{S}\cap T\right| \ge (1-2\varepsilon)N$ we can use Corollary G.1 from Zhu et al. (2022) to obtain the following result with probability at least $1-\delta$.

$$\left\|\mathbb{E}\left[\frac{o}{1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle)}\mathrm{x}\right] - \frac{1}{\left|\widehat{S}\cap T\right|}\sum_{i\in\widehat{S}\cap T}\frac{o^n}{1+\exp(o^n\cdot\langle\theta^\star,\mathrm{x}_n\rangle)}\mathrm{x}_n\right\|_2 \le \frac{CkL}{1-2\varepsilon}\left(\frac{(2\varepsilon)^{1-1/k}}{\delta^{1/k}} + \frac{1}{\delta}\sqrt{\frac{d}{N}}\right)$$

Now substituting $k = \sqrt{\log(\frac{1}{2\delta\varepsilon})}$ and assuming $N \ge d/\varepsilon^2$ we obtain the following result.

$$\frac{1}{\left|\widehat{S}\cap T\right|}\sum_{i\in\widehat{S}\cap T}\frac{o^n}{1+\exp(o^n\cdot\langle\theta^\star,\mathrm{x}_n\rangle)}\mathrm{x}_n = \mathbb{E}\left[\frac{o}{1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle)}\mathrm{x}\right] + \Delta$$

where

$$\|\Delta\|_2 \le \frac{4\varepsilon CL}{1-2\varepsilon}\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\left(\frac{1}{2\delta\varepsilon}\right)^{1/\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}} \le \frac{C_1\varepsilon L}{1-2\varepsilon}e^{\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}}$$

for some constant $C_1 > 0$. This lets us derive the following upper bound on T_1.

$$\begin{aligned}
T_1 &= -\frac{\left|\widehat{S}\cap T\right|}{N}\left(\mathbb{E}\left[\frac{o}{1+\exp(o\cdot\langle\theta^\star,\mathrm{x}\rangle)}\mathrm{x}\right] + \Delta\right)^\top\left(\widehat{\theta}-\theta^\star\right)\\
&= -\frac{\left|\widehat{S}\cap T\right|}{N}\left(\mathbb{E}_{\mathrm{x},o}\left[\nabla_\theta \log \mathbb{P}_{\theta^\star}(o \mid \mathrm{x})\right] + \Delta\right)^\top\left(\widehat{\theta}-\theta^\star\right)\\
&= -\frac{\left|\widehat{S}\cap T\right|}{N}\Delta^\top\left(\widehat{\theta}-\theta^\star\right)\\
&\le C_1\varepsilon L\exp\left(\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\right)\left\|\widehat{\theta}-\theta^\star\right\|_2
\end{aligned}$$

The second equality uses that the fact $\theta^\star$ optimizes the population logistic loss and hence the derivative is zero. The last inequality uses that $\left|\widehat{S}\cap T\right| \geq (1-2\varepsilon)N$.

We now bound the term T_2 defined in eq. (9). We use assumption (3) to show that the function $\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\nabla_\theta \log \mathbb{P}_\theta(o^n \mid \mathrm{x}_n)$ is strongly concave in θ. Indeed from the definition of $\mathbb{P}_\theta(o \mid \mathrm{x})$ we have the following result.

$$\begin{aligned}
\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\nabla_\theta^2 \log \mathbb{P}_\theta(o^n \mid \mathrm{x}_n) &= \frac{1}{N}\sum_{n\in\widehat{S}\cap T} -\frac{\exp(o^n\left\langle\theta,\mathrm{x}\right\rangle)}{(1+\exp(o^n\left\langle\theta,\mathrm{x}\right\rangle))^2}\mathrm{x}_n\mathrm{x}_n^\top \\
&= -\frac{1}{N}\sum_{n\in\widehat{S}\cap T}\frac{1}{(\exp(-o^n\left\langle\theta,\mathrm{x}\right\rangle/2)+\exp(o^n\left\langle\theta,\mathrm{x}\right\rangle/2))^2}\mathrm{x}_n\mathrm{x}_n^\top \\
&\preccurlyeq -\frac{1}{4N}\sum_{n\in\widehat{S}\cap T}\mathrm{x}_n\mathrm{x}_n^\top \\
&= -\frac{1}{4N}\left(\sum_{n=1}^N \mathrm{x}_n\mathrm{x}_n^\top - \sum_{n\notin\widehat{S}\cap T}\mathrm{x}_n\mathrm{x}_n^\top\right) \\
&\preccurlyeq -\frac{1}{4}\mathbb{E}\left[\mathrm{x}\mathrm{x}^\top\right] + c_1 L^2\sqrt{\frac{d+\log(1/\delta)}{N}}\cdot \mathrm{Id}_d + \frac{1}{2N}\sum_{n\notin\widehat{S}\cap T}\mathrm{x}_n\mathrm{x}_n^\top
\end{aligned}$$

The first inequality follows from the observation that $e^u + e^{-u} \geq 2$. The last inequality uses the concentration bound of a sample covariance matrix (lemma 7). For the third term in the last upper bound, note that $\left|\widehat{S}\cap T\right| \leq \varepsilon N$ and the L_2-norm of a feature is bounded by L. This implies that the last term is at most $\varepsilon L/2$. Now using assumption (3) and choosing $N \geq \frac{4c_1^2 L^3}{\varepsilon^2}\left(d+\log(1/\delta)\right)$ we obtain the following upper bound.

$$\frac{1}{N}\sum_{i\in\widehat{S}\cap T}\nabla_\theta^2 \log \mathbb{P}_\theta(o^n \mid \mathrm{x}_n) \preccurlyeq -\left(\frac{\xi}{4}-\varepsilon\right)L$$

Therefore, we get the following upper bound.

$$T_2 := \frac{1}{N}\sum_{i\in\widehat{S}\cap T}\left(\nabla_\theta \log \mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n) - \nabla_\theta \log \mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)\right)^\top\left(\widehat{\theta}-\theta^\star\right) \leq -\left(\frac{\xi}{4}-\varepsilon\right)L\left\|\widehat{\theta}-\theta^\star\right\|_2^2$$

This gives us the following upper bound on the right hand side of eq. (8).

$$-\left(\frac{\xi}{4}-\varepsilon\right)L\left\|\widehat{\theta}-\theta^\star\right\|_2^2 + \left(2 + C_1\exp\left(\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\right)\right)\varepsilon L\left\|\widehat{\theta}-\theta^\star\right\|_2 \tag{10}$$

We now provide a lower bound on the left hand side of eq. (8). From the definition of $\mathbb{P}_\theta(o \mid \mathrm{x})$ we obtain the following identity.

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in\widehat{S}\cap E}\nabla_\theta \log \mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)^\top\left(\widehat{\theta}-\theta^\star\right) = \frac{1}{N}\sum_{i\in\widehat{S}\cap E}\frac{o^n}{1+\exp\left(o^n\cdot\left\langle\widehat{\theta},\mathrm{x}_n\right\rangle\right)}\mathrm{x}_n^\top\left(\widehat{\theta}-\theta^\star\right) \\
&= \frac{1}{N}\sum_{i\in\widehat{S}\cap E}\left(1-\frac{1}{1+\exp\left(-o^n\cdot\left\langle\widehat{\theta},\mathrm{x}_n\right\rangle\right)}\right)o^n\cdot\mathrm{x}_n^\top\left(\widehat{\theta}-\theta^\star\right) \\
&= \frac{1}{N}\sum_{i\in\widehat{S}\cap E}\left(1-\mathbb{P}_{\widehat{\theta}}(o^n \mid \mathrm{x}_n)\right)o^n\cdot\mathrm{x}_n^\top\left(\widehat{\theta}-\theta^\star\right) \\
&\geq -\frac{1}{N}\sum_{i\in\widehat{S}\cap E}\left\|\mathrm{x}_n\right\|_2\left\|\widehat{\theta}-\theta^\star\right\|_2 \geq -\varepsilon L\left\|\widehat{\theta}-\theta^\star\right\|_2
\end{aligned}$$

The last inequality uses $\left|\widehat{S} \cap E\right| \leq \varepsilon N$. Now combining this lower bound with the upper bound established in eq. (10) we can obtain a bound on $\left\|\widehat{\theta} - \theta^\star\right\|_2$.

$$\begin{aligned}
&-\varepsilon L \left\|\widehat{\theta} - \theta^\star\right\|_2 \leq -\left(\frac{\xi}{4} - \varepsilon\right) L \left\|\widehat{\theta} - \theta^\star\right\|_2^2 + \left(2 + C_1 \exp\left(\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\right)\right) \varepsilon L \left\|\widehat{\theta} - \theta^\star\right\|_2 \\
&\Rightarrow \left\|\widehat{\theta} - \theta^\star\right\|_2 \leq \frac{3 + C_1 \exp\left(\sqrt{\log\left(\frac{1}{2\delta\varepsilon}\right)}\right)}{\xi/4 - \varepsilon} \cdot \varepsilon
\end{aligned}$$

□

B MISSING PROOFS FROM SECTION 4

Here we state a more general version of Lemma 2. Let us write $\mathbb{P}_\theta(o \mid \mathrm{x}) = \frac{1}{1+\exp(-o \cdot \mathrm{x}^\top \theta)}$. We will also write $\theta_N^\star$ to denote the parameter that maximizes empirical log-likelihood i.e.

$$\theta_N^\star \in \underset{\theta : \|\theta\|_2 \leq 1}{\arg\max} \frac{1}{N} \sum_n \log \mathbb{P}_\theta(o^n \mid x_n)$$

Lemma 4. *Suppose that $\|\theta\|_2 \leq B$ for any $\theta \in \Theta_B$, $\|\phi(\tau)\|_2 \leq L$ for any trajectory $\tau \in \mathcal{T}$, and $\log \mathbb{P}_\theta(\cdot)$ is a concave function of θ. Then with probability at least $1 - \delta$, we have*

$$\frac{1}{N} \sum_{n=1}^N \log\left(\frac{\mathbb{P}_{\widetilde{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)}\right) \leq 6\varepsilon L B + c \cdot \frac{d}{N} \log\left(\frac{LN}{\delta}\right)$$

for $\widetilde{\theta} = \widehat{\theta}$ or $\theta_N^\star$. Here $c > 0$ is a universal constant.

Proof. First note that we can express the difference in log-likelihood as follows.

$$\begin{aligned}
&\frac{1}{N} \sum_{n=1}^N \log \mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n) - \log \mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n) \\
&= \frac{1}{N} \sum_{n=1}^N \log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta_N^\star}(o^n \mid \mathrm{x}_n)}\right) + \frac{1}{N} \sum_{n=1}^N \log\left(\frac{\mathbb{P}_{\theta_N^\star}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)}\right)
\end{aligned} \tag{11}$$

For linear reward functions, we can use Lemma 1 of Zhan et al. (2023) to bound the second term. Let $T \subseteq [N]$ be the set of corrupted data points. Then we have,

$$\begin{aligned}
&\frac{1}{N} \sum_{n=1}^N \log\left(\frac{\mathbb{P}_{\theta_N^\star}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)}\right) \\
&= \frac{1}{N} \sum_{n \in T} \log\left(\frac{\mathbb{P}_{\theta_N^\star}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)}\right) + \frac{1}{N} \sum_{n \notin T} \log\left(\frac{\mathbb{P}_{\theta_N^\star}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star}(o^n \mid \mathrm{x}_n)}\right) \\
&\leq \varepsilon \cdot \log\left(\frac{1 + e^{LB}}{1 - e^{-LB}}\right) + O\left(\frac{d}{(1-\varepsilon)N} \log\left(\frac{LN}{\delta}\right)\right) \\
&\leq 2\varepsilon L B + O\left(\frac{d}{N} \log\left(\frac{LN}{\delta}\right)\right)
\end{aligned}$$

The first inequality uses Lemma 1 of Zhan et al. (2023) and $|T| \leq \varepsilon N$. Now, we consider bounding the first term in eq. (11).

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star_N}(o^n \mid \mathrm{x}_n)}\right) && (12)\\
&= \frac{1}{N}\sum_{n\notin\widehat{S}}\log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star_N}(o^n \mid \mathrm{x}_n)}\right) + \frac{1}{N}\sum_{n\in\widehat{S}}\log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star_N}(o^n \mid \mathrm{x}_n)}\right)\\
&\le \varepsilon\cdot\log\left(\frac{1+e^{LB}}{1-e^{-LB}}\right) + \frac{1}{N}\sum_{n\in\widehat{S}}\log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)}{\mathbb{P}_{\theta^\star_N}(o^n \mid \mathrm{x}_n)}\right)\\
&\le 2\varepsilon LB + \frac{1}{N}\sum_{n\in\widehat{S}}\log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n \mid x_n)}{\mathbb{P}_{\theta^\star_N}(o^n \mid \mathrm{x}_n)}\right)\\
&\le 2\varepsilon LB + \frac{1}{N}\sum_{n\in\widehat{S}}\nabla_\theta \log\mathbb{P}_{\hat{\theta}}(o^n \mid \mathrm{x}_n)^\top(\theta^\star - \hat{\theta})\\
&\le 2\varepsilon LB + \gamma\left\|\theta^\star_N - \widehat{\theta}\right\|_2 && (13)
\end{aligned}
$$

The first inequality uses that the size of $\widehat{S}$ is $(1-\varepsilon)N$ and the inner product between the parameter and the feature is bounded by LB. The second inequality uses that $\log\mathbb{P}_\theta(o^n \mid \mathrm{x}_n)$ is a concave function in θ.

