Adversarially-Robust TD Learning with Markovian Data: Finite-Time Rates and Fundamental Limits

Sreejeet Maity
North Carolina State University

Aritra Mitra
North Carolina State University

Abstract

One of the most basic problems in reinforcement learning (RL) is policy evaluation: estimating the long-term return, i.e., value function, corresponding to a given fixed policy. The celebrated Temporal Difference (TD) learning algorithm addresses this problem, and recent work has investigated finite-time convergence guarantees for this algorithm and variants thereof. However, these guarantees hinge on the reward observations being always generated from a well-behaved (e.g., sub-Gaussian) true reward distribution. Motivated by harsh, real-world environments where such an idealistic assumption may no longer hold, we revisit the policy evaluation problem from the perspective of *adversarial robustness*. In particular, we consider a Huber-contaminated reward model where an adversary can arbitrarily corrupt each reward sample with a small probability ϵ. Under this observation model, we first show that the adversary can cause the vanilla TD algorithm to converge to any arbitrary value function. We then develop a novel algorithm called `Robust-TD` and prove that its finite-time guarantees match that of vanilla `TD` with linear function approximation up to a small $O(\epsilon)$ term that captures the effect of corruption. We complement this result with a minimax lower bound, revealing that such an additive corruption-induced term is unavoidable. To our knowledge, these results are the first of their kind in the context of adversarial robustness of stochastic approximation schemes driven by Markov noise. The key new technical tool that enables our results is an analysis of the Median-of-Means estimator with corrupted, time-correlated data that might be of independent interest to the literature on robust statistics.

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

1 Introduction

In recent years, a significant body of research has focused on understanding the effects of adversarial corruption on deep learning (Goodfellow et al., 2014; Madry et al., 2017). While this line of work has contributed significantly to the design of reliable and trustworthy machine-learning models, the developments have primarily catered to supervised learning (Javanmard et al., 2020). Much less is understood when an adversary poisons data arriving in an online manner in the context of reinforcement learning (RL). Arguably, one of the most fundamental problems in RL is that of *policy evaluation*, where a learner unaware of the true underlying model of a Markov Decision Process (MDP) seeks to evaluate the long-term return (i.e., the value function) associated with a given fixed policy. To do so, at each time step, it plays an action based on the policy to be evaluated, observes as data a reward, and transitions to a new state. Importantly, the rewards are always generated based on the (unknown) reward functions of the MDP. Departing from this paradigm, we consider a scenario where a small fraction of the reward samples can be *arbitrarily* corrupted by a powerful adversary possessing complete knowledge of the MDP. One would ideally like to obtain guarantees on value function estimation that *degrade gracefully* with the corruption fraction. Whether this is possible is a hitherto unexplored question that we resolve in this paper.

To provide context, in the absence of adversarial corruption, the classical Temporal Difference (TD) learning algorithm of Sutton (1988) solves the policy evaluation problem. An asymptotic analysis of `TD`(0) - the simplest TD learning algorithm - with linear function approximation was provided in the seminal work of Tsitsiklis and Van Roy (1997). More

recently, a growing body of work has provided finite-time guarantees for TD learning with linear function approximation (Dalal et al., 2018; Bhandari et al., 2018; Srikant and Ying, 2019; Patil et al., 2023; Mitra, 2024), and more general nonlinear stochastic approximation (SA) schemes (Chen et al., 2019, 2022; Guannan Qu, 2020; Chen et al., 2024). The guarantees in each of the papers above assume that the rewards are always drawn from true reward distributions linked to the underlying MDP. Moreover, the rewards are either assumed to be deterministic or generated from light-tailed sub-Gaussian distributions.

Motivation. Unfortunately, such assumptions do not adequately capture harsh, real-world environments. For instance, in large-scale, complex systems such as the power grid (Kosut et al., 2011) or multi-robot networks (Gil et al., 2017), data is collected via imperfect sensors prone to unexpected failures and adversarial attacks. Motivated by the need to safeguard against such attacks that are common in cyber-physical systems (Dibaji et al., 2019), we consider a reward contamination model where, at each time step, with probability $1-\epsilon$, the reward is generated from a true reward distribution, and with probability ϵ, from an arbitrary error distribution controlled by an adversary. Here, ϵ captures the power of the adversary. Our data poisoning model is directly inspired by the Huber model from robust statistics (Huber, 1992, 2004). Similar reward contamination models have also been widely studied in the context of multi-armed bandits (Jun et al., 2018; Lykouris et al., 2018; Liu and Shroff, 2019; Gupta et al., 2019; Kapoor et al., 2019). However, beyond bandits, when it comes to *SA schemes in RL*, no prior work has provided a finite-time analysis of the effects of such attacks. Given this premise, we ask:

Is it possible to perform accurate value function estimation under the Huber-contaminated reward model? If so, what are the fundamental limits on performance imposed by this attack model?

The main difficulty in answering these questions arises from the need to deal with two different forms of uncertainty: the lack of knowledge of the MDP, and the uncertainty injected by the adversary. Furthermore, other than requiring the true reward distributions to have finite first and second moments, we make no assumptions of sub-Gaussianity. This makes it particularly challenging for the learner to distinguish between time-correlated, potentially heavy-tailed clean rewards (inliers) and adversarial outliers.

Our Contributions. In this paper, we systematically study adversarial robustness in the context of policy evaluation with linear function approximation. Our specific contributions are as follows.

• **Vulnerability of** `TD`. We start with a simple result (Theorem 1) showing that under the Huber-contaminated reward model, an adversary can cause the vanilla `TD`(0) algorithm to converge to any arbitrary point. This result directly motivates the need for adversarially robust variants of `TD`(0).

• **Robust Mean Estimation with Markov Data**. A key ingredient in our algorithmic development is that of robust mean estimation. While the literature on robust statistics has made significant advances in this regard (Lai et al., 2016; Chen et al., 2015), the results we know of all assume independent and identically (i.i.d.) distributed inliers. The same is true for some recent papers in RL (Zhuang and Sui, 2021; Zhu et al., 2024) that consider heavy-tailed i.i.d. rewards with no corruption. Since the reward samples in our setting are generated based on a Markov chain, we cannot directly appeal to such existing work. To overcome this difficulty, we provide the *first analysis of the Median-of-Means (`MoM`) estimator under Huber contamination and Markovian data.* In particular, our analysis carefully exploits the ergodicity of the underlying Markov chain along with a novel coupling argument. We also note that while the popular `MoM` scheme was known to be robust to heavy-tailed data, the fact that it is also robust to adversarial corruption appears to be new. As such, we believe that our main result in this context, namely Theorem 2, and its analysis in Appendix D, might be of independent interest to robust statistics.

• **Robust-TD Algorithm**. On the algorithmic front, our main contribution is the development of an adversarially robust variant of `TD`(0) called `Robust-TD`. `Robust-TD` relies on two main new ideas: (i) a robust mean estimation step that uses historical data to construct robust empirical TD update directions; and (ii) a dynamic thresholding step that provides a second layer of safety by accounting for low-probability events where the robust mean estimation guarantees might fail to hold. As we discuss in detail in Section 5, the design of the thresholding radius is a very delicate task: unless designed carefully, one may not achieve the near-optimal rates in this paper.