Lemma 1 shows that $\gamma \le \max\left\{2L\varepsilon, \frac{\varepsilon^2}{\left\|\theta^\star_N - \hat{\theta}\right\|_2}\right\}$. Substituting this upper bound in eq. (13) and using $\|\theta^\star_N\|_2, \left\|\hat{\theta}\right\|_2 \le B$ we get the following result: $\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{\mathbb{P}_{\hat{\theta}}(o^n|\mathrm{x}_n)}{\mathbb{P}_{\theta^\star_N}(o^n|\mathrm{x}_n)}\right) \le \max\left\{4\varepsilon LB, 2\varepsilon LB + \varepsilon^2\right\} \le 4\varepsilon LB.$ □

B.1 Proof of Theorem 4.1

Proof. Given a reward parameter θ let $V^\star(\theta) = \max_\pi V^\pi(\theta)$ be the optimal value function with reward parameter θ. We claim that $V^\star(\cdot)$ is a convex function. In order to see this, given a policy π let d be the corresponding occupancy measure i.e. $d_h(s,a) = \mathbb{P}_\pi(s_h = s, a_h = a)$. Then we can write the value function as $V^\pi(\theta) = \sum_{h,s,a} = d_h(s,a)\phi(s,a)^\top\theta = d^\top\Phi\theta$. This observation implies the following inequality.

$$\max_\pi V^\pi(\theta) \le \max_d d^\top\Phi\theta \tag{14}$$

On the other hand, given an occupancy measure d one can consider the following policy.

$$\pi^d_h(s,a) = \begin{cases} \frac{d_h(s,a)}{\sum_b d_h(s,b)} & \text{if } \sum_b d_h(s,b) > 0\\ \frac{1}{A} & \text{o.w.}\end{cases}$$

Moreover, it is known that occupancy measure induced by $\pi^d = (\pi^d_1,\ldots,\pi^d_H)$ is d. This implies the following inequality.

$$\max_\pi V^\pi(\theta) \ge \max_d d^\top\Phi\theta \tag{15}$$

Therefore, from equations (15) and (14) we conclude that

$$V^\star(\theta) = \max_\pi V^\pi(\theta) = \max_d d^\top\Phi\theta$$

Since $V^\star(\cdot)$ is a maximum of linear functions, it is a convex function. Moreover, by lemma (6) $V^\star(\cdot)$ is $\sqrt{Hd}$-Lipschitz. By a similar argument the function $\mathcal{R}(\theta) = \mathbb{E}_{\tau\sim\mu_{\text{ref}}}\left[\phi(\tau)^\top\theta\right]$ is $\sqrt{Hd}$-Lipshitz in θ. Therefore, $V^\star(\cdot) - \mathcal{R}(\cdot)$ is $2\sqrt{Hd}$-Lipschitz function.

Now observe that, algorithm (3) performs a projected sub-gradient descent of the function $V^\star(\cdot) - \mathcal{R}(\cdot)$ with biased zero oracle calls. In particular, since `RobRL` returns a $f(\varepsilon)$-robust estimate of the optimal value function, we are guaranteed that $\left|\widehat{V}(\theta) - V^\star(\theta)\right| \le f(\varepsilon)$. Therefore, we can apply the result of Theorem D.1 to obtain the following bound.

$$V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - \min_\theta\left(V^\star(\theta) - \mathcal{R}(\theta)\right) \le 5\sqrt{2f(\varepsilon)}(Hd)^{1/4}$$

Note that in order to apply theorem Theorem D.1, we need a lower bound on the number of iterations (T) and the number of calls to zero-order oracle (K) per iteration. For linear MDP we have the maximum norm of the parameter, $D \le \sqrt{Hd}$ and maximum value of the function $M \le H\sqrt{d}$. This implies the following lower bound on the number of samples.

$$N \ge T \cdot K \ge \widetilde{\Omega}\left(\frac{MD}{\varepsilon}\frac{M^2 d^3}{\varepsilon^2}\right) = \widetilde{\Omega}\left(\frac{H^{3/2} d^5}{\varepsilon^3}\right)$$

Since $\widetilde{\pi}$ is $f(\varepsilon)$-approximately optimal with respect to the reward parameter $\overline{\theta}$ we are guaranteed that,

$$V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - f(\varepsilon) \le V^{\widetilde{\pi}}(\overline{\theta}) - \mathcal{R}(\overline{\theta}) \le V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) + 8\sqrt{f(\varepsilon)}(Hd)^{1/4}. \tag{16}$$

Now using lemma (8) (i.e. $\min_\theta \max_\pi V^\pi(\theta) - \mathcal{R}(\theta) = \max_\pi \min_\theta V^\pi(\theta) - \mathcal{R}(\theta)$ for linear reward models) we obtain the following inequality.

$$\begin{aligned}
\max_\pi \min_\theta \left(V^\pi(\theta) - \mathcal{R}(\theta)\right) - f(\varepsilon) = \min_\theta \max_\pi \left(V^\pi(\theta) - \mathcal{R}(\theta)\right) &\le V^{\widetilde{\pi}}(\overline{\theta}) \le V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - f(\varepsilon) \\
&\le V^{\widetilde{\pi}}(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - f(\varepsilon) \\
&\le \max_\pi \min_\theta \left(V^\pi(\theta) - \mathcal{R}(\theta)\right) + 8\sqrt{f(\varepsilon)}(Hd)^{1/4}
\end{aligned} \tag{17}$$

We claim that this implies that $\widetilde{\pi}$ approximately optimizes the objective $\max_\pi \min_\theta V^\pi(\theta) - \mathcal{R}(\theta)$ i.e.

$$\min_\theta \left(V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta)\right) \ge \max_\pi \min_\theta \left(V^\pi(\theta) - \mathcal{R}(\theta)\right) - f(\varepsilon) - 8\sqrt{f(\varepsilon)}(Hd)^{1/4} \tag{18}$$

Let $(\pi^\star, \theta^\star)$ be an optimal solution of the optimization problem $\max_\pi \min_\theta V^\pi(\theta) - \mathcal{R}(\theta)$. Then the observation above follows from the following set of inequalities.

$$\begin{aligned}
&\min_\theta \left(V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta)\right) - \max_\pi \min_\theta \left(V^\pi(\theta) - \mathcal{R}(\theta)\right) \\
&= \min_\theta \left(V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta)\right) - \min_\theta \left(V^{\pi^\star}(\theta) - \mathcal{R}(\theta)\right) \\
&\ge -\min_\theta \left|V^{\widetilde{\pi}}(\theta) - V^{\pi^\star}(\theta)\right| \\
&= -\min_\theta \left|\left(V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta)\right) - \left(V^{\pi^\star}(\theta) - \mathcal{R}(\theta)\right)\right| \\
&\ge \underbrace{-\min_\theta \left|\left(V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta)\right) - \left(V^{\widetilde{\pi}}(\overline{\theta}) - \mathcal{R}(\overline{\theta})\right)\right|}_{:=T_1} - \underbrace{\min_\theta \left|\left(V^{\widetilde{\pi}}(\overline{\theta}) - \mathcal{R}(\overline{\theta})\right) - \left(V^{\pi^\star}(\theta) - \mathcal{R}(\theta)\right)\right|}_{:=T_2} \\
&\ge -\left|\left(V^{\widetilde{\pi}}(\overline{\theta}) - \mathcal{R}(\overline{\theta})\right) - \left(V^{\pi^\star}(\theta^\star) - \mathcal{R}(\theta^\star)\right)\right| \\
&\ge -f(\varepsilon) - 8\sqrt{f(\varepsilon)}(Hd)^{1/4}
\end{aligned}$$

The first inequality follows since $\min_\theta V^{\pi^\star}(\theta) \le \min_\theta \left|V^{\pi^\star}(\theta) - V^{\widetilde{\pi}}(\theta)\right| + V^{\widetilde{\pi}}(\theta) \le \min_\theta \left|V^{\pi^\star}(\theta) - V^{\widetilde{\pi}}(\theta)\right| + \min_\theta V^{\widetilde{\pi}}(\theta)$. The third inequality follows by substituting $\theta = \overline{\theta}$ in the term T_1 and $\theta = \overline{\theta}$ in the term T_2. Finally, the last inequality uses eq. (17). Now we can apply lemma (2) with $\eta = f(\varepsilon) + 8\sqrt{f(\varepsilon)}(Hd)^{1/4}$ to complete the proof. □

B.2 Proof of Proposition 2

Proof. For linear MDP, the parameter $\theta = [\theta_1; \theta_2; \ldots; \theta_H]$ and the feature of a trajectory τ is constructed by concatenating the features of H state, action pairs. Therefore, $\|\theta\|_2 \le \sqrt{Hd}$ and $\|\phi(\tau)\|_2 \le \sqrt{H}$ for any trajectory τ. So we substitute $L = \sqrt{H}$,$B = \sqrt{Hd}$, and $M \le LB = H\sqrt{d}$.

We will use `R-LSVI` from Zhang et al. (2022) as the corruption robust offline RL oracle `RobRL`. Note that if $N \ge \Omega(H \cdot \mathrm{poly}(d)/\varepsilon)$ we have $f(\varepsilon) \le \widetilde{O}\left(H^2 d\sqrt{\alpha\varepsilon}\right)$. Now using the upper bound provided in theorem (4.1) we obtain the following bound.

$$
\begin{aligned}
&V^{\star}(\theta^{\star}) - V^{\widetilde{\pi}}(\theta^{\star}) \\
&\le O\left(\kappa\sqrt{\alpha}\left(\sqrt{\varepsilon}H d^{1/4} + \sqrt{\frac{d}{N}\log\left(\frac{HdN}{\delta}\right)}\right)\right) \\
&+ \widetilde{O}(H^2 d\kappa\sqrt{\alpha\varepsilon}) + \widetilde{O}\left(H^{5/4}d^{3/4}(\alpha\varepsilon)^{1/4}\right)
\end{aligned}
$$

Now observe that if $N \ge \Omega(H \cdot \mathrm{poly}(d)/\varepsilon)$ the term $\widetilde{O}(\sqrt{d/N})$ can be bounded by $O(\sqrt{\varepsilon})$. Finally, we need a lower bound of $N \ge \widetilde{\Omega}\left(\frac{H^{3/2}d^5}{\varepsilon^3}\right)$ in order to apply theorem (4.1). □

Lemma 5. *Suppose assumption (5) holds, and* $\sup_{p\in[0,1]}\left|\frac{d\Phi^{-1}(p)}{dp}\right| \le \kappa$. *Let* π *be a policy so that*

$$
\min_{\theta\in\Theta(\widehat{\mathcal{D}}_1)}\left(V^{\widetilde{\pi}}(\theta) - \mathbb{E}_{\tau\sim\mu_{ref}}\left[\phi(\tau)^\top\theta\right]\right) \ge \max_{\pi}\min_{\theta\in\Theta(\widehat{\mathcal{D}}_1)}\left(V^{\pi}(\theta) - \mathbb{E}_{\tau\sim\mu_{ref}}\left[\phi(\tau)^\top\theta\right]\right) - \eta
$$

then for any target policy $\pi^\dagger$, *with probability at least* $1-\delta$, *we have*

$$
V^{\pi^\dagger}(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) \le c\kappa\sqrt{\alpha}\left(\sqrt{\varepsilon}H d^{1/4} + \sqrt{\frac{d}{N}\log\left(\frac{HdN}{\delta}\right)}\right) + \eta.
$$

Proof. The proof follows a similar approach to the proof of theorem 1 in Zhan et al. (2023), except for the fact that we need to account for the approximation error η and corrupted dataset. We will write $\mathcal{R}(\theta) = \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}\left[\phi(\tau)^\top\theta\right]$. Moreover, let $\theta^\dagger \in \arg\min_{\theta\in\Theta(\widehat{\mathcal{D}}_1)} V^{\pi^\dagger}(\theta) - \mathcal{R}(\theta)$.

$$
\begin{aligned}
V^{\pi^\dagger}(\theta^\star) - V^{\widetilde{\pi}}(\theta^\star) &= \left(V^{\pi^\dagger}(\theta^\star) - \mathcal{R}(\theta^\star)\right) - \left(V^{\widetilde{\pi}}(\theta^\star) - \mathcal{R}(\theta^\star)\right) \\
&\le \left(V^{\pi^\dagger}(\theta^\star) - \mathcal{R}(\theta^\star)\right) - \left(V^{\pi^\dagger}(\theta^\dagger) - \mathcal{R}(\theta^\dagger)\right) + \eta \\
&= \mathbb{E}_{\substack{\tau\sim\mu^{\pi^\dagger} \\ \tau_0\sim\mu_{\mathrm{ref}}}}\left[(\phi(\tau) - \phi(\tau_0))^\top(\theta^\star - \theta^\dagger)\right] \\
&\le \mathbb{E}_{\substack{\tau\sim\mu^{\pi^\dagger} \\ \tau_0\sim\mu_{\mathrm{ref}}}}\left[\left|(\phi(\tau) - \phi(\tau_0))^\top(\theta^\star - \theta^\dagger)\right|\right] + \eta \\
&\le \sqrt{\mathbb{E}_{\substack{\tau\sim\mu^{\pi^\dagger} \\ \tau_0\sim\mu_{\mathrm{ref}}}}\left[(\theta^\star - \theta^\dagger)^\top(\phi(\tau) - \phi(\tau_0))(\phi(\tau) - \phi(\tau_0))^\top(\theta^\star - \theta^\dagger)\right]} + \eta \\
&\le \sqrt{\alpha}\sqrt{\mathbb{E}_{\substack{\tau_0\sim\mu_0 \\ \tau_1\sim\mu_1}}\left[(\theta^\star - \theta^\dagger)^\top(\phi(\tau_0) - \phi(\tau_1))(\phi(\tau_0) - \phi(\tau_1))^\top(\theta^\star - \theta^\dagger)\right]} + \eta \\
&= \sqrt{\alpha}\sqrt{\mathbb{E}_{\substack{\tau_0\sim\mu_0 \\ \tau_1\sim\mu_1}}\left[\left|(\theta^\star - \theta^\dagger)^\top(\phi(\tau_0) - \phi(\tau_1))\right|^2\right]} + \eta \\
&= \sqrt{\alpha}\sqrt{\mathbb{E}_{\substack{\tau_0\sim\mu_0 \\ \tau_1\sim\mu_1}}\left[\left|\Phi^{-1}\left(P_{\theta^\star}(o=1\mid\tau_1,\tau_0)\right) - \Phi^{-1}\left(P_{\theta^\dagger}(o=1\mid\tau_1,\tau_0)\right)\right|^2\right]} + \eta \\
&\le \sqrt{\alpha}\kappa\sqrt{\mathbb{E}_{\substack{\tau_0\sim\mu_0 \\ \tau_1\sim\mu_1}}\left[\left|P_{\theta^\star}(o=1\mid\tau_1,\tau_0) - P_{\theta^\dagger}(o=1\mid\tau_1,\tau_0)\right|^2\right]} + \eta \\
&= \frac{\sqrt{\alpha}\kappa}{\sqrt{2}}\sqrt{\mathbb{E}_{\substack{\tau_0\sim\mu_0 \\ \tau_1\sim\mu_1}}\left[\left\|P_{\theta^\star}(\cdot\mid\tau_1,\tau_0) - P_{\theta^\dagger}(\cdot\mid\tau_1,\tau_0)\right\|^2\right]} + \eta
\end{aligned}
$$