• **Main Convergence Result.** Our main convergence result for `Robust-TD` establishes a mean-square error bound of the form $\tilde{O}(\bar{\tau}_{mix}/T) + O(\epsilon)$, where $\bar{\tau}_{mix}$ is the mixing time of the underlying Markov chain, T is the number of iterations, and ϵ is the corruption probability. For a specific statement of this result, see Theorem 3. When $\epsilon = 0$, our result is consistent with prior finite-time guarantees for `TD`(0) with linear function approximation (Bhandari et al., 2018; Srikant and Ying, 2019). Thus, `Robust-TD` is provably robust to adversarial reward contamination, and its guarantees match that of vanilla `TD`(0) up to a small

$O(\epsilon)$ term. Establishing this result is non-trivial as we need to contend with the complex interplay between Markovian noise, adversarial perturbations, and function approximation. We elaborate on these challenges in Sections 2 and 6. To our knowledge, Theorem 3 is the first result of its kind for stochastic approximation schemes in RL driven by Markov noise and subject to adversarial outliers.

• **Minimax Lower Bound.** To complement the upper-bound in Theorem 3, we provide an algorithm-independent lower bound in Theorem 4. The main message conveyed by this result is that the additive $O(\epsilon)$ term is *unavoidable*, and captures the fundamental price of adversarial contamination.

Overall, our algorithmic and theoretical contributions above provide a fairly complete characterization of the effects of Huber-reward-contamination on policy evaluation in general, and TD learning in particular.

Related Work. We discuss the most relevant works on *adversarial robustness in RL* below. A more elaborate survey appears in Appendix A. Data corruption in online finite-horizon episodic RL problems is studied by Lykouris et al. (2021) and Wei et al. (2022), where the notion of performance is measured by cumulative regret. Our setting is *fundamentally different* in that we consider an infinite horizon, discounted single-trajectory setting, where performance is captured by the mean-squared error w.r.t. the solution to the projected Bellman equation. Furthermore, our algorithm builds on stochastic approximation and differs significantly from the Upper-Confidence-Based (UCB)/Action-Elimination type algorithms employed in Lykouris et al. (2021); Wei et al. (2022). Corruption-robust algorithms in the offline setting are considered in Zhang et al. (2022), where data tuples are collected offline in an i.i.d. manner, and the true rewards are assumed to be sub-Gaussian. In sharp contrast, data arrives sequentially in our setting, and, as such, *we need to contend with corruption in heavy-tailed Markovian data* - a much more challenging setting. Different from the SA problem we consider here, outlier-robust policy gradient (PG) algorithms have been explored in Zhang et al. (2021), where the issue of Markovian sampling does not arise. Finally, a very recent work (Cayci and Eryilmaz, 2024) considers heavy-tailed rewards *with no adversarial corruption* in TD learning. The analysis in their paper requires a strong realizability assumption and relies on a projection step in the algorithm to control the iterates; we require neither, making it much more challenging to tackle both heavy-tailed data and adversarial perturbations. Moreover, our proposed algorithm differs considerably from that in Cayci and Eryilmaz (2024). **In summary, our work is the first to study the topic of adversarial reward corruption in the context of TD learning with function approximation and Markovian data.** As such, we do not focus on other potential forms of attack. In fact, as we shall see, the reward attack model we consider here is rich enough to merit non-trivial and subtle algorithmic ideas and analysis techniques.

2 Model and Problem Formulation

We start by reviewing the essentials of policy evaluation with linear function approximation, and then proceed to set up our problem of interest.

The Policy Evaluation Problem. We consider a Markov Decision Process (MDP) denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a finite state space of size m, $\mathcal{A}$ is a finite action space, $\mathcal{P}$ is a set of action-dependent Markov transition kernels, $\mathcal{R}$ is a reward function, and $\gamma \in (0,1)$ is the discount factor. We consider deterministic policies, where each deterministic policy $\mu : \mathcal{S} \to \mathcal{A}$ maps a state to an action.[1] A fixed policy μ induces a Markov reward process (MRP) characterized by a transition matrix P_μ, and a reward function $R_\mu : \mathcal{S} \to \mathbb{R}$. Here, $P_\mu(s, s')$ denotes the probability of transitioning from state s to state s' under the action $\mu(s)$. Associated with each state $s \in \mathcal{S}$ is a conditional reward distribution $\mathcal{D}_\mu(\cdot|s)$: whenever action $\mu(s)$ is played in state s, a noisy *random* reward $r(s)$ drawn from $\mathcal{D}_\mu(\cdot|s)$ is observed, such that $\mathbb{E}_{r(s)\sim\mathcal{D}_\mu(\cdot|s)}[r(s)] = R_\mu(s)$, and $\mathbb{E}_{r(s)\sim\mathcal{D}_\mu(\cdot|s)}[(r(s) - R_\mu(s))^2] \leq \rho^2$, where ρ is assumed to be finite. In words, upon playing $\mu(s)$ in state s, the observed noisy reward has mean $R_\mu(s)$ and variance upper-bounded by ρ^2. We assume that the mean reward at each state is uniformly bounded, i.e., $\exists \bar{r} > 0$ such that $|R_\mu(s)| \leq \bar{r}, \forall s \in \mathcal{S}$. The long-term value of a state s in the MRP induced by μ is captured by a value function $V_\mu(s)$. Formally, $V_\mu(s)$ is the discounted expected cumulative reward obtained by playing policy μ starting from initial state s:

$$V_\mu(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_\mu(s_t) | s_0 = s\right], \tag{1}$$

where s_t represents the state of the Markov chain (induced by μ) at time t, when initiated from $s_0 = s$. The primary goal of this paper is to study the *policy evaluation* problem, i.e., the problem of evaluating the value function V_μ corresponding to a given policy μ, when R_μ and P_μ are *unknown* to the learner.

Linear Function Approximation. In practice, the size of the state space $\mathcal{S}$ can be extremely large. This renders the task of estimating V_μ *exactly* (based on observations of rewards and state

[1] The assumption of deterministic policies is only for simplicity of exposition. Our results can be easily extended to account for stochastic policies.

transitions) intractable. One common approach to tackle this difficulty is to consider a parametric approximation $\hat{V}_\theta$ of V_μ in the linear subspace spanned by a set $\{\phi_k\}_{k\in[K]}$ of $K \ll m$ basis vectors, where $\phi_k = [\phi_k(1), \ldots, \phi_k(m)]^\top \in \mathbb{R}^m$. Specifically, we have $\hat{V}_\theta(s) = \sum_{k=1}^K \theta(k)\phi_k(s)$, where $\theta = [\theta(1), \ldots, \theta(K)]^\top \in \mathbb{R}^K$ is a weight vector. Let $\Phi \in \mathbb{R}^{m\times K}$ be a matrix with ϕ_k as its k-th column; we will denote the s-th row of Φ by $\phi(s) \in \mathbb{R}^K$, and refer to it as the feature vector for state s. Compactly, $\hat{V}_\theta = \Phi\theta$, and for each $s \in \mathcal{S}$, we have that $\hat{V}_\theta(s) = \langle \phi(s), \theta\rangle$. As is standard (Bhandari et al., 2018; Srikant and Ying, 2019), we assume that the columns of Φ are linearly independent, and that the feature vectors are normalized, i.e., for each $s \in \mathcal{S}$, $\|\phi(s)\|_2^2 \le 1$.

The TD(0) Algorithm. Given the above setup, the goal is to find the best parameter vector θ^* that minimizes the distance (in a suitable norm) between $\hat{V}_\theta$ and V_μ. To achieve this, we will focus on the classical TD(0) algorithm Sutton (1988) within the family of TD learning algorithms. At each time-step $t = 0, 1, \ldots$, this algorithm receives as observation a data tuple $X_t = (s_t, s_{t+1}, r_t = r(s_t))$ comprising of the current state s_t, the next state s_{t+1} reached by playing action $\mu(s_t)$, and the instantaneous reward $r_t \sim \mathcal{D}_\mu(\cdot|s_t)$. Next, we define the TD(0) update direction $g_t(\theta) = g(X_t, \theta)$ as:

$$g_t(\theta) \triangleq (r_t + \gamma\langle\phi(s_{t+1}), \theta\rangle - \langle\phi(s_t), \theta\rangle)\,\phi(s_t), \forall\theta \in \mathbb{R}^K.$$

The TD(0) update to the current parameter θ_t then takes the following form:

$$\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t), \tag{2}$$

where $\alpha_t \in (0, 1)$ is the step-size/learning rate. Under some mild technical conditions, it was shown in Tsitsiklis and Van Roy (1997) that the iterates generated by TD(0) converge almost surely to the unique solution θ^* of the projected Bellman equation $\Pi_D\mathcal{T}_\mu(\Phi\theta^*) = \Phi\theta^*$. Here, D is a diagonal matrix with entries given by the elements of the stationary distribution π of the kernel P_μ. Moreover, $\Pi_D(\cdot)$ is the projection operator onto the subspace spanned by $\{\phi_k\}_{k\in[K]}$ with respect to the inner product $\langle\cdot,\cdot\rangle_D$.

Policy Evaluation with a Corrupted Reward Model. We depart from the standard policy evaluation setting reviewed above by considering an observation model where an adversary can *arbitrarily* perturb a small fraction $\epsilon \in [0, 1/2)$ of the rewards as per the classical Huber contamination model in robust statistics (Huber, 1992, 2004; Lai et al., 2016; Chen et al., 2015). Specifically, at each time-step t, a reward $\tilde{r}_t$ is generated as follows. First, a Bernoulli random variable Z_t that takes value 1 with probability $(1-\epsilon)$ and value 0 with probability ϵ is generated independently of all prior history. If $Z_t = 1$, then $\tilde{r}_t$ is sampled from the true reward distribution $\mathcal{D}_\mu(\cdot|s_t)$. If $Z_t = 0$, then $\tilde{r}_t$ is generated from an unconstrained and unknown error distribution $\mathcal{Q}$ controlled by an adversary. Compactly, we have $\tilde{r}_t \sim (1-\epsilon)\mathcal{D}_\mu(\cdot|s_t)+\epsilon\mathcal{Q}$, where we use $(1-\epsilon)\mathcal{P}_1 + \epsilon\mathcal{P}_2$ to denote the mixture of two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$. Here, ϵ is the proportion of contamination and captures the power of the adversary. The distribution $\mathcal{Q}$ could potentially be both state- and time-dependent; furthermore, the adversarial bias injected when $\tilde{r}_t$ is sampled from $\mathcal{Q}$ is allowed to be *arbitrary*. Let us note that under the corrupted observation model, the learner is presented with a modified sequence $\{\tilde{X}_t\}$ of observations, where $\tilde{X}_t = (s_t, s_{t+1}, \tilde{r}_t)$. Given this setup, we are interested in providing precise answers to the following questions.

Q1. *Under the corrupted observation model, can we still hope to obtain a reliable estimate of the value function V_μ? If so, how can this be achieved algorithmically?*

Q2. *What are the fundamental limits on policy evaluation imposed by the Huber-contaminated reward model?*

Technical Challenges. As it turns out, answering the above questions is quite non-trivial due to several technical challenges that we outline below.

• **Noisy Heavy-Tailed Rewards.** Even in the absence of corruption, note that our observation model allows the rewards to be noisy/random. Furthermore, we do not assume that the true reward distributions are sub-Gaussian; instead, we only require them to have finite mean and variance. This is unlike recent works on TD learning (Bhandari et al., 2018; Srikant and Ying, 2019), where the rewards are assumed to be deterministic (conditioned on the state). The possibility of *heavy-tailed* uncorrupted rewards, in tandem with the lack of knowledge of the reward function R_μ, *significantly complicates the learner's task of distinguishing between clean and corrupted data.*

• **Temporal Correlations.** In the standard robust statistics literature (Huber, 1992; Lai et al., 2016; Chen et al., 2015), the inliers (i.e., clean data) are generated i.i.d from an unknown distribution. However, the observations in our setting are all part of one single Markovian trajectory and, as such, exhibit *temporal correlations*. Even in the non-adversarial setting, obtaining finite-time results under Markovian data is known to be highly non-trivial. *Moreover, contending with data that is simultaneously temporally correlated and adversarially contaminated has not been previously explored before*, thereby requiring the development of

novel algorithmic and analysis techniques - this is one of the most challenging aspects of our problem.

Additionally, the function approximation setting we consider here is much harder to tackle relative to a tabular setting. Despite the complex interplay between the challenges listed above, we will provide a precise finite-time analysis of a robust variant of TD(0) to be developed later in Section 5. In the next section, we justify the need for such a development.

3 Motivation: Vulnerability of TD(0)

The purpose of this section is to formally establish that the basic TD(0) algorithm is not robust to reward-poisoning attacks. To proceed, we make the following assumption that is standard in the analysis of RL algorithms (Tsitsiklis and Van Roy, 1997; Bhandari et al., 2018; Srikant and Ying, 2019).

Assumption 1. *The Markov chain induced by the policy μ is aperiodic and irreducible.*

Under the above assumption, when the rewards are uncorrupted, TD(0) converges to $\theta^* = -\bar{A}^{-1}\bar{b}$, where $\bar{A} = \Phi^\top D\left(\gamma P_\mu - I\right)\Phi$, and $\bar{b} = \Phi^\top D R_\mu$ are the steady-state versions of A_t and b_t, respectively, and $R_\mu \in \mathbb{R}^m$ is a reward vector stacking up the mean rewards for the different states (Tsitsiklis and Van Roy, 1997). To isolate the effect of Huber contamination, it suffices to consider a *noiseless* reward model where, whenever in state $s \in \mathcal{S}$, with probability $1-\epsilon$, the learner observes the true mean reward $R_\mu(s)$.[2] To capture corruption, consider the following error model: for each state $s \in \mathcal{S}$, whenever in state s, with probability ϵ, the learner receives a bounded, deterministic signal $C(s)$. Let $C \in \mathbb{R}^m$ be the corrupted reward vector that stacks up $C(s)$ for each $s \in \mathcal{S}$. We then have the following simple result that characterizes the limit point of TD(0) under the above reward contamination model; see Appendix C for its proof.

Theorem 1. ***(Vulnerability of TD(0))*** *Suppose Assumption 1 holds, and the step-size sequence $\{\alpha_t\}$ of TD(0) is chosen to satisfy $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$. Then, under the Huber-contaminated reward model described above, the iterates of vanilla TD(0) converge with probability 1 to $\tilde{\theta}^*$, where*

$$\tilde{\theta}^* = (1-\epsilon)\theta^* + \epsilon(-\bar{A}^{-1}\Phi^\top DC). \tag{3}$$

Furthermore, for every point $w \in \mathbb{R}^K$, there exists a corresponding choice of corrupted reward vector C_w that ensures $\tilde{\theta}^* = w$.

[2] We emphasize here that the assumption of noiseless rewards is only made for this motivating section. In the sequel, when we consider the problem of defending against reward contamination, we will work under the more general heavy-tailed noisy reward model described in Section 2.

Discussion. The above result reveals that under the standard conditions for the convergence of TD(0), the limit point of TD(0) with reward contamination is a convex combination of the true solution θ^* and a vector $-\bar{A}^{-1}\Phi^\top DC$ that can be controlled by the adversary by tuning C. The result also tells us that by carefully designing C, the adversary can cause the perturbed limit point $\tilde{\theta}^*$ to be any arbitrary point in $\mathbb{R}^K$.