The first inequality follows from the following observation – $V^{\widetilde{\pi}}(\theta^\star) - \mathcal{R}(\theta^\star) \ge \min_{\theta\in\Theta(\widehat{\mathcal{D}}_1)} V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta) \ge \left(V^{\pi^\dagger}(\theta^\dagger) - \mathcal{R}(\theta^\dagger)\right) - \eta$. The second inequality uses Jensen's inequality. The third inequality uses the assumption of finite relative condition number (5). Now we can proceed similar to the proof of proposition 14 in Liu et al. (2022) to establish the following bound (with probability at least $1-\delta$).

$$
\mathbb{E}_{\substack{\tau_0\sim\mu_0 \\ \tau_1\sim\mu_1}}\left[\left\|P_{\theta^\star}(\cdot\mid\tau_0,\tau_1) - P_{\theta^\dagger}(\cdot\mid\tau_0,\tau_1)\right\|_1^2\right] \le \frac{c}{N}\left(\sum_{n=1}^{N}\log\left(\frac{P_{\theta^\dagger}(o^n\mid\widetilde{\tau}^{0,n},\widetilde{\tau}^{1,n})}{P_{\theta^\star}(o^n\mid\widetilde{\tau}^{0,n},\widetilde{\tau}^{1,n})}\right) + \log\left(\frac{\mathcal{N}(\Theta,1/N)}{\delta}\right)\right)
$$

Here $\mathcal{N}(\Theta, 1/N)$ is the number of elements in an ε-net of the set Θ for $\varepsilon = 1/N$. Since $\|\theta\|_2 \le H\sqrt{d}$ for each $\theta \in \Theta$ we are guaranteed that $|\mathcal{N}(\Theta, 1/N)| \le (2H\sqrt{d}N)^d$. Additionally, observe that we are using the clean data $\{\widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n}\}_{n=1}^N$ in the bound on the ratio of the log-likelihood. Now, let S be the set of clean trajectories that have been corrupted by the adversary. Then we can bound the difference in log-likelihood as follows.

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^N \log\left(\frac{P_{\theta^\dagger}(o^n \mid \widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n})}{P_{\theta^\star}(o^n \mid \widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n})}\right) &= \frac{1}{N}\sum_{n \notin S} \log\left(\frac{P_{\theta^\dagger}(o^n \mid \widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n})}{P_{\theta^\star}(o^n \mid \widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n})}\right) + \frac{1}{N}\sum_{n \in S} \log\left(\frac{P_{\theta^\dagger}(o^n \mid \widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n})}{P_{\theta^\star}(o^n \mid \widetilde{\tau}^{0,n}, \widetilde{\tau}^{1,n})}\right) \\
&\le \frac{1}{N}\sum_{n=1}^N \log\left(\frac{P_{\theta^\dagger}(o^n \mid \tau^{0,n}, \tau^{1,n})}{P_{\theta^\star}(o^n \mid \tau^{0,n}, \tau^{1,n})}\right) + \varepsilon \cdot \log\left(\frac{1 + e^{Hd}}{1 + e^{-Hd}}\right) \\
&\le \frac{1}{N}\sum_{n=1}^N \log\left(\frac{P_{\theta^\star_N}(o^n \mid \tau^{0,n}, \tau^{1,n})}{P_{\theta^\star}(o^n \mid \tau^{0,n}, \tau^{1,n})}\right) + 2\varepsilon H\sqrt{d} \\
&\le 8\varepsilon H\sqrt{d} + c \cdot \frac{d}{N}\log\left(\frac{Hd}{\delta}\right)
\end{aligned}$$

The first inequality uses $|S| \le \varepsilon N$ and $\left|\phi(\tau)^\top \theta\right| \le H\sqrt{d}$. The second inequality uses the fact that $\theta^\star_N$ maximizes the log-likelihood over the corrupted dataset, and the final inequality uses lemma 2. □

Lemma 6. *For linear MDP, the optimal value function i.e. $V^\star(\theta) = \max_\pi V^\pi(\theta)$ is $\sqrt{H}d$-Lipschitz in the reward parameter θ.*

Proof. We use the occupancy measure characterization of Markov decision process. Given a probability transition function P let $\mathcal{C}$ be the set of all feasible occupancy measures with respect to P. Then $V^\star(\theta) = \sup_{d \in \mathcal{C}} \sum_{h=1}^H d_h^\top \Phi \theta_h$.

$$\begin{aligned}
V^\star(\theta) - V^\star(\theta') &= \sup_{d \in \mathcal{C}} \sum_h d_h^\top \Phi \theta_h - \sup_{d \in \mathcal{C}} \sum_h d_h^\top \Phi \theta'_h \\
&\le \sup_{d \in \mathcal{C}} \left| \sum_{h=1}^H d_h^\top \Phi \theta_h - \sum_{h=1}^H d_h^\top \Phi \theta'_h \right| \\
&\le \sup_{d \in \mathcal{C}} \left| \sum_{h=1}^H \sum_{s,a} d_h(s,a) \phi(s,a)^\top (\theta_h - \theta'_h) \right| \\
&\le \sup_{d \in \mathcal{C}} \sum_{h=1}^H \sum_{s,a} d_h(s,a) \|\phi(s,a)\|_2 \|\theta_h - \theta'_h\|_2 \\
&\le \sqrt{d} \sum_{h=1}^H \|\theta - \theta'\|_2
\end{aligned}$$

The last inequality uses $\|\phi(s,a)\|_2 \le \sqrt{d}$ and $\sum_{s,a} d_h(s,a) = 1$ for any h. Now the claim follows from the following observation $\sum_{h=1}^H \|\theta - \theta'\|_2 \le \sqrt{H}\sqrt{\sum_{h=1}^H \|\theta - \theta'\|_2^2} = \sqrt{H}\|\theta - \theta'\|_2$. □

Lemma 7. *Suppose $X_1, \ldots, X_n$ are drawn i.i.d. from a d-dimensional distribution with covariance Σ and sub-Gaussian norm at most K. Then with probability at least $1 - \delta$ we have,*

$$\left\| \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \Sigma \right\| \le c_1 K^2 \|\Sigma\| \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n} \right).$$

Proof. See Vershynin (2018) for a proof. □

Lemma 8. *For linear models, $\min_\theta \max_\pi V^\pi(\theta) - \mathbb{E}_{\tau \sim \mu_{ref}}\left[\phi(\tau)^\top \theta\right] = \max_\pi \min_\theta V^\pi(\theta) - \mathbb{E}_{\tau \sim \mu_{ref}}\left[\phi(\tau)^\top \theta\right]$.*

Proof. We will write $\mathcal{R}(\theta) = \mathbb{E}_{\tau\sim\mu_{\text{ref}}}\left[\phi(\tau)^\top\theta\right]$. There are two cases to consider.

Case 1: First, we consider the linear MDP setting. Given a policy π let d be the corresponding occupancy measure i.e. $d_h(s,a) = \mathbb{P}_\pi(s_h = s, a_h = a)$. Then we can write the value function as $V^\pi(\theta) = \sum_{h,s,a} = d_h(s,a)\phi(s,a)^\top\theta = d^\top\Phi\theta$. This observation implies the following inequality.

$$\min_\theta \max_\pi V^\pi(\theta) - \mathcal{R}(\theta) \le \min_\theta \max_d d^\top\Phi\theta - \mathcal{R}(\theta) \tag{19}$$

On the other hand, given an occupancy measure d one can consider the following policy.

$$\pi_h^d(s,a) = \begin{cases} \frac{d_h(s,a)}{\sum_b d_h(s,b)} & \text{if } \sum_b d_h(s,b) > 0 \\ \frac{1}{A} & \text{o.w.} \end{cases}$$

Moreover, it is known that occupancy measure induced by $\pi^d = (\pi_1^d, \ldots, \pi_H^d)$ is d. This implies the following inequality.

$$\min_\theta \max_\pi V^\pi(\theta) - \mathcal{R}(\theta) \ge \min_\theta \max_d d^\top\Phi\theta - \mathcal{R}(\theta) \tag{20}$$

Therefore, from equations (20) and (19) we conclude that

$$\min_\theta \max_\pi V^\pi(\theta) - \mathcal{R}(\theta) = \min_\theta \max_d d^\top\Phi\theta - \mathcal{R}(\theta)$$

Now observe that the objective $d^\top\Phi\theta - \mathcal{R}(\theta)$ is linear in both d and θ. Therefore, strong duality holds and we can exchange the order of min and max.

$$\min_\theta \max_\pi V^\pi(\theta) - \mathcal{R}(\theta) = \min_\theta \max_d d^\top\Phi\theta - \mathcal{R}(\theta) = \max_d \min_\theta d^\top\Phi\theta - \mathcal{R}(\theta)$$

Finally, by an argument very similar to the first part of the proof (correspondence between policy and occupancy measure) we can prove the following identity.

$$\max_d \min_\theta d^\top\Phi\theta - \mathcal{R}(\theta) = \max_\pi \min_\theta V^\pi(\theta) - \mathcal{R}(\theta)$$

Case 2: We now consider the case of trajectory based linear MDP. Let $\mathcal{C}$ be the set of all valid probability distributions over the trajectories i.e. $\mathcal{C} = \{p : \sum_\tau p_\tau = 1, p_\tau \ge 0\ \forall \tau\}$. Given any policy π, one can consider the probability distribution $p^\pi \in \mathcal{C}$ induced by π so that $V^\pi(\theta) = \sum_\tau p_\tau^\pi \phi(\tau)^\top\theta = {p^\pi}^\top\Phi\theta$. This gives us the following inequality.

$$\min_\theta \max_\pi V^\pi(\theta) - \mathcal{R}(\theta) \le \min_\theta \max_{p\in\mathcal{C}} p^\top\Phi\theta - \mathcal{R}(\theta) \tag{21}$$

On the other hand, given any probability distribution $p \in \mathcal{C}$, one can consider the following non-Markovian policy.

$$\pi_h^p(a \mid h) = \begin{cases} \frac{\sum_\tau p_{h,a,\tau}}{\sum_{b,\tau} p_{h,b,\tau}} & \text{if } \sum_{b,\tau} p_{h,b,\tau} > 0 \\ \frac{1}{A} & \text{o.w.} \end{cases}$$

We will also write $P_M(\tau')$ to denote the marginal probability of a sub-trajectory τ' which is defined as $P_M(\tau') = \sum_{\tau''} p_{\tau',\tau''}$. Now given any trajectory $\tau = (s_0, a_0, s_1, a_1, s_2, \ldots, s_{H-1}, a_{H-1}, s_H)$ the probability that the τ is generated under π^p is given as,

$$\begin{aligned}
\mathbb{P}(\tau) &= \mu(s_0)\pi_0^p(a_0 \mid s_0)\mathbb{P}(s_1 \mid s_0, a_0)\mu_1^p(a_1 \mid s_0, a_0, s_1) \\
&\quad \ldots \mu_{H-1}^p(a_{H-1} \mid s_0, \ldots, s_{H-1})\mathbb{P}(s_H \mid s_{H-1}, a_{H-1})\mu_H^p(a_H \mid s_0, \ldots, s_H) \\
&= \mu(s_0)\frac{P_M(s_0,a_0)}{P_M(s_0)}\mathbb{P}(s_1 \mid s_0, a_0)\frac{P_M(s_0,a_0,s_1,a_1)}{P_M(s_0,a_0,s_1)} \\
&\quad \ldots \frac{P_M(s_0,\ldots,s_{H-1},a_{H-1})}{P_M(s_0,\ldots,s_{H-1})}\mathbb{P}(s_H \mid s_{H-1}, a_{H-1})\frac{P_M(s_0,\ldots,s_H,a_H)}{P_M(s_0,\ldots,s_H)} \\
&= \mu(s_0)\frac{P_M(s_0,a_0,s_1)}{P_M(s_0)}\frac{P_M(s_0,a_0,s_1,a_1,s_2)}{P_M(s_0,a_0,s_1)}\ldots\frac{P_M(s_0,\ldots,s_{H-1},a_{H-1},s_H)}{P_M(s_0,\ldots,s_{H-1})}\frac{P_M(s_0,\ldots,s_H,a_H)}{P_M(s_0,\ldots,s_H)} \\
&= P_M(\tau)
\end{aligned}$$

Therefore, policy π^p induces the same probability distribution over the trajectories as $p \in \mathcal{C}$. This implies the following inequality.

$$\min_{\theta} \max_{\pi} V^{\pi}(\theta) - \mathcal{R}(\theta) \geq \min_{\theta} \max_{p \in \mathcal{C}} p^{\top} \Phi \theta - \mathcal{R}(\theta) \tag{22}$$

Inequalities (21) and (21) imply the following identity.

$$\min_{\theta} \max_{\pi} V^{\pi}(\theta) - \mathcal{R}(\theta) = \min_{\theta} \max_{p \in \mathcal{C}} p^{\top} \Phi \theta - \mathcal{R}(\theta)$$

The rest of the proof is very similar to case 1 as we can again use strong duality to exchange the order of min and max. □

C MISSING PROOFS FROM SECTION 5

C.1 Proof of Theorem 5.1

Proof. As shown in the proof of Theorem 4.1, $V^\star(\theta) = \max_\pi V^\pi(\theta)$ is a convex function in θ. Let $\mathcal{R}(\theta) = \mathbb{E}_{\tau \sim \mu_{\text{ref}}}\left[\phi(\tau)^\top \theta\right]$. Then $V^\star(\theta) - \mathcal{R}(\theta)$ is convex in θ.