The key takeaway from the above result is that even when the corruption fraction ϵ is small (but non-zero), the adversary can cause the true limit point θ^* of TD(0) to get perturbed to *any* point in $\mathbb{R}^K$. To complement this result, we illustrate in Fig. 1 a scenario where, even when the corruption fraction ϵ is merely 0.001, the mean-square error of TD(0) can be large. Motivated by the finding from Theorem 1, we will systematically design a robust version of TD(0) in the sequel. As our first step towards this goal, we develop a robust univariate mean estimator for time-correlated data in the next section.

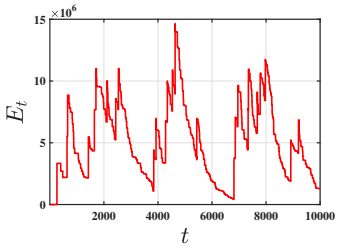

Figure 1: Plot of mean-square error E_t showing the effect of reward corruption on TD(0), with corruption probability $\epsilon = 0.001$.

4 Robust Markovian Mean Estimation

This section investigates the problem of robust mean estimation given *dependent* data samples generated by a Markov chain and corrupted as per the Huber attack model. The resulting developments will provide the key technical tools needed to tackle the robust TD learning problem. To provide context, suppose we are given N i.i.d. real-valued random variables $X_1, X_2, \ldots, X_N$ with finite mean $\mathbb{E}[X_1] = \bar{X}$, and finite variance ρ^2. The goal is to construct an estimator $\hat{X}_N = \hat{X}_N(X_1, \ldots, X_N)$, i.e., a measurable function of $X_1, \ldots, X_N$, that provides a high-probability estimate of the mean $\bar{X}$. It is well known that the empirical mean fails to provide optimal error rates in this setting (Lugosi and Mendelson, 2019). Instead, a simple yet powerful estimator known as the Median-of-Means (MoM) estimator yields the following optimal rate: given any $\delta \in (0,1)$, with probability (w.p.) at least $1-\delta$, the MoM estimate $\hat{X}_N$ satisfies: $|\hat{X}_N - \bar{X}| \leq O\left(\rho\sqrt{\log(1/\delta)/N}\right)$ (Lugosi and Mendelson, 2019). We depart from this setting by considering the *dependent-data* observation model below.

The Setting. Consider an ergodic, time-homogeneous,

and stationary Markov chain $X_1, X_2, \ldots$ with stationary distribution π. The Markov chain takes values over a finite state space $\mathcal{X}$. Let $F : \mathcal{X} \mapsto \mathbb{R}$ be a bounded function such that $|F(x)| \leq \psi, \forall x \in \mathcal{X}$, and define $\bar{f} \triangleq \mathbb{E}_{X\sim\pi}[F(X)]$. Now consider a noisy observation model characterized by a set of conditional distributions $\{\mathcal{D}(\cdot|x), x \in \mathcal{X}\}$, such that whenever in state x, the learner gets to observe a random sample $f(x)$ with mean $\mathbb{E}_{f(x)\sim\mathcal{D}(\cdot|x)}[f(x)] = F(x)$ and finite variance ρ^2. Our goal is to obtain a high-probability estimate of $\bar{f} = \mathbb{E}_{X_i\sim\pi, f(X_i)\sim\mathcal{D}(\cdot|X_i)}[f(X_i)], i \in [N]$ for the following observation model.

Noisy and Corrupted Markovian Data: The learner observes N noisy and Huber-contaminated samples $\tilde{f}_1, \ldots, \tilde{f}_N$, where $\tilde{f}_i \sim (1-\epsilon)\mathcal{D}(\cdot|X_i) + \epsilon\mathcal{Q}, i \in [N]$, and $\mathcal{Q}$ is an unknown and unconstrained *adversarial* error distribution.

For the model above, can we expect mean-estimation bounds similar to those known under i.i.d data? In what follows, we will provide an answer in the affirmative by developing a simple variant of the `MoM` estimator. Our estimator and its guarantees will depend on the notion of a mixing time τ_{mix}. To define this object, let $\mathbb{P}(X_{t+1} \in \cdot|X_1 = X)$ denote the conditional distribution of X_{t+1} given that $X_1 = X$. Next, following Dorfman and Levy (2022), define $d_{mix}(t) \triangleq \sup_{X\in\mathcal{X}} D_{TV}\left(\mathbb{P}(X_{t+1} \in \cdot|X_1 = X), \pi\right)$, where $D_{TV}(\mathcal{P}_1, \mathcal{P}_2)$ denotes the total variation distance between two probability measures $\mathcal{P}_1$ and $\mathcal{P}_2$. We then define the mixing time τ_{mix} as follows:

$$\tau_{mix}(\eta) \triangleq \inf\{t : d_{mix}(t) \leq \eta\}, \ \tau_{mix} \triangleq \tau_{mix}(1/4). \quad (4)$$

With the above model and notation in place, we are now ready to describe our estimator called `RUMEM`: Robust Univariate Mean Estimator for Markovian Data.

Description of `RUMEM`: The algorithm takes as input a data set $\mathcal{S} = \{\tilde{f}_1, \ldots, \tilde{f}_N\}$ generated as per the noisy and corrupted Markovian data model, a confidence parameter δ, the mixing time τ_{mix}, and the corruption fraction ϵ. It then performs the following operations.

1) Subsampling. The first step is to create a subsampled set $\mathcal{S}_{sub} = \{\tilde{f}_1, \tilde{f}_{\tau+1}, \ldots, \tilde{f}_{(n-1)\tau+1}\}$ by selecting every τ-th element from the original data set $\mathcal{S}$; here, $n = \lfloor (N-1)/\tau \rfloor + 1$. The rationale behind this step is to create a data set comprising approximately independent samples by choosing the parameter τ appropriately; we will specify this choice in the statement of Theorem 2.

2) Partitioning the Subsampled Set: Next, the subsampled set $\mathcal{S}_{sub}$ is split into L equal buckets denoted $\{\mathcal{B}_1, \ldots, \mathcal{B}_L\}$, where each bucket $\mathcal{B}_i$ has size given by $\lfloor n/L \rfloor$.

3) Calculating the Median-of-Means Estimate: Let the mean of the samples in the i-th bucket $\mathcal{B}_i$ be denoted $\hat{\mu}_i$. The algorithm then returns $\hat{\mu} = \texttt{Median}\{\hat{\mu}_1, \ldots, \hat{\mu}_L\}$. Other than the sub-sampling step, the rest of `RUMEM` is the same as a standard `MoM` estimator; *as such, we claim no novelty in the design of* ***`RUMEM`****. Instead, our main contribution here lies in the analysis of* ***`RUMEM`*** *under heavy-tailed, corrupted, and Markovian data.* It is this analysis that guides the choice of the two main parameters in the algorithm, namely the sub-sampling gap τ and the number of buckets L. The main result of this section is as follows.