Now observe that, algorithm (4) performs a projected sub-gradient descent of the function $V^\star(\cdot) - \mathcal{R}(\cdot)$ with first order oracle calls. Since, `RobRL` returns a $f(\varepsilon)$ approximate subgradient of the optimal value function $V^\star(\cdot)$, $g_t + \mathbb{E}_{\tau \sim \mu_{\text{ref}}}[\phi(\tau)]$ is also an $f(\varepsilon)$ approximate sub-gradient of $V^\star(\theta_t) - \mathcal{R}(\theta_t)$. Moreover, $\|g_t + \mathbb{E}_{\tau\sim\mu_{\text{ref}}}[\phi(\tau)]\|_2 \leq \|g_t\|_2 + \|\mathbb{E}_{\tau\sim\mu_{\text{ref}}}[\phi(\tau)]\|_2 \leq G + \sqrt{H}$, and for any $\theta = (\theta_1, \ldots, \theta_H)$ we have $\|\theta\|_2 \leq \sqrt{Hd}$. Therefore, we can apply theorem C.1 to obtain the following bound.

$$V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - \min_{\theta}\left(V^\star(\theta) - \mathcal{R}(\theta)\right) \leq \frac{\sqrt{Hd(G+\sqrt{H})}}{\sqrt{T}} + f(\varepsilon)$$

If $T \geq \frac{Hd(G+\sqrt{H})}{f(\varepsilon)^2}$, we have

$$V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - \min_{\theta}\left(V^\star(\theta) - \mathcal{R}(\theta)\right) \leq 2 \cdot f(\varepsilon).$$

Since $\widetilde{\pi}$ is approximately optimal with respect to the reward parameter $\overline{\theta}$ we are guaranteed that,

$$V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) - f(\varepsilon) \leq V^{\widetilde{\pi}}(\overline{\theta}) - \mathcal{R}(\overline{\theta}) \leq V^\star(\overline{\theta}) - \mathcal{R}(\overline{\theta}) + 2f(\varepsilon)$$

We can now proceed similar to the proof of Theorem 4.1, and establish that $\widetilde{\pi}$ approximately optimizes the objective $\max_\pi \min_\theta V^\pi(\theta) - \mathcal{R}(\theta)$ i.e.

$$\min_{\theta}\left(V^{\widetilde{\pi}}(\theta) - \mathcal{R}(\theta)\right) \geq \max_{\pi}\min_{\theta}\left(V^{\pi}(\theta) - \mathcal{R}(\theta)\right) - 2f(\varepsilon)$$

Now we can apply Lemma 5 to complete the proof. □

C.2 Subgradient Descent with Biased First-Order Oracle

Setting: Our goal is to minimize a L-Lipschitz convex function $f : S \to [-M, M]$ where S is a convex and bounded set. The function f might not be differentiable, and we have access to a (first-order) noisy oracle, that given a point $x \in S$ returns a sub-gradient vector g such that

$$f(y) \geq f(x) - \beta + \langle g, y - x\rangle \,\forall y \in E.$$

We will also write $g \in \delta_\beta f(x)$ to denote such a noisy subgradient vector. The next theorem is well-known, but we provide a short proof for completeness.

Theorem C.1. *Consider the iterates of projected subgradient descent i.e.* $\theta_{t+1} = Proj_S\left(\theta_t - \eta g_t\right)$ *for* $t = 0, 1, \ldots, T-1$. *Suppose* $g_t \in \delta_\beta f(\theta_t)$ *for all* t, $\|g_t\|_2 \leq G$ *for all* t, *and* $\sup_{\theta\in S}\|\theta\|_2 \leq D$. *Then*

$$f(\overline{\theta}) - f(\theta^\star) \leq \frac{D\sqrt{G}}{\sqrt{T}} + \beta$$

Proof.

$$\left\|\theta_{t+1}-\theta^\star\right\|_2^2 \le \left\|\theta_t-\eta g_t-\theta^\star\right\|_2^2 = \left\|\theta_t-\theta^\star\right\|_2^2+\eta^2\left\|g_t\right\|_2^2-2\eta\left\langle g_t,\theta_t-\theta^\star\right\rangle$$

After rearranging and dividing by 2η, we obtain the following inequality.

$$\left\langle g_t,\theta_t-\theta^\star\right\rangle \le \frac{1}{2\eta}\left(\left\|\theta_t-\theta^\star\right\|_2^2-\left\|\theta_{t+1}-\theta^\star\right\|_2^2\right)+\frac{\eta}{2}\left\|g_t\right\|_2^2$$

Since $g_t \in \delta_\beta f(\theta_t)$ is a noisy subgradient, using convexity we obtain,

$$f(\theta_t)-f(\theta^\star)\le\left\langle\theta_t-\theta^\star,g_t\right\rangle+\beta.$$

Now using $\overline{\theta}=\frac{1}{T}\sum_{t=1}^T\theta_t$ and convexity of the function $f(\cdot)$ we obtain the following upper bound.

$$\begin{aligned}f(\overline{\theta})-f(\theta^\star)&\le\frac{1}{T}\sum_{t=1}^T f(\theta_t)-f(\theta^\star)\le\frac{1}{T}\sum_{t=1}^T\left\langle\theta_t-\theta^\star,g_t\right\rangle+\beta\\&\le\frac{1}{T}\sum_{t=1}^T\frac{1}{2\eta}\left(\left\|\theta_t-\theta^\star\right\|_2^2-\left\|\theta_{t+1}-\theta^\star\right\|_2^2\right)+\frac{\eta}{2T}\sum_{t=1}^T\left\|g_t\right\|_2^2+\beta\\&\le\frac{D^2}{2\eta T}+\frac{\eta G}{2}+\beta\end{aligned}$$

Now choosing $\eta=\frac{D}{\sqrt{GT}}$ we obtain the desired bound. □

C.3 Proof of Proposition 3

Proof. For linear MDP, the parameter $\theta=[\theta_1;\theta_2;\ldots;\theta_H]$ and the feature of a trajectory τ is constructed by concatenating the features of H state, action pairs. Therefore, $\left\|\theta\right\|_2\le\sqrt{Hd}$ and $\left\|\phi(\tau)\right\|_2\le\sqrt{H}$ for any trajectory τ.

We will use robust offline RL oracle provided by theorem 5.2. Note that if $N\ge\widetilde{\Omega}(H^2d^4\nu^4/\varepsilon^2)$ we have $f(\varepsilon)\le O\left(\nu\sqrt{\varepsilon}H^2d^{3/2}\right)$. Now using the upper bound provided in theorem (5.1) we obtain the following bound.

$$\begin{aligned}&V^\star(\theta^\star)-V^{\widetilde{\pi}}(\theta^\star)\le O\left(\nu\sqrt{\varepsilon}H^2d^{3/2}\right)\\&\le O\left(\kappa\sqrt{\alpha}\left(\sqrt{\varepsilon H}d^{1/4}+\sqrt{\frac{d}{N}\log\left(\frac{HdN}{\delta}\right)}\right)\right)\end{aligned}$$

Now observe that if $N\ge\Omega(H\cdot\mathrm{poly}(d)/\varepsilon)$ the term $\widetilde{O}(\sqrt{d/N})$ can be bounded by $O(\sqrt{\varepsilon})$. Finally, we need a lower bound of $N\ge\Omega\left(\frac{H^{3/2}dG}{f(\varepsilon)^2}\right)=\Omega(d/\varepsilon^2)$ in order to apply theorem (5.1). □

D PROJECTED SUBGRADIENT DESCENT WITH BIASED ZERO-ORDER ORACLE

Setting: Our goal is to minimize a L-Lipschitz convex function $f:S\to[-M,M]$ where S is a convex and bounded set. The function f might not be differentiable, and we only have access to a noisy oracle $\widetilde{f}$ that guarantees $\left|\widetilde{f}(x)-f(x)\right|\le\varepsilon$ for any $x\in S$. We consider a projected subgradient descent based algorithm where algorithm 6 is used to construct a biased subgradient.

Theorem D.1. *Suppose algorithm* (5) *is run for* $T\ge\frac{4DM}{\varepsilon}$ *iterations, and we set* $K\ge\frac{256CM^2d^3}{\varepsilon^2}\ln\left(\frac{16DM}{\varepsilon\delta}\right)$ *and* $\mu=\frac{\sqrt{\varepsilon}}{\sqrt{8d}}$. *Then the output* $\overline{\theta}$ *of algorithm* (5) *satisfies*

$$f(\overline{\theta})-f(\theta^\star)\le5\sqrt{\varepsilon L}$$

ALGORITHM 5: Biased Subgradient Descent

```
Input: Stepsize $\eta$, $\theta_0 \in \mathbb{R}^d$, number of iterations $T$.
for $t = 0, 1, \ldots, T-1$ do
    Construct subgradient $g_t = \widetilde{\nabla} f_\mu(\theta_t)$ using algorithm (6).
    $\theta_{t+1} = \mathrm{Proj}_S\,(\theta_t - \eta g_t)$.
$\overline{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$.
```

Proof.

$$\|\theta_{t+1} - \theta^\star\|_2^2 \le \|\theta_t - \eta g_t - \theta^\star\|_2^2 = \|\theta_t - \theta^\star\|_2^2 + \eta^2 \|g_t\|_2^2 - 2\eta \langle g_t, \theta_t - \theta^\star\rangle$$

After rearranging and dividing by 2η, we obtain the following inequality.

$$\langle g_t, \theta_t - \theta^\star\rangle \le \frac{1}{2\eta}\left(\|\theta_t - \theta^\star\|_2^2 - \|\theta_{t+1} - \theta^\star\|_2^2\right) + \frac{\eta}{2}\|g_t\|_2^2$$

Since $g_t = \widetilde{\nabla} f_\mu(\theta_t)$ is a noisy subgradient constructed by algorithm (6), using lemma (9) we get,

$$f(\theta_t) - f(\theta^\star) \le \langle \theta_t - \theta^\star, g_t\rangle + b_t$$

where

$$b_t = \sqrt{\frac{C}{K}}\frac{4M}{\mu}\sqrt{2d\ln(2/\delta)} + \frac{2\varepsilon}{\mu}\mathrm{diam}(E) + \mu L\sqrt{d}$$

Summing over $t = 0, 1, \ldots, T-1$ we obtain the following upper bound.

$$\begin{aligned}
\sum_{t=0}^{T-1} f(\theta_t) - f(\theta^\star) &\le \sum_{t=0}^{T-1} \langle \theta_t - \theta^\star, g_t\rangle + \sum_{t=1}^{T} b_t \\
&\le \frac{1}{2\eta}\sum_{t=0}^{T-1}\left(\|\theta_t - \theta^\star\|_2^2 - \|\theta_{t+1} - \theta^\star\|_2^2\right) + \frac{\eta}{2}\sum_{t=0}^{T-1}\|g_t\|_2^2 + \sum_{t=0}^{T-1} b_t \\
&\le \frac{1}{2\eta}\left(\|\theta_0 - \theta^\star\|_2^2 - \|\theta_T - \theta^\star\|_2^2\right) + \frac{\eta}{2}\sum_{t=0}^{T-1}\|g_t\|_2^2 + \sum_{t=0}^{T-1} b_t
\end{aligned}$$

From the construction of subgradient in algorithm (6) it is clear that $\|g_t\|_2 \le \frac{2M}{\mu}\mathrm{diam}(E)$. Moreover, diameter of S is at most D. This gives us the following result.

$$f(\overline{\theta}) - f(\theta^\star) \le \frac{1}{T}\sum_{t=0}^{T-1} f(\theta_t) - f(\theta^\star) \le \frac{2D^2}{\eta T} + \eta\frac{2M^2}{\mu^2 T}\mathrm{diam}^2(E) + \sqrt{\frac{C}{K}}\frac{4M}{\mu}\sqrt{2d\ln(2/\delta)} + \frac{2\varepsilon}{\mu}\mathrm{diam}(E) + \mu L\sqrt{d}$$

We now substitute $\eta = \frac{D\mu}{M\mathrm{diam}(E)}$.

$$f(\overline{\theta}) - f(\theta^\star) \le \frac{4DM}{\mu T}\mathrm{diam}(E) + \sqrt{\frac{C}{K}}\frac{4M}{\mu}\sqrt{2d\ln(2/\delta)} + \frac{2\varepsilon}{\mu}\mathrm{diam}(E) + \mu L\sqrt{d}$$

We further substitute $\mu = \frac{\sqrt{\varepsilon}}{\mathrm{diam}(E)\sqrt{L}}$ and choose $T \ge \frac{4DM}{\varepsilon}$ and $K \ge \frac{32CM^2\mathrm{diam}^2(E)}{\varepsilon^2} d\ln(2T/\delta)$.

$$f(\overline{\theta}) - f(\theta^\star) \le 4\sqrt{\varepsilon L} + \frac{\sqrt{\varepsilon L d}}{\mathrm{diam}(E)} \tag{23}$$

Now recall that lemma (9) requires that the set E be such that $\int_E \exp\left(-\frac{1}{4}\|u\|_2^2\right) du \ge \frac{1}{2}$. We choose a simple set $E = [-\ell, \ell]^d$ and show that one can pick $\ell = O(1)$. Then we have,

$$\int_E \exp\left(-\frac{1}{4}\|u\|_2^2\right) du = \left\{\int_{\ell}^{-\ell} \exp(-1/4v^2)dv\right\}^d = \left\{2\int_{\ell/2}^{-\ell/2} \exp(-1/2t^2)dt\right\}^d = \{2(2\Phi(\ell/2) - 1)\}^d.$$

Here $\Phi(t) = \mathbb{P}(X \le t)$ with X being a standard Gaussian random variable. Substituting $\Phi(t) \ge 1 - e^{-t^2/2}$ we get the following lower bound.

$$\int_E \exp\left(-\frac{1}{4}\left\|u\right\|_2^2\right) du \ge \left\{2(1 - 2e^{-\ell^2/8})\right\}^d$$

It can be checked that picking $\ell > \sqrt{8 \ln 4}$ satisfies $\int_E \exp\left(-\frac{1}{4}\left\|u\right\|_2^2\right) du \ge 1/2$. Therefore, we choose $E = [-4, 4]^d$. This also implies that $\mathrm{diam}(E) = \sqrt{8d}$ and substituting this bound in eq. (23) we obtain the following upper bound.

$$f(\overline{\theta}) - f(\theta^\star) \le 4\sqrt{\varepsilon L} + \frac{\sqrt{\varepsilon L}}{\sqrt{8}}$$

□

D.1 Gradient Construction

Given a convex function $f : E \to \mathbb{R}^d$, let f_μ be defined as its Gaussian approximation.

$$f_\mu(x) = \frac{1}{\kappa}\int_E f(x + \mu u) e^{-\frac{1}{2}\|u\|_2^2} du$$

where $\kappa = \int_E e^{-\frac{1}{2}\|u\|_2^2} du$. Suppose f is L-Lipschitz then the following results are well known Nesterov and Spokoiny (2017).

1. For any $x \in E$, $|f_\mu(x) - f(x)| \le \mu L \sqrt{d}$.
2. $\nabla f_\mu(x) = \frac{1}{\kappa}\int_E \frac{f(x+\mu u) - f(x)}{\mu} e^{-\frac{1}{2}\|u\|_2^2} u du$.
3. $\nabla f_\mu(x) \in \delta_\alpha f(x)$ for $\alpha = \mu L \sqrt{d}$ i.e. $f(y) \ge f(x) - \mu L\sqrt{d} + \langle \nabla f_\mu(x), y - x\rangle$ for all $y \in E$.

ALGORITHM 6: Gradient Construction

```
Input: Noisy oracle $\widetilde{f}$, number of iterations $K$, input $x$.
Generate $u_1, \dots, u_K$ uniformly at random from the standard normal distribution (restricted to the set $E$).
Let $\widetilde{\nabla} f_\mu(x) = \frac{1}{K}\sum_{k=1}^K \frac{\widetilde{f}(x+\mu u_k) - \widetilde{f}(x)}{\mu} u_k$.
$\widetilde{\nabla} f_\mu(x)$.
```

Lemma 9. *Suppose the set E is chosen so that $\int_E e^{-\frac{1}{4}\|u_k\|_2^2} \ge \frac{1}{2}$, and $\left|\widetilde{f}(x) - f(x)\right| \le \varepsilon$ for any x. Then the gradient estimate returned by algorithm (6) satisfies*

$$\widetilde{\nabla} f_\mu(x) \in \delta_\alpha f(x) \quad \textit{for} \quad \alpha = \sqrt{\frac{C}{K}}\frac{4M}{\mu}\sqrt{2d\ln(2/\delta)} + \frac{2\varepsilon}{\mu} \textit{diam}(E) + \mu L \sqrt{d}$$

with probability at least $1 - \delta$.

Proof. Let $\widehat{\nabla} f_\mu(x) = \frac{1}{K}\sum_{k=1}^K \frac{f(x+\mu u_k) - f(x)}{\mu} u_k$. Then we have,

$$\begin{aligned}\left\|\widehat{\nabla} f_\mu(x) - \widetilde{\nabla} f_\mu(x)\right\|_2 &= \frac{1}{K}\left\|\sum_{k=1}^K \frac{\left(\widetilde{f}(x+\mu u_k) - f(x+\mu u_k)\right) - \left(\widetilde{f}(x) - f(x)\right)}{\mu} u_k\right\|_2 \\ &\le \frac{2\varepsilon}{\mu K}\sum_{k=1}^K \|u_k\|_2 \\ &\le \frac{2\varepsilon}{\mu}\mathrm{diam}(E) \qquad (24)\end{aligned}$$

We now show that $\widehat{\nabla} f_\mu(x)$ concentrates around $\nabla f_\mu(x)$. Let $V_k = \frac{f(x+\mu u_k)-f(x)}{\mu} u_k$. We claim that the sub-Gaussian norm of V_k is at most $\frac{4M}{\mu}$. This follows from two observations. First, $\left|\frac{f(x+\mu u_k)-f(x)}{\mu}\right| \le \frac{2M}{\mu}$. Second, we show that the sub-Gaussian norm of the random vector u_k is at most 2. Since $\left\|u_k\right\|_{\psi_2} = \sup_{v \in S_{d-1}} \left\|u_k^\top v\right\|_{\psi_2}$, consider any $v \in \mathbb{R}^d$ with $\left\|v\right\|_2 = 1$.

$$\mathbb{E}\left[e^{\frac{(u_k^\top v)^2}{4}}\right] = \frac{1}{\kappa}\int_E e^{\frac{(u_k^\top v)^2}{4}} e^{-\frac{1}{2}\left\|u_k\right\|_2^2} du_k \le \frac{1}{\kappa}\int_E e^{-\frac{1}{4}\left\|u_k\right\|_2^2} du_k = \frac{\int_E e^{-\frac{1}{4}\left\|u_k\right\|_2^2}}{\int_E e^{-\frac{1}{2}\left\|u_k\right\|_2^2}} \le \frac{1}{\int_E e^{-\frac{1}{4}\left\|u_k\right\|_2^2}} \le 2$$

The first inequality uses $\left\|v\right\|_2 = 1$, and the second inequality uses Jensen's inequality. We can now use proposition 2.6.1 from Vershynin (2018) to bound the sub-Gaussian norm of the average vector.

$$\left\|\sum_{k=1}^K \frac{f(x+\mu u_k)-f(x)}{\mu} u_k\right\|_{\psi_2} \le \sqrt{C \sum_{k=1}^K \left\|\frac{f(x+\mu u_k)-f(x)}{\mu} u_k\right\|_{\psi_2}^2} \le \sqrt{CK}\frac{4M}{\mu}$$

for some universal constant $C > 0$. Therefore, $\left\|\widehat{\nabla} f_\mu(x)\right\|_{\psi_2} \le \sqrt{\frac{C}{K}}\frac{4M}{\mu}$. This also means that $\left\|\widehat{\nabla} f_\mu(x)\right\|_{\psi_2}$ is $\sqrt{\frac{C}{K}}\frac{4M}{\mu}\sqrt{d}$ norm sub-Gaussian Jin et al. (2019) and from the definition of norm sub-Gaussian random vectors (definition 3 from Jin et al. (2019)) we have the following bound.

$$\Pr\left(\left\|\widehat{\nabla} f_\mu(x) - \nabla f_\mu(x)\right\|_2 \ge \sqrt{\frac{C}{K}}\frac{4M}{\mu}\sqrt{2d\ln(2/\delta)}\right) \le \delta \tag{25}$$

Finally, we can combine eq. (24), and eq. (25) and use item 3 to obtain the desired bound. □

E A NEW CORRUPTION ROBUST OFFLINE RL METHOD

We adopt the linear programming based formulation of reinforcement learning Manne (1960). We will write $\Phi \in \mathbb{R}^{SA \times d}$ to write the feature matrix, and $P_h \in \mathbb{R}^{S \times SA}$ to be the transition probability matrix at time-step h, which is defined as $P_h(s,(s',b')) = P_h(s \mid s', b')$. Note that we can write $P_h = \Psi_h \Phi^\top$ where $\boldsymbol{\mu}_h \in \mathbb{R}^{S \times d}$ is the $\boldsymbol{\mu}_h$ is the d-dimensional measure matrix.

$$\begin{aligned}
\max_q \ & \sum_{h=1}^H q_h^\top \Phi \theta_h \\
\text{s.t. } & \sum_a q_1(s,a) = \rho(s) \ \forall s \\
& E q_{h+1} = \boldsymbol{\mu}_h \Phi^\top q_h \ \forall h \in \{1,2,\ldots,H-1\} \\
& q_h \ge 0 \ \forall h \in [H]
\end{aligned}$$

The matrix $E \in \mathbb{R}^{S \times SA}$ is defined as $E(s,(s',a')) = \mathbb{1}\{s = s'\}$. We make the following substitution $\lambda_h = \Phi^\top q_h$ to obtain the following equivalent LP.

$$\begin{aligned}
\max_{\{q_h\}_{h=1}^H, \{\lambda_h\}_{h=1}^H} \ & \sum_{h=1}^H \lambda_h^\top \theta_h \\
\text{s.t. } & E q_1 = \rho \\
& E q_{h+1} = \boldsymbol{\mu}_h \lambda_h \ \forall h \in \{1,2,\ldots,H-1\} \\
& q_h \ge 0 \ \forall h \in [H] \\
& \lambda_h = \Phi^\top q_h \ \forall h \in [H]
\end{aligned} \tag{26}$$

The dual problem of the above optimization problem is the following optimization problem.

$$\begin{aligned}
\min_{\{v_h\}_{h=1}^H, \{w_h\}_{h=1}^H} \ & \rho^\top v_1 \\
\text{s.t. } & E^\top v_h \ge \Phi w_h \ \forall h \in [H] \\
& w_h \ge \theta_h + \boldsymbol{\mu}_h^\top v_{h+1} \ \forall h \in [H-1] \\
& w_H \ge \theta_H
\end{aligned} \tag{27}$$

The corresponding Lagrangian is given as $\mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w})$ where

$$\begin{aligned}\mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w}) &= \rho^\top v_1 + \sum_{h=1}^{H} \left\langle q_h, -E^\top v_h + \Phi w_h \right\rangle + \sum_{h=1}^{H-1} \left\langle \theta_h + \boldsymbol{\mu}_h^\top v_{h+1} - w_h, \lambda_h \right\rangle + \langle \theta_H - w_H, \lambda_H \rangle \\ &= \sum_{h=1}^{H} \lambda_h^\top \theta_h + \langle v_1, -E q_1 + \rho \rangle + \sum_{h=2}^{H} \langle v_h, -E q_{h+1} + \boldsymbol{\mu}_h \lambda_h \rangle + \sum_{h=1}^{H} \left\langle w_h, \Phi^\top q_h - \lambda_h \right\rangle\end{aligned}$$

We aim to solve a saddle point of the Lagrangian through gradient descent-ascent method. Note that each of λ_h and w_h is d-dimensional. So we will only perform gradient steps over these variables, whereas we will represent high-dimensional (possible infinite) v_h and q_h implicitly. The gradient with respect to λ_h is given through the following expression.

$$\nabla_{\lambda_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w}) = \begin{cases} \theta_h + \boldsymbol{\mu}_h^\top v_{h+1} - w_h & \text{if } h \in [H-1] \\ \theta_h - w_h & \text{if } h = H \end{cases}$$

Now we introduce a transformation of variables suggested by Gabbianelli et al. (2024). Let $\Lambda_h = \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}^h}\left[\phi(s,a)\phi(s,a)^\top\right]$ be the covariance matrix under the reference policy μ_{ref} at time step h. Then we can rewrite the gradient as follows.

$$\begin{aligned}\nabla_{\lambda_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\lambda}; \boldsymbol{v}, \boldsymbol{w}) &= \Lambda_h^{-1} \Lambda_h \left(\theta_h + \boldsymbol{\mu}_h^\top v_{h+1} - w_h\right) = \Lambda_h^{-1} \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}^h}\left[\phi(s,a)\phi(s,a)^\top \left(\theta_h + \boldsymbol{\mu}_h^\top v_{h+1} - w_h\right)\right] \\ &= \Lambda_h^{-1} \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}^h, s'\sim P_h(\cdot|s,a)}\left[\phi(s,a)\left(r_h(s,a) + v_{h+1}(s') - w_h^\top \phi(s,a)\right)\right]\end{aligned}$$

We can build an estimator of the expectation from samples, however the covariance matrix Λ_h might be unknown. Therefore, as proposed by Gabbianelli et al. (2024), we substitute $\beta_h = \Lambda_h^{-1}\lambda_h$ for any $h \in [H]$ in the Lagrangian.

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = \rho^\top v_1 + \sum_{h=1}^{H} \left\langle q_h, -E^\top v_h + \Phi w_h \right\rangle + \sum_{h=1}^{H-1} \left\langle \Lambda_h \left(\theta_h + \boldsymbol{\mu}_h^\top v_{h+1} - w_h\right), \beta_h \right\rangle + \langle \Lambda_H \left(\theta_H - w_H\right), \beta_H \rangle \tag{28}$$

Gradient with respect to β_h is given as follows.

$$\nabla_{\beta_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = \begin{cases} \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}^h, s'\sim P_h(\cdot|s,a)}\left[\phi(s,a)\left(r_h(s,a) + v_{h+1}(s') - w_h^\top \phi(s,a)\right)\right] & \text{if } h \in [H-1] \\ \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}^h}\left[\phi(s,a)\left(r_h(s,a) - w_h^\top \phi(s,a)\right)\right] & \text{if } h = H \end{cases}$$

Therefore, given any data point (s_h, a_h, s'_h, r_h) we can define the following estimate of the gradient.

$$\widetilde{g}_{\beta_h} = \widehat{\nabla}_{\beta_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = \begin{cases} \phi(s_h, a_h)\left(r_h + v_{h+1}(s'_h) - w_h^\top \phi(s_h, a_h)\right) & \text{if } h \in [H-1] \\ \phi(s_h, a_h)\left(r_h - w_h^\top \phi(s_h, a_h)\right) & \text{if } h = H \end{cases}$$

On the other hand, the gradient with respect to w_h is the following.

$$\nabla_{w_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = \Phi^\top q_h - \Lambda_h \beta_h = \Phi^\top q_h - \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}}\left[\phi(s,a) \cdot \beta_h^\top \phi(s,a)\right]$$

This leads to the following estimate of the gradient with respect to w_h.

$$\widetilde{g}_{w_h} = \widehat{\nabla}_{w_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = \Phi^\top q_h - \phi(s_h, a_h) \cdot \beta_h^\top \phi(s_h, a_h)$$

We will also use the following symbolic representation for policy, value, and occupancy measure.