Theorem 2. ***(Performance of `RUMEM`)*** *Consider the Huber-contaminated Markov data model described in this section with corruption fraction ϵ. Let* ***`RUMEM`*** *be run on this data set with parameters chosen as follows:*

$$\tau = \lceil \log_2(6N/\delta)\, \tau_{mix} \rceil; \ \epsilon' = \epsilon + \frac{32}{3n}\log(24/\delta); \\ L = \lceil 12\epsilon' n + \frac{256}{7}\log\left(\frac{N}{\delta}\right) \rceil.^{3} \quad (5)$$

There exists a universal constant $C \geq 1$ such that given any $\delta \in (0,1)$, if $N \geq \max\{2, 4L\tau\}$, then with probability at least $1-\delta$, the output $\hat{\mu}$ of ***`RUMEM`*** *satisfies*

$$|\hat{\mu} - \bar{f}| \leq C\max\{\psi, \rho\}\left(\sqrt{\epsilon} + \sqrt{\frac{\tau}{N}\log\left(\frac{N}{\delta}\right)}\right). \quad (6)$$

Discussion. To appreciate the bound in Theorem 2, consider first the case when the data is generated in an i.i.d. manner; here, the subsampling gap τ is simply 1. With $\tau = 1$, our bound in Eq. (6) is consistent with that known for robust univariate mean estimation under i.i.d. data with a trimmed mean estimator; see (Lugosi and Mendelson, 2021, Theorem 1). While the `MoM` estimator is known to be robust against heavy-tailed i.i.d. noise, the fact that it is also robust to adversarial noise is new. Under Markov data, our bound gets inflated by a factor of $\sqrt{\tau}$. This is also consistent with mean estimation results under Markov data since one essentially has N/τ "effective" samples (Nagaraj et al., 2020; Dorfman and Levy, 2022). The significance of Theorem 2 lies in **providing the first guarantees of robust mean estimation under both Markovian and adversarial data.** This result might be of independent interest to robust statistics, and we conjecture that it will find use in dealing with outliers in time-series data (beyond our TD learning setting). The proof of Theorem 2 is provided in Appendix D. One subtle aspect of the proof is that it needs to account for the fact that the number of "good" uncorrupted buckets is a random object. The other key ingredients in the analysis involve carefully exploiting the geometric mixing property of the underlying Markov chain along with a coupling idea in the recent paper of Dorfman and Levy (2022).

Algorithm 1 `Robust-TD` Algorithm

1: **Input:** Policy to be evaluated μ, initial estimate $\theta_0 \in \mathbb{R}^K$, corruption fraction ϵ, total number of iterations T, and burn-in time $\bar{T}$.
2: **for** $t = 0, 1, \ldots$ **do**
3: Play $\mu(s_t)$ and observe tuple $\tilde{X}_t = (s_t, s_{t+1}, \tilde{r}_t)$, where $\tilde{r}_t \sim (1-\epsilon)\mathcal{D}_\mu(\cdot|s_t) + \epsilon\mathcal{Q}$.
4: If $t \leq \bar{T}$, then maintain $\theta_t = \theta_0$.
5: **if** $t > \bar{T}$ **then**
6: Set $[\hat{b}_t]_i = \texttt{RUMEM}\left(\{y_{i,k}\}_{0\leq k\leq t}; \delta = 1/(KT^2)\right)$ with $y_{i,k} = [\phi(s_k)]_i \tilde{r}_k$ to compute $\hat{b}_t$.
7: Compute threshold G_t as per Eq. (7) and set $\sigma_1 = \max\{1, \bar{r}, \rho\}$.
8: If $\|\hat{b}_t\|_2 > G_t + \sigma_1$, then set: $\hat{b}_t \leftarrow 0$.
9: Compute Robust TD direction $\tilde{g}_t(\theta_t)$:

$$\tilde{g}_t(\theta_t) = A_t\theta_t + \hat{b}_t,\ A_t = \gamma\phi(s_t)\phi^\top(s_{t+1}) - \phi(s_t)\phi^\top(s_t).$$

10: Update parameter: $\theta_{t+1} = \theta_t + \alpha\tilde{g}_t(\theta_t)$.
11: **end if**
12: **end for**

5 Robust TD Learning Algorithm

In this section, we develop our proposed algorithm called `Robust-TD`; the steps of our method are outlined as Algorithm 1. As we shall see later, despite the presence of Huber-contaminated rewards, `Robust-TD` yields guarantees that are consistent with those provided by TD in the absence of attacks, up to a small *unavoidable* $O(\sqrt{\epsilon})$ term that captures the price of corruption. Our approach rests on two new ideas: (i) Using historical information of rewards along with the `MoM` estimator in Section 4 to construct robust TD update directions; and (ii) a carefully designed dynamic thresholding scheme to account for rare (i.e., low-probability) events. We describe these ideas below.

• **Constructing Robust TD update directions.** To build intuition, let us start by noting from Eq. (2) that the vanilla `TD`(0) update direction (without corruption) can be expressed in affine form: $g_t(\theta) = A_t\theta + b_t$, where $A_t = \gamma\phi(s_t)\phi^\top(s_{t+1}) - \phi(s_t)\phi^\top(s_t)$ and $b_t = \phi(s_t)r_t$. The main observation here is that the rewards only affect the term b_t; as such, our goal will be to obtain a robust estimate of this object. Due to Assumption 1, b_t will eventually approach its stationary value $\bar{b} = \mathbb{E}_{s_t\sim\pi, r(s_t)\sim\mathcal{D}_\mu(\cdot|s_t)}[b_t] = \sum_{s\in\mathcal{S}}\pi(s)\phi(s)R_\mu(s)$. We would thus ideally like our robust estimate to be "close" to $\bar{b}$. There are a few subtleties here. To explain them, let us consider a couple of candidate strategies. Given the structure of $\bar{b}$, one natural idea could be to use all prior observations collected for each state $s \in \mathcal{S}$ to construct estimates of $R_\mu(s)$ and $\pi(s)$ individually. However, this would require maintaining vectors of dimension equal to the size $|\mathcal{S}|$ of the state space, defeating the purpose of function approximation. Since only the rewards are corrupted, yet another strategy could be to apply the `RUMEM` estimator from Section 4 to the set of reward observations $\{\tilde{r}_k\}_{0\leq k\leq t}$ collected up to time t. While this will yield a robust estimate of $\sum_{s\in\mathcal{S}}\pi(s)R_\mu(s)$, our goal instead is to get an estimate of $\sum_{s\in\mathcal{S}}\pi(s)\phi(s)R_\mu(s)$. The main message here is that some care needs to be taken while devising an approach for estimating $\bar{b}$.

Our approach is to maintain component-wise estimates of $\bar{b}$. To that end, let $[\bar{b}]_i$ and $[\phi(s)]_i$ represent the i-th components of $\bar{b}$ and $\phi(s)$, respectively; here, $i \in [K]$. Moreover, let $\hat{b}_t$ denote the estimate of $\bar{b}$ at time t. Then, the i-th component of $\hat{b}_t$ is constructed by applying the `RUMEM` estimator in Section 4 to the data set $\{y_{i,k}\}_{0\leq k\leq t}$ with a confidence parameter $\delta = 1/(KT^2)$, where $y_{i,k} = [\phi(s_k)]_i\tilde{r}_k$, and T is the total number of iterations. We represent this operation succinctly as $[\hat{b}_t]_i = \texttt{RUMEM}\left(\{y_{i,k}\}_{0\leq k\leq t}; \delta = 1/(KT^2)\right)$; see line 6 of Algo. 1. The intuition here is simple: if a sample at time k is uncorrupted (i.e., $\tilde{r}_k = r(s_k)$), then we have $[\bar{b}]_i = \mathbb{E}_{s_k\sim\pi, r(s_k)\sim\mathcal{D}_\mu(\cdot|s_k)}[[\phi(s_k)]_i r(s_k)]$. In words, if the underlying Markov chain is stationary, then each uncorrupted sample $y_{i,k}$ provides an unbiased estimate of $[\bar{b}]_i$. Notably, however, these samples are *not independent* - this is precisely why we need to appeal to the results from Section 4.

• **Thresholding mechanism.** We now explain that the design of $\hat{b}_t$ as described above is not enough for our purpose. From Theorem 2, note that the guarantees afforded by `RUMEM` do not hold deterministically, rather only with high probability. Since we seek to obtain mean-square error bounds, *we need to thus provide an additional layer of safety for low-probability events on which the guarantees from `RUMEM` do not hold.* Accordingly, if $\|\hat{b}_t\|_2$ exceeds a carefully designed threshold $G_t + \sigma_1$, we reset $\hat{b}_t$ to 0. Here, $\forall t \geq \bar{T}$,

$$G_t \triangleq C\sqrt{K}\sigma_1\left(\sqrt{\epsilon} + 2\log(12KT^3)\sqrt{\frac{\tau_{mix}}{t}}\right), \quad (7)$$

where $\sigma_1 = \max\{1, \bar{r}, \rho\}$, C is as in Eq. (6) of Theorem 2, $K << |\mathcal{S}|$ is the number of feature vectors, $\bar{T}$ is an initial burn-in time, and $\tau_{mix} = \tau_{mix}(1/4)$ is as in Eq. (4), and corresponds to the mixing-time of the Markov chain induced by the policy μ. The design of the thresholding radius $G_t + \sigma_1$ is very delicate: if the radius is too loose, it may lead to sub-optimal bounds; if it is too tight, we might reset $\hat{b}_t$ to 0 too often unnecessarily, leading again to vacuous bounds. Our analysis in Appendix E reveals that if G_t is designed as per Eq. (7), then the resetting operation in line 8 of Algorithm 1 will get bypassed with high probability, and $\hat{b}_t$ will remain the output of the `RUMEM` operation

in line 6. In other words, the resetting operation will take place only under extreme (low-probability) events, exactly as desired. Lemma 1 in Section 6 shows that the RUMEM-based estimation scheme in conjunction with the thresholding mechanism described above yields an accurate estimate of $\bar{b}$. Finally, the iterates are updated only after an initial burn-in time $\bar{T}$ that is logarithmic in T; the exact form of $\bar{T}$ will be specified in Section 6. This ensures enough data samples have been collected for the guarantees in Section 4 to kick in.

6 Main Results

The goal of this section is to state and discuss our main results: (i) a finite-time bound for Robust-TD, and (ii) a minimax lower-bound establishing the near-optimality of the guarantee from Robust-TD. To do so, we will need to introduce a bit of notation and terminology. Let $\Sigma = \Phi^\top D \Phi$. Since Φ is full column rank, Σ is full rank with a strictly positive smallest eigenvalue $\omega < 1$. Next, recall from Section 5 that the TD(0) update direction can be expressed as $g_t(\theta) = A_t(X_t)\theta + b_t(X_t)$. Also, recall the steady-state version of A_t denoted by $\bar{A} = \mathbb{E}_{s_t \sim \pi, s_{t+1} \sim P_\mu(\cdot|s_t)}[A_t(X_t)]$. Let us now introduce the following definition of mixing time, which will play a key role in our analysis.

Definition 1. *Define* $\tau'_{mix}(\eta) \triangleq \min\{t \geq 1 : \|\mathbb{E}\left[A_k(X_k)|X_0\right] - \bar{A}\|_2 \leq \eta, \forall k \geq t, \forall X_0\}$.

Assumption 1 implies that the total variation distance between the conditional distribution $\mathbb{P}\left(s_t = \cdot | s_0 = s\right)$ and the stationary distribution π decays geometrically fast for all $t \geq 0$, regardless of the initial state $s \in \mathcal{S}$ (Levin and Peres, 2017). As a result of this geometric mixing of the Markov chain, one can show that $\tau'_{mix}(\eta)$ in Definition 1 is $O\left(\log(1/\eta)\right)$ (Chen et al., 2019). For our purpose, we set $\tau' = \tau'_{mix}(\alpha)$, where α is the step-size. Define $\bar{\tau}_{mix} \triangleq \max\{\tau', \tau_{mix}\}$, $d_t \triangleq \|\theta_t - \theta^*\|_2^2$, and $\sigma \triangleq \max\{\|\theta^*\|_2, \|\theta_0\|_2, \sigma_1\}$, where recall that $\sigma_1 = \max\{1, \bar{r}, \rho\}$. Our main result is then as follows.

Theorem 3. ***(Performance of Robust-TD)*** *Suppose Assumption 1 holds, and the initial distribution of s_0 is the steady-state distribution π. Let $G = K/(\omega^2(1-\gamma)^2)$. There exist universal constants $c_1, c_2 \geq 1$ such that if the step-size α, the burn-in time $\bar{T}$, and the number of iterations T are chosen as follows:*

$$\begin{aligned} &\alpha = \frac{4}{\omega(1-\gamma)} \frac{\log(T)}{T}, \bar{T} = \lceil c_1 \tau_{mix} \log^2(KT) \rceil, \\ &T \geq \max\left\{\bar{T} + \tau', \frac{c_2 \tau' \log(T)}{\omega^2(1-\gamma)^2}\right\}, \end{aligned} \tag{8}$$

then Robust-TD guarantees:

$$\mathbb{E}[d_T] \leq \tilde{O}\left(\frac{\bar{\tau}_{mix}\sigma^2 G}{T}\right) + O\left(\epsilon \sigma_1^2 G\right). \tag{9}$$

Lower Bound Analysis. From Theorem 3, we infer that despite adversarial contamination, the iterates generated by Robust-TD converge (in the mean-square sense) at a rate of $\tilde{O}(1/T)$ to a small ball of radius $O(\epsilon)$ around the optimal parameter θ^*. Our next goal is to prove an information-theoretic lower bound to establish that the $O(\epsilon)$ additive error *cannot be avoided, in general.* To do so, it suffices to consider a simpler i.i.d observation model where at each time-step t, s_t is sampled independently from the stationary distribution π, and s_{t+1} from $P_\mu(\cdot|s_t)$. We also consider a simpler tabular setting where the feature matrix Φ is the identity matrix of dimension $|\mathcal{S}|$. Next, we use $\mathcal{M}(\epsilon, \rho, \mathcal{Q})$ to represent the set of all MRPs with finite state and action spaces and bounded mean rewards, where the reward random variable $\tilde{r}(s)$ is sampled as $\tilde{r}(s) \sim (1-\epsilon)\mathcal{D}_\mu(\cdot|s) + \epsilon\mathcal{Q}$, and the noise distribution $\mathcal{D}_\mu(\cdot|s)$ has variance at most ρ^2. We will use the shorthand $V \in \mathcal{M}(\epsilon, \rho, \mathcal{Q})$ to mean that V is the true value function associated with some MRP in the set $\mathcal{M}(\epsilon, \rho, \mathcal{Q})$. Now, suppose the learner is presented with T independent samples $\tilde{X}_1, \ldots, \tilde{X}_T$, where $\tilde{X}_t = (s_t, s_{t+1}, \tilde{r}(s_t)), t \in [T]$. An estimator $\hat{V}_T$ is some measurable function of these T samples. We then have the following *fundamental* lower bound.

Theorem 4. ***(Lower Bound)*** *There exists a universal constant $\tilde{c} > 0$ such that*

$$\inf_{\hat{V}_T} \sup_{V \in \mathcal{M}(\epsilon,\rho,\mathcal{Q})} \mathbb{P}\left(\|\hat{V}_T - V\|_2 \geq \frac{\tilde{c}\rho\sqrt{\epsilon}}{(1-\gamma)}\right) \geq \frac{1}{2}.$$

Before providing proof sketches for our main results, several remarks are in order.

Discussion. To put our result in Theorem 3 into perspective, let us note that in the absence of corruption, i.e., when $\epsilon = 0$, our convergence bound in Eq. (9) is consistent - up to log factors - with prior results on TD(0) with linear function approximation (Bhandari et al., 2018; Srikant and Ying, 2019). In particular, the dependence of the first term in the R.H.S. of Eq. (9) on each of the parameters $\bar{\tau}_{mix}, \omega, \gamma$, and T match those for vanilla TD(0) in (Bhandari et al., 2018, Theorem 3).