$$\pi_h(a \mid s) = \frac{\exp(\phi(s,a)^\top w_h)}{\sum_b \exp(\phi(s,b)^\top w_h)}$$

$$v_h(s) = \sum_a \pi_h(a \mid s)\phi(s,a)^\top w_h$$

ALGORITHM 7: Corruption Robust Offline Primal-Dual

```
Input: (a) Corrupted dataset $\mathcal{D}$, (b) corruption parameter $\varepsilon$, (c) Step sizes $\eta_w$, $\eta_b$, and $\alpha$, and (d) Number of iterations $T$.
Partition dataset $\mathcal{D}$ uniformly at random into two datasets $\mathcal{D}_m$ and $\mathcal{D}_c$, where $\mathcal{D}_c = \Theta(H \cdot d^2/\varepsilon^2 \log^2(d))$.
Partition dataset $\mathcal{D}_m$ uniformly at random into $2HT$ groups $\{\mathcal{D}_1^{t,h}, \mathcal{D}_2^{t,h}\}_{h\in[H],t\in[T]}$.
Initialize $w^0 = \{w_h^0\}_{h=1}^H$ and $\beta^0 = \{\beta_h^0\}_{h=1}^H$.
for $t = 0, \dots, T-1$ do
    for $h = 1, \dots, H$ do
        /* Take a gradient step for $w_h$ */
        Set $\pi_h^t(a \mid s) \propto \exp\left(\alpha\phi(s,a)^\top w_h^t\right)$.
        For each $j \in [K]$, set (symbolically)
            $$q_{h,j}^t(\widetilde{s}, \widetilde{b}) = \begin{cases} \pi_h^t(\widetilde{b} \mid \widetilde{s}) \cdot \mathbb{1}\left\{s_j^{'} = \widetilde{s}\right\} \phi(s_{h,j}, a_{h,j})^\top \beta_{h-1}^t & \text{if } h > 1 \\ \pi_h^t(\widetilde{b} \mid \widetilde{s}) \cdot \rho(\widetilde{s}) & \text{if } h = 1 \end{cases}$$
         Set $\widetilde{g}_{w_h}^t = \text{RobMean}\left(\left\{\Phi^\top q_{h,j}^t - \phi(s_{h,j}^2, a_{h,j}^2) \cdot \left\langle \beta_h^t, \phi(s_{h,j}^2, a_{h,j}^2)\right\rangle\right\}_{j=1}^K\right)$.
        $w_h^{t+1} \leftarrow \text{Proj}_{\mathcal{W}}(w_h^t - \eta_w \cdot \widetilde{g}_{w_h}^t)$
    for $h = 1, \dots, H$ do
        /* Take a gradient step for $\beta_h$ */
        Set $\pi_h^t(a \mid s) \propto \exp\left(\alpha\phi(s,a)^\top w_h^t\right)$.
        Set $v_h^t(s) = \sum_a \pi_h^t(a \mid s) \cdot \phi(s,a)^\top w_h^t$.
        Set $\widetilde{g}_{\beta_h}^t = \widehat{\nabla}_{\beta_h} \mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w})$ defined as
            $$\widetilde{g}_{\beta_h}^t = \begin{cases} \text{RobMean}\left(\left\{\phi(s_{h,j}, a_{h,j})\left(r_{h,j} + v_{h+1}(s'_{h,j}) - \left\langle w_h^t, \phi(s_h, a_h)\right\rangle\right)\right\}_{j=1}^K\right) & \text{if } h \in [H-1] \\ \text{RobMean}\left(\left\{\phi(s_{h,j}, a_{h,j})\left(r_{h,j} - \left\langle w_h^t, \phi(s_{h,j}, a_{h,j})\right\rangle\right)\right\}_{j=1}^K\right) & \text{if } h = H \end{cases}$$
         $\beta_h^{t+1} \leftarrow \text{Proj}_{\mathcal{B}}(\beta_h^t + \eta_b \cdot \widetilde{g}_{\beta_h}^t)$
Partition dataset $\mathcal{D}_c$ uniformly at random into $H$ groups $\{\mathcal{D}_c^h\}_{h\in[H]}$.
for $h = 1, \dots, H$ do
    Set $\overline{w}_h = \frac{1}{T}\sum_{t=1}^T w_h^t$ and $\overline{\beta}_h = \frac{1}{T}\sum_{t=1}^T \beta_h^t$.
    Set $\widehat{v}_h = \text{RobCovariance}(\mathcal{D}_c^h) \cdot \overline{\beta}_h$
return $\overline{\pi} = (\overline{\pi}_1, \dots, \overline{\pi}_H)$ and $\widehat{v} = (\widehat{v}_1, \dots, \widehat{v}_H)$.
```

and

$$q_1(s) = \rho(s) \quad \text{and} \quad q_{h+1}(s') = \boldsymbol{\mu}_h(s')^\top \Lambda_h \beta_h = \mathbb{E}_{(s,a)\sim\mu_{\text{ref}}^h}\left[P_h(s' \mid s,a)\phi(s,a)^\top \beta_h\right]$$

Given w_h, β_h we define policy π_h as

$$\pi_h(a \mid s) = \frac{\exp(\phi(s,a)^\top w_h)}{\sum_b \exp(\phi(s,b)^\top w_h)}.$$

We also define $q_h^{\pi,\beta}$ as

$$q_h^{\pi,\beta}(s,a) = \begin{cases} \pi_h(a \mid s) \cdot \rho(s) & \text{if } h = 1 \\ \pi_h(a \mid s) \cdot \boldsymbol{\mu}_h(s)^\top \Lambda_{h-1}\beta_{h-1} & \text{o.w.} \end{cases}$$

After substituting $q_h = q_h^{\pi,\beta}$ we obtain the following form of the Lagrangian.

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{w}) = \sum_{h=1}^H \langle \Lambda_h \theta_h, \beta_h \rangle + \sum_{h=1}^H \left\langle w_h, \Phi^\top q_h^{\pi,\beta} - \Lambda_h \beta_h \right\rangle \tag{29}$$

This also gives us the following expression for derivative with respect to w_h.

$$\nabla_{w_h} f(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{w}) = \Phi^\top q_h^{\pi,\beta} - \Lambda_h \beta_h \tag{30}$$

Additionally, if we write $v_h^{\pi,w}(s) = \sum_a \pi_h(a \mid s) \cdot w_h^\top \phi(s,a)$ and $d_h^\beta = Eq_h^{\pi,\beta}$ then we obtain the following form of the Lagrangian.

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{\beta}; \boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{w}) = \sum_{h=1}^H \langle \Lambda_h(\theta_h - w_h), \beta_h \rangle + \sum_{h=1}^H \left\langle d_h^\beta, v_h^{\pi,w} \right\rangle \tag{31}$$

And, we can write down the derivative with respect to β_h for any $h > 1$ as

$$\nabla_{\beta_h} f(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{w}) = \Lambda_h(\theta_h - w_h) + \sum_{s'} v_{h+1}^{\pi,w}(s') \nabla_{\beta_h} d_{h+1}^\beta(s') = \Lambda_h(\theta_h - w_h) + \sum_{s'} v_{h+1}^{\pi,w}(s') \Lambda_h \boldsymbol{\mu}_h(s'). \tag{32}$$

And, for $h = 1$ we have,

$$\nabla_{\beta_h} f(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{w}) = \Lambda_h(\theta_h - w_h). \tag{33}$$

Following Gabbianelli et al. (2024) we define the following notion of regret.

$$\mathcal{R}(\boldsymbol{\beta}^\star, \boldsymbol{\pi}^\star, \boldsymbol{w}_{1:T}^\star) = \frac{1}{T} \sum_{t=1}^T f(\boldsymbol{\beta}^\star, \boldsymbol{\pi}^\star, \boldsymbol{w}_t) - f(\boldsymbol{\beta}_t, \boldsymbol{\pi}_t, \boldsymbol{w}_t^\star) \tag{34}$$

Lemma 10. *Suppose $\pi^\star = (\pi_1^\star, \ldots, \pi_H^\star)$ be a policy and $q^{\pi^\star}$ be its state, action occupancy measure. If we set $\beta_h^\star = \Lambda_h^{-1} \Phi^\top q_h^\star$ for each $h = 1, \ldots, H$, and $w_{t,h}^\star = w_h^t$ for each $t \in [T]$ and $h \in [H]$, the the policy $\overline{\pi}$ output by algorithm (7) satisfies*

$$\mathbb{E}\left[(q^{\pi^\star} - q^{\overline{\pi}})^\top r \right] \leq \mathcal{R}(\boldsymbol{\beta}^\star, \boldsymbol{\pi}^\star, \boldsymbol{w}_{1:T}^\star)$$

Proof. The proof is very similar to the proof of lemma 4.1 of Gabbianelli et al. (2024). □

Lemma 11. *With the choice of the parameters as in Lemma 10, we have the following regret decomposition.*

$$\begin{aligned}\mathcal{R}(\boldsymbol{\beta}^\star, \boldsymbol{\pi}^\star, \boldsymbol{w}_{1:T}^\star) = \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \langle w_{t,h} - w_h^\star, \nabla_{w_h} f(\boldsymbol{\pi}_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t) \rangle + \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \left\langle \beta_h^\star - \beta_{t,h}^\star, \nabla_{\beta_h} f(\boldsymbol{\pi}_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t) \right\rangle \\ + \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \sum_s q_h^{\pi^\star}(s) \sum_a (\pi_h^\star(a \mid s) - \pi_{t,h}(a \mid s)) \langle w_{t,h}, \phi(s,a) \rangle\end{aligned}$$

Proof. The proof is very similar to the proof of lemma 4.2 of Gabbianelli et al. (2024). □

E.1 Formal Statement and Proof of Theorem 5.2

Theorem E.1. *Suppose assumptions (7) holds, and $N \geq \Omega\left(\frac{H^2 d^4 \nu^4}{\varepsilon^2} (\log^2 d + \log^2 A) \right)$. Then the policy $\overline{\pi}$ output by algorithm (7) is approximately optimal i.e.*

$$\max_\pi V^\pi(\theta) - \mathbb{E}\left[V^{\overline{\pi}}(\theta) \right] \leq O\left(\nu \sqrt{\varepsilon} H^2 d^{3/2} \right),$$

and the vector $\widehat{v} = (\widehat{v}_1, \ldots, \widehat{v}_H)$ is an approximate sub-gradient to $V^\star(\theta) = \max_\pi V^\pi(\theta)$ i.e.

$$V^\star(\theta') \geq V^\star(\theta) + \sum_{h=1}^H \langle \widehat{v}_h, \theta_h \rangle - O\left(\nu \sqrt{\varepsilon} H^2 d^{3/2} \right) \ \forall \theta'.$$

Proof. Let Λ_h be the feature covariance matrix under the offline policy π_{ref} at time step h. Moreover, let $d_h^\star = \mathbb{E}_{(s,a) \sim \pi^\star}[\phi(s,a)]$ and $\beta_h^\star = d_h^\star \Lambda_h^{-1}$. Then by assumption (7), $\|\beta^\star\|_2 \leq \nu$. Therefore, it is sufficient to take diameter of the set $\mathcal{B}$ to be ν. We now bound the diameter of the set $\mathcal{W}$ from the feasiblity condition in the optimization problem (27). It can be easily seen that given any optimal solution $(\{v_h\}_{h=1}^H, \{w_h\}_{h=1}^H)$, we can always choose $w_h = \theta_h + \boldsymbol{\mu}_h^\top v_{h+1}$ for any $h \in [H-1]$, and $w_H = \theta_H$. Indeed, if this condition is not satisfied, then we can define the following new set of variables.

$$\widetilde{w}_H = \theta_H \;\text{ and }\; \widetilde{w}_h = \theta_h + \boldsymbol{\mu}_h^\top \widetilde{v}_{h+1},\; \widetilde{v}_h(s) = \sum_a \phi(s,a)^\top \widetilde{w}_h \;\text{ for } h = H-1,\dots,1$$

This new set of variables is feasible to the optimization problem (27) and has objective value bounded above by $\rho^\top v_1$. For linear MDP, the reward at every step is at most $\sqrt{d}$, and hence the value function $v_h(s)$ is at most $H\sqrt{d}$. This implies that for any h, $\|w_h\|_2 \le \|\theta_h\| + \left\|\boldsymbol{\mu}_h^\top \widetilde{v}_{h+1}\right\|_2 \le \sqrt{d} + H\sqrt{d}\,\|\boldsymbol{\mu}_h\|_2 \le 2Hd$. Therefore, $\|w\|_2^2 = \sum_{h=1}^H \|w_h\|_2^2 \le 2H^2 d$, and we can take the diameter of the set $\mathcal{W}$ to be at most $2H\sqrt{d}$.