When $\epsilon \neq 0$, the R.H.S. of Eq. (9) features an additive $O(\epsilon)$ term. At this stage, it is instructive to compare this term with the analogous $O(\epsilon)$ term in Eq. (3). Crucially, with the basic TD(0) algorithm, the $O(\epsilon)$ term is affected by the **magnitude** of the attack corruption through the corruption vector C (see (3)). In contrast, with Robust-TD, the $O(\epsilon)$ term in the mean square error is *completely unaffected by the magnitude of the attack inputs* and depends only on instance-dependent parameters. Specifically, the $O(\epsilon)$ term in Eq. (9) is scaled by the "variance" ρ^2 of our noisy observation model; recall here that $\sigma_1 = \max\{1, \bar{r}, \rho\}$. We note

such a $O(\epsilon)$ term - inflated by the noise variance - has been proven to be unavoidable in general for robust mean estimation (Chen et al., 2015; Lai et al., 2016; Cheng et al., 2019; Dalalyan and Minasyan, 2022). Similar unavoidable terms that capture the price of adversarial contamination also show up for multi-armed bandits with reward corruptions (Lykouris et al., 2018; Gupta et al., 2019; Kapoor et al., 2019). **Our work complements these results, and its significance lies in providing the first provable guarantees of adversarial robustness for TD learning.**

However, one might still ask: *In the context of policy evaluation in RL, is the $O(\epsilon)$ term inevitable or simply an artifact of our analysis?* Theorem 4 settles this question by establishing that it is the former. Comparing the lower bound in Theorem 4 with the upper bound in Eq. (9), we also infer that the dependencies on the noise variance ρ and the discount factor γ via the $(1-\gamma)^{-1}$ term, as they appear in the $O(\epsilon)$ term of Eq. (9), are tight.

Finally, observe that the corruption-affected $O(\epsilon)$ term in Eq. (9) is inflated by $(1/\omega^2)$, where $\omega > 0$ is the smallest eigenvalue of the steady-state feature covariance matrix $\Sigma = \Phi^\top D \Phi$. To gain some intuition, suppose for the moment that Φ is the identity matrix of order m. In this case, ω is simply the smallest entry in the steady-state distribution vector π. A small value of ω implies that the corresponding state is visited infrequently, that is, there is a paucity of data for such a state. Intuitively, this should favor the adversary, and make it harder to get a reliable estimate of the value function corresponding to the state that gets visited least frequently. Our upper bound captures this intuitive phenomenon; however, at the moment, we do not have a lower bound to support the dependence on $1/(\omega^2)$. Aside from this shortcoming, Theorems 3 and 4 collectively paint a fairly complete picture of the problem of TD learning with Huber-contaminated adversarial rewards. To corroborate our theory, we provide various experiments on synthetic data in Appendix G.

We conclude this section by providing proof sketches for our main results.

Proof Sketch for Theorem 3. There are two different bias terms that affect the learning dynamics of `Robust-TD`: the term $\langle \theta_t - \theta^*, (A_t - \bar{A})\theta_t \rangle$, capturing the effect of Markov sampling; and the term $\langle \theta_t - \theta^*, \hat{b}_t - \bar{b} \rangle$, capturing the effect of adversarial perturbations. Our setting is particularly complicated because these two terms are *coupled*: the adversarial bias term affects the iterate θ_t, which shows up in the Markovian bias term. *Controlling this coupling in a way that leads to near-optimal guarantees - as in Theorem 3 - has not appeared in prior RL work.* The key new technical ingredient unique to our analysis is the following lemma; its proof exploits the bound in Theorem 2.

Lemma 1. ***(Adversarial Perturbation Bound)*** *Under the conditions in the statement of Theorem 3, the following is true for all $t \geq \bar{T}$:*

$$\mathbb{E}[\|\hat{b}_t - \bar{b}\|_2^2] \leq O\left(\epsilon + \frac{\log^2(KT)\tau_{mix}}{t}\right) K\sigma_1^2.$$

While this lemma helps us control the effect of the adversarial bias term, we need additional work to handle the effects of Markovian bias and function approximation. The thresholding operation in line 8 of `Robust-TD` helps us in this regard. The rest of the analysis proceeds by carefully leveraging the mixing properties of the underlying Markov chain in tandem with the bounds on each of the bias terms; the detailed steps are provided in Appendix E. Notably, using the ergodicity of the Markov chain to handle outliers in time-correlated data is novel to our analysis, and might be of broader interest to both RL and robust statistics. Note that the assumption that the initial state is distributed as per the steady-state distribution ensures that the resulting Markov chain is stationary - this is a standard assumption made in prior work (Bhandari et al., 2018) to simplify some of the expressions.

Proof Sketch for Theorem 4. The proof of this result, detailed in Appendix F, relies on carefully constructing two different MRPs and associated attack distributions, such that (i) the value functions in the two MRPs differ in magnitude by $\Omega(\rho\sqrt{\epsilon}/(1-\gamma))$; and (ii) the distributions of the samples in the two MRPs is indistinguishable to a learner. We then leverage ideas to prove minimax lower bounds (Wainwright, 2019a, Chapter 15) from statistical learning theory.

7 Conclusion and Future Work

We conducted the first principled study of TD learning with linear function approximation under a Huber-contaminated reward model. We started by showing that the basic TD algorithm is vulnerable to reward poisoning. We then developed a robust TD algorithm by drawing on median-of-means estimators, and by constructing a novel dynamic thresholding scheme. By establishing nearly matching upper and lower bounds, we showed that our proposed approach is provably robust to adversarial reward contamination. As future work, we plan to generalize our algorithm and results to nonlinear stochastic approximation schemes such as Q-learning; some preliminary results in this regard are reported in Maity and Mitra (2024). We also plan to consider other attack models and derive finer lower bounds that account for Markov sampling.

References

Adibi, A., Dal Fabbro, N., Schenato, L., Kulkarni, S., Poor, H. V., Pappas, G. J., Hassani, H., and Mitra, A. (2024). Stochastic approximation with delayed updates: Finite-time rates under markovian sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR.

Bogunovic, I., Krause, A., and Scarlett, J. (2020). Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1071–1081. PMLR.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. (2021). Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Borkar, V. S. and Meyn, S. P. (2000). The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469.

Boucheron, S. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Cayci, S. and Eryilmaz, A. (2024). Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36.

Cayci, S., Satpathi, S., He, N., and Srikant, R. (2023). Sample complexity and overparameterization bounds for temporal difference learning with neural network approximation. *IEEE Transactions on Automatic Control*.

Chen, M., Gao, C., and Ren, Z. (2015). Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2024). A lyapunov theory for finite-sample guarantees of markovian stochastic approximation. *Operations Research*, 72(4):1352–1367.

Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. (2022). Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. (2019). Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, page 4.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019). High-dimensional robust mean estimation in nearly-linear time. In *Proc. of the thirtieth annual ACM-SIAM symp. on discrete algorithms*, pages 2755–2771. SIAM.

Chung, F. and Lu, L. (2006). Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127.

Dal Fabbro, N., Adibi, A., Mitra, A., and Pappas, G. J. (2024). Finite-time analysis of asynchronous multi-agent td learning. In *2024 American Control Conference (ACC)*, pages 2090–2097. IEEE.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Dalalyan, A. S. and Minasyan, A. (2022). All-in-one robust estimator of the gaussian mean. *The Annals of Statistics*.

Dibaji, S. M., Pirani, M., Flamholz, D. B., Annaswamy, A. M., Johansson, K. H., and Chakrabortty, A. (2019). A systems and control perspective of cps security. *Annual reviews in control*, 47:394–411.