By lemma (10) and (11) we can express the suboptimality of value function as follows.

$$V^{\pi^\star}(\theta) - \mathbb{E}\left[V^{\overline{\pi}}(\theta)\right] \le \underbrace{\frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \left\langle w_{t,h} - w_h^\star, \nabla_{w_h} f(\boldsymbol{\pi}_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t)\right\rangle}_{:=\mathrm{Reg}_1} + \underbrace{\frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \left\langle \beta_h^\star - \beta_{t,h}^\star, \nabla_{\beta_h} f(\boldsymbol{\pi}_t, \boldsymbol{\beta}_t, \boldsymbol{w}_t)\right\rangle}_{:=\mathrm{Reg}_2}$$
$$+ \underbrace{\frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \sum_s q_h^{\pi^\star}(s) \sum_a \left(\pi_h^\star(a \mid s) - \pi_{t,h}(a \mid s)\right)\left\langle w_{t,h}, \phi(s,a)\right\rangle}_{:=\mathrm{Reg}_3}$$

We now apply Lemma 12 with $W = 2H\sqrt{d}$, $B = \nu$, and $\eta_w = \frac{W}{Bd}\frac{1}{\sqrt{T}} = \frac{H}{\nu\sqrt{dT}}$ to obtain the following bound on the term Reg_1.

$$\mathrm{Reg}_1 \le O\left(\nu\sqrt{\varepsilon}dH\sum_{h=1}^H \|\Lambda_h\|_2 + \frac{\nu H^2 d^{3/2}}{\sqrt{T}}\right) \tag{35}$$

We apply Lemma 13 with $W = 2H\sqrt{d}$, $B = \nu$ and $\eta_b = \sqrt{\frac{HB^2}{2T}} \cdot \frac{1}{\sqrt{(d+W^2)Hd^2}} = \frac{\nu}{d^{3/2}\sqrt{2(H^2+1)}}\frac{1}{\sqrt{T}}$ to obtain the following bound on the term Reg_2.

$$\mathrm{Reg}_2 \le O\left(\sqrt{\varepsilon}dH\sum_{h=1}^H \|\Lambda_h\|_2 + \frac{H^2\nu d^{3/2}}{\sqrt{T}}\right) \tag{36}$$

For the third term, we apply Lemma 15 separately for each $h \in [H]$. In particular, we set $q_t^h = \Phi w_{t,h}$, and $D = \left\|q_t^h\right\|_\infty \le W$.

$$\mathrm{Reg}_3 \le \frac{1}{T}\sum_{h=1}^H \frac{\mathcal{H}(\pi_h^\star \| \pi_1^h)}{\alpha} + \frac{H\alpha W^2}{2}$$

We now substitute $W = H\sqrt{d}$, $\mathcal{H}(\pi_h^\star \| \pi_1^h) \le \log A$ and $\alpha = \frac{1}{H} \cdot \sqrt{\frac{2\log A}{dT}}$ to obtain the following bound.

$$\mathrm{Reg}_3 \le O\left(H^2\sqrt{\frac{d\log A}{T}}\right) \tag{37}$$

Using the upper bounds on Reg_1, Reg_2, and Reg_3, we obtain the following upper bound on the suboptimality gap.

$$V^{\pi^\star}(\theta) - \mathbb{E}\left[V^{\overline{\pi}}(\theta)\right] \le O\left(\nu\sqrt{\varepsilon}dH\sum_{h=1}^H \|\Lambda_h\|_2 + \frac{\nu H^2 d^{3/2}}{\sqrt{T}} + H^2\sqrt{\frac{d\log A}{T}}\right)$$

Now we substitute $\|\Lambda_h\|_2 \le \mathrm{Trace}(\Lambda_h) \le d$, for any $h \in [H]$. Moreover, we must have $K \ge \Theta((d/\varepsilon)\log d)$ and $N \ge KTH$. If we use $T = \sqrt{N}$ then we need $N \ge \widetilde{O}\left(\frac{H^2 d^2}{\varepsilon^2}\right)$. This substitution gives us the following upper bound.

$$V^{\pi^\star}(\theta) - \mathbb{E}\left[V^{\overline{\pi}}(\theta)\right] \le O\left(\nu\sqrt{\varepsilon}H^2 d^{3/2} + \frac{\nu H^2 d^{3/2} + H^2\sqrt{d\log A}}{N^{1/4}}\right)$$

If $N \geq \frac{(\nu d+\sqrt{\log A})^4}{\nu^4 \varepsilon^2}$ then the second term dominates the first term and we get the following bound.

$$V^{\pi^\star}(\theta) - \mathbb{E}\left[V^{\overline{\pi}}(\theta)\right] \leq O\left(\nu\sqrt{\varepsilon} H^2 d^{3/2}\right)$$

For any $h \in [H]$, the average of the feature distribution at time-step h is $\mathbb{E}_{(s,a)\sim\overline{\pi}_h}\left[\phi(s,a)\right] = \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{(s,a)\sim\pi_h^t}\left[\phi(s,a)\right] = \frac{1}{T}\sum_{t=1}^T \Phi^\top q_h^{\pi^t} = \frac{1}{T}\sum_{t=1}^T \Lambda_h \beta_h^t = \Lambda_h \overline{\beta}_h$. Algorithm (7) performs a robust covariance estimation of Λ_h, and then multiplies this estimator to $\overline{\beta}_h$ to obtain the average feature distribution. Give any feature $\phi = \phi(s,a)$, let X be the flattened vector $\phi\phi^\top$. Then each entry of the matrix $XX^\top$ can be expressed as $\phi_i\phi_j\phi_k\phi_\ell$ where $1 \leq i,j,k,\ell \leq m$. This means that $\left\|XX^\top\right\|_F^2 = \sum_{i,j,k,\ell}\phi_i^2\phi_j^2\phi_k^2\phi_\ell^2 = \|\phi\|_2^8 \leq 1$, and $\mathrm{cov}(X) \leq 2\cdot\mathrm{Id}$. So we can apply Lemma 17 and conclude that $\left\|\hat{\Lambda}_h - \Lambda_h\right\|_2 \leq O(\sqrt{\varepsilon})$ for any $h \in [H]$. Therefore, for any $h \in [H]$, $\left\|\hat{v}_h - \Lambda_h\overline{\beta}_h\right\|_2 \leq \left\|\hat{\Lambda}_h - \Lambda_h\right\|_2 \left\|\overline{\beta}_h\right\|_2 \leq O\left(\sqrt{\varepsilon}\nu\right)$. This bound also implies that $\left\|(\widehat{v}_1,\ldots,\widehat{v}_H) - \left(\Lambda_1\overline{\beta}_1,\ldots,\Lambda_H\overline{\beta}_H\right)\right\|_2 \leq O\left(\nu\sqrt{H\varepsilon}\right)$.

Now recall that we can write $V^{\overline{\pi}}(\theta) = \sum_{h=1}^H \left\langle \Lambda_h\overline{\beta}_h, \theta_h\right\rangle \geq \sum_{h=1}^H \left\langle \widehat{v}_h, \theta_h\right\rangle - O\left(\nu\sqrt{H\varepsilon}\right)\|(\theta_1,\ldots,\theta_H)\|_2 \geq \sum_{h=1}^H \left\langle \widehat{v}_h, \theta_h\right\rangle - O\left(\nu H\sqrt{d\varepsilon}\right)$. Since $\overline{\pi}$ is an approximate $O\left(\nu\sqrt{\varepsilon}H^2 d^{3/2}\right)$ optimal policy, and $\widehat{v} = (\widehat{v}_1,\ldots,\widehat{v}_H)$ is an approximate $O\left(\nu\sqrt{\varepsilon}H d^{1/2}\right)$ subgradient of $V^{\overline{\pi}}(\theta)$, we can apply lemma (18) to conclude that $\widehat{v} = (\widehat{v}_1,\ldots,\widehat{v}_H)$ is also an approximate $O\left(\nu\sqrt{\varepsilon}H^2 d^{3/2}\right)$ of the optimal value function with respect to the reward parameter θ. □

We now bound the three terms appearing in lemma (11).

Lemma 12. *Assume diam$(\mathcal{B}) \leq B$, diam$(\mathcal{W}) \leq W$, and $K \geq \Theta\left((d/\varepsilon)\log d\right)$. Then we have,*

$$\frac{1}{T}\sum_{t=1}^T\sum_{h=1}^H \left\langle w_{t,h} - w_h^\star, \nabla_{w_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\rangle \leq O\left(\sqrt{\varepsilon}WB\sum_{h=1}^H \|\Lambda_h\|_2 + \frac{HW^2}{\eta_w T} + \eta_w B^2 \sum_{h=1}^H \|\Lambda_h\|_2^2\right)$$

with constant probability.

Proof. Let $\overline{g}_{w_h}^t = \frac{1}{K}\sum_{j=1}^K \Phi^\top q_{h,j}^t - \phi(s_{h,j}^2, a_{h,j}^2)\cdot\left\langle \beta_h^t, \phi(s_{h,j}^2, a_{h,j}^2)\right\rangle$. From the definition of q_h^t in algorithm (7), we have for any $h > 1$,

$$\begin{aligned}
\mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[q_{h,j}^t(\widetilde{s},\widetilde{b})\right] &= \mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[\pi_h^t(\widetilde{b}\mid\widetilde{s})\cdot\mathbb{1}\left\{s'_{h,j} = \widetilde{s}\right\}\phi(s_{h,j},a_{h,j})^\top\beta_{h-1}^t\right] \\
&= \pi_h^t(\widetilde{b}\mid\widetilde{s})\cdot\mathbb{E}_{(s,a)\sim\mu_{\mathrm{ref}}^h}\left[P_{h-1}(\widetilde{s}\mid s,a)\phi(s,a)^\top\beta_{h-1}^t\right] \\
&= \pi_h^t(\widetilde{b}\mid\widetilde{s})\cdot\boldsymbol{\mu}_{h-1}(\widetilde{s})^\top\mathbb{E}_{(s,a)\sim\mu_{\mathrm{ref}}^{h-1}}\left[\phi(s,a)\phi(s,a)^\top\beta_{h-1}^t\right] \\
&= \pi_h^t(\widetilde{b}\mid\widetilde{s})\cdot\boldsymbol{\mu}_{h-1}(\widetilde{s})^\top\Lambda_{h-1}\beta_{h-1}^t = q_h^{\pi^t,\beta_t}(\widetilde{s},\widetilde{b})
\end{aligned}$$

Additionally $\mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[q_{h,j}^t(\widetilde{s},\widetilde{b})\right] = \pi_h^t(\widetilde{b}\mid\widetilde{s})\cdot\rho(\widetilde{s}) = q_h^{\pi^t,\beta^t}(\widetilde{s},\widetilde{b})$. We now bound on the deviation of the estimator $\widetilde{g}_{w_h}^t$ from $\nabla_{w_h} f(\boldsymbol{\pi}_t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)$.

$$\begin{aligned}
\mathbb{E}_{\mu_{\mathrm{ref}}^h,\mathcal{D}^{t,h}}\left[\overline{g}_{w_h}^t\right] &= \frac{1}{K}\sum_{j=1}^K \cdot\mathbb{E}_{\mu_{\mathrm{ref}}^h,\mathcal{D}^{t,h}}\left[\Phi^\top q_{h,j}^t - \phi(s_{h,j},a_{h,j})\cdot\left\langle\beta_h^t,\phi(s_{h,j},a_{h,j})\right\rangle\right] \\
&= \Phi^\top q_h^{\pi^t,\beta^t} - \mathbb{E}_{(s,a)\sim\mu_{\mathrm{ref}}^h}\left[\phi(s,a)\phi(s,a)^\top\beta_h^t\right] \\
&= \Phi^\top q_h^{\pi^t,\beta^t} - \Lambda_h\beta_h^t \\
&= \nabla_{w_h} f(\boldsymbol{\pi}_t, \boldsymbol{\beta}^t, \boldsymbol{w}_t) \quad \text{[By eq. (30)]}
\end{aligned}$$

Let $\phi_{h,j} = \phi(s_{h,j}, a_{h,j})$. Then we have,

$$\begin{aligned}
&\mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[\left\|\Phi^\top q_{h,j}^t - \phi_{h,j}\cdot\left\langle\phi_{h,j},\beta_h^t\right\rangle\right\|_2^2\right] \leq 2\mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[\left\|\Phi^\top q_{h,j}^t\right\|_2^2\right] + 2\mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[\left\|\phi_{h,j}\cdot\left\langle\phi_{h,j},\beta_h^t\right\rangle\right\|_2^2\right] \\
&\leq 2 + 2\cdot\mathbb{E}_{\mu_{\mathrm{ref}}^h}\left[(\beta_h^t)^\top\phi_{h,j}\phi_{h,j}^\top\beta_h^t\right] = 2 + 2\cdot\left\|\beta_h^t\right\|_{\Lambda_h}^2 \leq 2\cdot\left(1 + B^2\|\Lambda_h\|_2^2\right)
\end{aligned}$$

The second inequality uses the fact that the norm of the features is bounded by one, and exactly one entry of $q^t_{h,j}$ is set to one. The above bound also implies that $\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\overline{g}^t_{w_h}\right\|_2^2\right] \le 2 \cdot \left(1 + B^2 \left\|\Lambda_h\right\|_2^2\right)$. Now, observe that ε-fraction of the dataset $\mathcal{D}^{t,h}_1$ is corrupted, and we apply robust mean to obtain the estimator $\widetilde{g}^t_{w_h}$. Therefore, we can apply lemma 16 with $\sigma^2 = 4 \cdot \left(1 + B^2 \left\|\Lambda_h\right\|_2^2\right)$ to obtain the following bound (as long as $K \ge \Theta((d/\varepsilon)\log d)$).

$$\left\|\widetilde{q}^t_{w_h} - \nabla_{w_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\|_2 \le O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) \tag{38}$$

The above bound also implies the following upper bound on the L_2-norm $\widetilde{g}^t_{w_h}$.

$$\begin{aligned}
\left\|\widetilde{g}^t_{w_h}\right\|_2 &\le O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) + \left\|\nabla_{w_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\|_2 \\
&\le O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) + \left\|\Phi^\top q_h^{\pi^t,\beta^t} - \Lambda_h \beta^t_h\right\|_2 \\
&\le O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) + \left\|\Phi^\top q_h^{\pi^t,\beta^t}\right\|_2 + \left\|\Lambda_h \beta^t_h\right\|_2 \\
&\le O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) + \sum_{s,a} q_h^{\pi^t,\beta^t}(s,a) \left\|\phi(s,a)\right\|_2 + \left\|\Lambda_h\right\|_2 \left\|\beta^t_h\right\|_2 \\
&\le O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) + 1 + B\left\|\Lambda_h\right\|_2 = O\left(B\left\|\Lambda_h\right\|_2\right)
\end{aligned}$$

The penultimate inequality uses the fact that $q_h^{\pi^t,\beta^t}$ is a probability distribution over the state, action pairs and the feature norms are bounded by one.

Let us write $\widetilde{g}^t_{\mathbf{w}} = (\widetilde{g}^t_{w_1}, \ldots, \widetilde{g}^t_{w_H})$. Then $\|\widetilde{g}^t_{\mathbf{w}}\|_2^2 \le O\left(B^2 \sum_{h=1}^H \left\|\Lambda_h\right\|_2^2\right)$. Furthermore, for any t and h, $\|w^t_h\|_2^2 \le W^2$. Therefore, $\|\mathbf{w}^t\|_2^2 \le HW^2$. So we can apply lemma (14) to obtain the following bound.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \left\langle w_{t,h} - w^\star_h, \nabla_{w_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\rangle \\
&\le \frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \left\langle w_{t,h} - w^\star_h, \mathbb{E}\left[\widetilde{g}^t_{w_h}\right]\right\rangle + \frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \left\|w_{t,h} - w^\star_h\right\|_2 \cdot O\left(\sqrt{\varepsilon} B \left\|\Lambda_h\right\|_2\right) \\
&\le O\left(\sqrt{\varepsilon} W B \sum_{h=1}^H \left\|\Lambda_h\right\|_2\right) + \frac{HW^2}{2\eta_w T} + O\left(\eta_w B^2 \sum_{h=1}^H \left\|\Lambda_h\right\|_2^2\right)
\end{aligned}$$

□

Lemma 13. *Assume diam($\mathcal{B}$) $\le B$, diam($\mathcal{W}$) $\le W$, and $K \ge \Theta((d/\varepsilon)\log d)$. Then we have,*

$$\frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \left\langle \beta^\star_h - \beta_{t,h}, \nabla_{\beta_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\rangle \le O\left(\sqrt{\varepsilon}(\sqrt{d} + W) \sum_{h=1}^H \left\|\Lambda_h\right\|_2\right) + \frac{HB^2}{2\eta_b T} + O\left(\eta_b (d + W^2) \sum_{h=1}^H \left\|\Lambda_h\right\|_2^2\right)$$

with constant probability.

Proof. Recall that algorithm (7) defines $v^t_h(s) = \sum_a \pi^t_h(a \mid s) \cdot \phi(s,a)^\top w^t_h$. Let us define the gradient $\overline{g}^t_{\beta_h}$ as follows.

$$\overline{g}^t_{\beta_h} = \begin{cases} \frac{1}{K}\sum_{j=1}^K \phi(s_{h,j}, a_{h,j}) \left(r_{h,j} + v_{h+1}(s'_{h,j}) - \langle w^t_h, \phi(s_h, a_h)\rangle\right) & \text{if } h \in [H-1] \\ \frac{1}{K}\sum_{j=1}^K \phi(s_{h,j}, a_{h,j}) \left(r_{h,j} - \langle w^t_h, \phi(s_{h,j}, a_{h,j})\rangle\right) & \text{if } h = H \end{cases}$$

We will write $\phi_{h,j} = \phi(s_{h,j}, a_{h,j})$. Then for any $h \in [H-1]$ we have,

$$\begin{aligned}
&\mathbb{E}_{\mu^h_{\text{ref}}}\left[\frac{1}{K}\sum_{j=1}^{K}\phi_{h,j}\left(\theta_h^\top \phi_{h,j} + v^t_{h+1}(s'_{h,j}) - \langle w^t_h, \phi_{h,j}\rangle\right)\right] \\
&= \mathbb{E}_{\mu^h_{\text{ref}}}\left[\phi_{h,j}\left(\theta_h^\top \phi_{h,j} + v^t_{h+1}(s'_{h,j}) - \langle w^t_h, \phi_{h,j}\rangle\right)\right] \\
&= \mathbb{E}_{(s,a)\sim\mu^h_{\text{ref}}}\left[\phi(s,a)\phi(s,a)^\top\theta_h\right] + \mathbb{E}_{(s,a)\sim\mu^h_{\text{ref}}}\left[\sum_{s'} P_h(s' \mid s,a) v^t_{h+1}(s')\phi(s,a)\right] - \mathbb{E}_{(s,a)\sim\mu^h_{\text{ref}}}\left[\phi(s,a)\phi(s,a)^\top w^t_h\right] \\
&= \Lambda_h(\theta_h - w^t_h) + \sum_{s'} \Lambda_h \boldsymbol{\mu}_h(s') v^{\pi^t,w}_{h+1}(s') \\
&= \nabla_{\beta_h} f(\boldsymbol{\beta}^t, \boldsymbol{\pi}^t, \boldsymbol{w}^\star_t) \quad \text{[By eq. (32)]}
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\phi_{h,j}\left(\theta_h^\top \phi_{h,j} + v^t_{h+1}(s'_{h,j}) - \langle w^t_h, \phi_{h,j}\rangle\right)\right\|_2^2\right] \\
&\le 2\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\phi_{h,j}\left(\theta_h^\top \phi_{h,j} + v^t_{h+1}(s'_{h,j})\right)\right\|_2^2\right] + 2\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\phi_{h,j}\left\langle w^t_h, \phi_{h,j}\right\rangle\right\|_2^2\right] \\
&\le 4\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\phi_{h,j}\cdot\theta_h^\top \phi_{h,j}\right\|_2^2\right] + 4\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\phi_{h,j}\cdot v^t_{h+1}(s'_{h,j})\right\|_2^2\right] + 2\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\phi_{h,j}\left\langle w^t_h, \phi_{h,j}\right\rangle\right\|_2^2\right] \\
&\le 4\mathbb{E}_{\mu^h_{\text{ref}}}\left[\|\phi(s,a)\|_2^2\,\theta_h^\top\phi(s,a)\phi(s,a)^\top\theta_h\right] + 4\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\sum_{s'} P_h(s' \mid s,a) v^t_{h+1}(s')\cdot\phi(s,a)\right\|_2^2\right] \\
&+ 2\mathbb{E}_{\mu^h_{\text{ref}}}\left[(w^t_h)^\top\phi(s,a)\phi(s,a)^\top w^t_h\right] \\
&\le 4\left\|\theta^t_h\right\|^2_{\Lambda_h} + 2\left\|w^t_h\right\|^2_{\Lambda_h} + 4\mathbb{E}_{\mu^h_{\text{ref}}}\left[(w^t_h)^\top\phi(s,a)\phi(s,a)^\top w^t_h\right] \\
&\le 4\left\|\theta^t_h\right\|^2_{\Lambda_h} + 6\left\|w^t_h\right\|^2_{\Lambda_h} \le \left(4d + 6W^2\right)\|\Lambda_h\|_2^2
\end{aligned}$$

The fourth inequality uses the definition of v^t_{h+1} and $\|\phi(s,a)\|_2 \le 1$. The final inequality uses $\|\theta^t_h\|_2 \le \sqrt{d}$ and $\|w^t_h\|_2 \le W$. The above bound implies that for any $h \in [H]$, $\mathbb{E}_{\mu^h_{\text{ref}}}\left[\left\|\overline{g}^t_{\beta_h}\right\|_2^2\right] \le \left(4d + 6W^2\right)\|\Lambda_h\|_2^2$. Now, observe that ε-fraction of the dataset $\mathcal{D}^{t,h}_2$ is corrupted, and we apply robust mean to obtain the estimator $\widetilde{g}^t_{\beta_h}$. Therefore, we can apply Lemma (16) with $\sigma^2 = \left(4d + 6W^2\right)\|\Lambda_h\|_2^2$ to obtain the following bound (as long as $K \ge \Theta((d/\varepsilon)\log d)$.

$$\left\|\widetilde{g}^t_{\beta_h} - \nabla_{\beta_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\|_2 \le O(\sqrt{\varepsilon(d + W^2}\,\|\Lambda_h\|_2) \tag{39}$$

Furthermore, the above bound also implies the following upper bound on the L_2-norm of $\widetilde{g}^t_{\beta_h}$.

$$\begin{aligned}
\left\|\widetilde{g}^t_{\beta_h}\right\|_2 &\le O\left(\sqrt{\varepsilon}(\sqrt{d} + W)\,\|\Lambda_h\|_2\right) + \left\|\nabla_{\beta_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\|_2 \\
&\le O\left(\sqrt{\varepsilon}(\sqrt{d} + W)\,\|\Lambda_h\|_2\right) + \left\|\Lambda_h(\theta_h - w_h) + \sum_{s'} v^{\pi,w}_{h+1}(s')\Lambda_h\boldsymbol{\mu}_h(s')\right\|_2
\end{aligned}$$

From the definition of value function we have $v^{\pi,w}_{h+1}(s') \le \left|\sum_{b'} \pi^t_{h+1}(b' \mid s')\phi(s',b')^\top w^t_{h+1}\right| \le \sum_{b'} \pi^t_{h+1}(b' \mid s')\,\|\phi(s',b')\|_2\,\left\|w^t_{h+1}\right\|_2 \le W$ as feature norms are bounded by one. This result gives us the following upper bound.

$$\left\|\widetilde{g}^t_{\beta_h}\right\|_2 \le O\left(\sqrt{\varepsilon}(\sqrt{d} + W)\,\|\Lambda_h\|_2\right) + \|\Lambda_h(\theta_h - w_h)\|_2 + \|\boldsymbol{\mu}_h\Lambda_h\|_2 \le O\left((\sqrt{d} + W)\,\|\Lambda_h\|_2\right)$$

Let us now write $\widetilde{g}^t_{\boldsymbol{\beta}} = (\widetilde{g}^t_{w_1}, \ldots, \widetilde{g}^t_{w_H})$. Then $\left\|\widetilde{g}^t_{\boldsymbol{\beta}}\right\|_2^2 \le O\left((d + W^2)\sum_{h=1}^{H}\|\Lambda_h\|_2^2\right)$. Furthermore, for any t and

h, $\|\beta_h\|_2 \le B$. Therefore, $\|\boldsymbol{\beta}\|_2^2 \le HB^2$. So we can apply Lemma 14 to obtain the following bound.

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\left\langle \beta_h^\star - \beta_{t,h}, \nabla_{\beta_h} f(\boldsymbol{\pi}^t, \boldsymbol{\beta}^t, \boldsymbol{w}_t)\right\rangle \\
&\le \frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\left\langle \beta_h^\star - \beta_{t,h}, \mathbb{E}\left[\widetilde{g}^t_{\beta_h}\right]\right\rangle + \frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\|\beta_{t,h} - \beta_h^\star\|_2 \cdot O\left(\sqrt{\varepsilon}(\sqrt{d} + W)\,\|\Lambda_h\|_2\right) \\
&\le O\left(\sqrt{\varepsilon}(\sqrt{d} + W)\sum_{h=1}^{H}\|\Lambda_h\|_2\right) + \frac{HB^2}{2\eta_b T} + O\left(\eta_b(d + W^2)\sum_{h=1}^{H}\|\Lambda_h\|_2^2\right)
\end{aligned}
$$

□

Lemma 14 (Online Stochastic Gradient Descent)**.** *Let $y_1 \in W$, and $\eta > 0$. Define the sequence $y_2, \ldots, y_{n+1}$ and $h_1, \ldots, h_n$ such that for $k = 1, \ldots, n$*

$$
y_{k+1} = \mathit{Proj}_W\left(y_k + \eta \widehat{h}_k\right)
$$

and $\widehat{h}_k$ satisfies $\mathbb{E}\left[\widehat{h}_k \mid \mathcal{F}_{k-1}\right] = h_k$ and $\mathbb{E}\left[\left\|\widehat{h}_k\right\|_2^2 \mid \mathcal{F}_{k-1}\right] \le G^2$. Then for any $y^\star \in W$,

$$
\mathbb{E}\left[\sum_{k=1}^{n}\langle y^\star - y_k, h_k\rangle\right] \le \frac{\|y_1 - y^\star\|_2^2}{2\eta} + \frac{\eta n G^2}{2}.
$$

Lemma 15 (Mirror Descent, Lemma D.2 of Gabbianelli et al. (2024))**.** *Let $q_1, q_2, \ldots, q_T$ be a sequence of functions from $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ so that $\|q_t\|_\infty \le D$. Given an initial policy π_1, and a learning rate $\alpha > 0$, define a sequence of policies*

$$
\pi_{t+1}(a \mid s) \propto \pi_t(a \mid s)e^{\alpha q_t(s,a)}
$$

for $t = 1, 2, \ldots, T-1$. Then for any comparator policy $\pi^\star$,

$$
\frac{1}{T}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} q^{\pi^\star}(s)\left\langle \pi^\star(\cdot \mid s) - \pi_t(\cdot \mid s), q_t(s,\cdot)\right\rangle \le \frac{\mathcal{H}(\pi^\star\|\pi_1)}{T\alpha} + \frac{\alpha D^2}{2}
$$

Lemma 16 (Diakonikolas et al. (2017), Theorem 3.2)**.** *Let P be a distribution on $\mathbb{R}^d$ with unknown mean vector μ and unknown covariance matrix $\Sigma \preccurlyeq \sigma^2 \cdot \mathrm{Id}$. Let S be an ε-corrupted set of samples from P of size $\Theta((d/\varepsilon)\log d)$. There exists an efficient algorithm that, on input S and $\varepsilon > 0$, with probability $9/10$ outputs $\widehat{\mu}$ with $\|\widehat{\mu} - \mu\|_2 \le O\left(\sqrt{\varepsilon}\sigma\right)$.*

Lemma 17. *Let P be a distribution on $\mathbb{R}^d$ with unknown mean vector μ and unknown covariance matrix Σ. Suppose $\mathrm{cov}_{X\sim P}(XX^\top) \preccurlyeq \sigma^4\mathrm{Id}$. Let S be an ε-corrupted set of samples from P of size $\Theta((d^2/\varepsilon^2)\log^2 d)$. There exists an efficient algorithm that, on input S and $\varepsilon > 0$, with probability $9/10$ outputs $\widehat{\mu}$ with $\left\|\widehat{\Sigma} - \Sigma\right\|_2 \le O\left(\sqrt{\varepsilon}\sigma^2\right)$.*

Proof. Apply robust mean estimation on the set of flattened vectors $\left\{xx^\top : x \in S\right\}$. See also Diakonikolas and Kane (2019), subsection 3.2. □

Lemma 18 (Approximate Subgradient)**.** *Let $f(x) = \max_{i\in[m]} f_i(x)$ where each f_i is closed and convex. Let $j \in [m]$ be a β_1-approximate optimizer i.e. $f_j(x) \ge f(x) - \beta_1$. If v is a β_2-approximate subgradient of f_j at x, then v is a $(\beta_1 + \beta_2)$-approximate subgradient of f at x.*

Proof. Since v is a β_2-approximate subgradient of f_j at x, for any y we have,

$$
f(y) = \max_i f_i(y) \ge f_j(y) \ge f_j(x) - \beta_2 + \langle v, y - x\rangle \ge f(x) - (\beta_1 + \beta_2) + \langle v, y - x\rangle.
$$

□