Dorfman, R. and Levy, K. Y. (2022). Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR.

Garcelon, E., Roziere, B., Meunier, L., Tarbouriech, J., Teytaud, O., Lazaric, A., and Pirotta, M. (2020). Adversarial attacks on linear contextual bandits. *arXiv preprint arXiv:2002.03839*.

Gil, S., Kumar, S., Mazumder, M., Katabi, D., and Rus, D. (2017). Guaranteeing spoof-resilient multi-robot networks. *Autonomous Robots*, 41:1383–1400.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guannan Qu, A. W. (2020). Finite-time analysis of asynchronous stochastic approximation and q-learning. *Proceedings of Machine Learning Research*, 125:1–21.

Gupta, A., Koren, T., and Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR.

He, J., Zhou, D., Zhang, T., and Gu, Q. (2022). Nearly optimal algorithms for linear contextual

bandits with adversarial corruptions. *arXiv preprint arXiv:2205.06811*.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. (2020). Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR.

Jun, K.-S., Li, L., Ma, Y., and Zhu, X. (2018). Adversarial attacks on stochastic bandits. *arXiv preprint arXiv:1810.12188*.

Kapoor, S., Patel, K. K., and Kar, P. (2019). Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715.

Korda, N. and La, P. (2015). On td(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International conference on machine learning*, pages 626–634. PMLR.

Kosut, O., Jia, L., Thomas, R. J., and Tong, L. (2011). Malicious data attacks on the smart grid. *IEEE Transactions on Smart Grid*, 2(4):645–658.

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE.

Lakshminarayanan, C. and Szepesvári, C. (2017). Linear stochastic approximation: Constant step-size and iterate averaging. *arXiv preprint arXiv:1709.04073*.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2024). Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236.

Liu, F. and Shroff, N. (2019). Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR.

Liu, R. and Olshevsky, A. (2021). Temporal difference learning as gradient splitting. In *International Conference on Machine Learning*, pages 6905–6913. PMLR.

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

Lugosi, G. and Mendelson, S. (2021). Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410.

Lykouris, T., Mirrokni, V., and Paes Leme, R. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122.

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.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Maity, S. and Mitra, A. (2024). Robust q-learning under corrupted rewards. *arXiv preprint arXiv:2409.03237*.

Minsker, S. (2018). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

Mitra, A. (2024). A simple finite-time analysis of td learning with linear function approximation. *IEEE Transactions on Automatic Control*, 70(2):1388–1394.

Mitra, A., Pappas, G. J., and Hassani, H. (2023). Temporal difference learning with compressed updates: Error-feedback meets reinforcement learning. *arXiv preprint arXiv:2301.00944*.

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. (2020). Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676.

Narayanan, C. and Szepesvári, C. (2017). Finite time bounds for temporal difference learning with function approximation: Problems with some "state-of-the-art" results. Technical report, Technical report.

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585.

Patil, G., Prashanth, L., Nagaraj, D., and Precup, D. (2023). Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5438–5448. PMLR.

Prashanth, L., Korda, N., and Munos, R. (2021). Concentration bounds for temporal difference learning with linear function approximation: the case of batch data and uniform sampling. *Machine Learning*, 110(3):559–618.

Shah, D. and Xie, Q. (2018). Q-learning with nearest neighbors. *Advances in Neural Information Processing Systems*, 31.

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

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

Szepesvári, C. (2022). *Algorithms for reinforcement learning*. Springer nature.

Tian, H., Paschalidis, I. C., and Olshevsky, A. (2023). On the performance of temporal difference learning with neural networks. *arXiv preprint arXiv:2312.05397*.

Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. In *IEEE Transactions on Automatic Control*.

Wainwright, M. J. (2019a). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wainwright, M. J. (2019b). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. *arXiv preprint arXiv:1905.06265*.

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.

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, J., Wan, R., Qi, Z., Luo, S., and Shi, C. (2024). Robust offline reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 541–549. PMLR.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No.

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, we provide complete proofs in the Supplementary Material, and a proof sketch of our main results in Section 5.
   (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. Yes, while we do not provide a code, we provide all the details needed to reproduce our "toy" simulations in the Appendix.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, these are provided in the Appendix.
   (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.

A Additional Literature Review

In this section, we provide a more detailed discussion of the relevant threads of literature.

1. **Temporal Difference Learning and Stochastic Approximation.** The family of TD learning methods was introduced by Sutton in Sutton (1988). The initial analysis of this algorithm with linear function approximation was carried out in Tsitsiklis and Van Roy (1997) by drawing on tools from the rich area of stochastic approximation theory (Borkar, 2009; Borkar and Meyn, 2000). The nature of the analysis in these works was *asymptotic*, i.e., no convergence rates were provided. The next set of results (Korda and La, 2015; Lakshminarayanan and Szepesvári, 2017; Dalal et al., 2018; Narayanan and Szepesvári, 2017; Prashanth et al., 2021) on this topic did manage to provide finite-time rates for TD methods; however, these results were obtained assuming that the samples are drawn i.i.d. from the steady state distribution of the underlying Markov chain. The i.i.d. assumption was first relaxed in Bhandari et al. (2018), where the authors relied on a projection step in their analysis. An analysis without the projection step was then provided in Srikant and Ying (2019), and more recently by Mitra (2024) based on a novel inductive proof technique. An interesting interpretation of the TD update direction was provided in Liu and Olshevsky (2021) by introducing the notion of "gradient-splitting". We note that in Bhandari et al. (2018); Srikant and Ying (2019); Liu and Olshevsky (2021); Mitra (2024), the authors characterize finite-time bounds under linear function approximation in terms of an ℓ_2-error metric. Our setting is similar. Complementary to ℓ_2-error bounds, the work of Pananjady and Wainwright (2020) provides ℓ_∞ bounds for the least squares temporal difference learning (LSTD) algorithm for a tabular setting, under the assumption of a generative data/observation model. For a more textbook treatment of the subject, we refer the interested reader to Sutton and Barto (2018); Szepesvári (2022).

   While TD learning with linear function approximation is an instance of linear stochastic approximation, the analysis of TD learning with neural network approximation has been recently carried out in Tian et al. (2023); Cayci et al. (2023). Finite-time analysis of general nonlinear stochastic approximation schemes (that subsume variants of Q-learning) can be found in Wainwright (2019b); Shah and Xie (2018); Guannan Qu (2020); Chen et al. (2019, 2022); Li et al. (2024).

   Each of the papers mentioned above studies the basic versions of the concerned algorithms, where updates are made using noisy versions of some true underlying operator. Our work analyzes the robustness of these algorithms to adversarial perturbations. On a related note, we mention here that other types of perturbations resulting from communication-induced challenges (e.g., delays and compression) have been explored recently in Mitra et al. (2023); Adibi et al. (2024); Dal Fabbro et al. (2024).

2. **Reward Contamination in Multi-Armed Bandits.** A large body of work has explored the effects of reward contamination on the performance of stochastic bandit problems, both for the unstructured multi-armed bandit (MAB) setting (Jun et al., 2018; Liu and Shroff, 2019; Kapoor et al., 2019; Lykouris et al., 2018; Gupta et al., 2019), and also for structured linear bandits (Bogunovic et al., 2020; Garcelon et al., 2020; Bogunovic et al., 2021; He et al., 2022). The basic premise in these papers is that an adversary can modify the true stochastic reward/feedback on certain rounds; a corruption budget C captures the total corruption injected by the adversary over the horizon T. In particular, the authors in Kapoor et al. (2019) study a Huber-contaminated reward model like us, where in each round, with probability η (independently of the other rounds), the attacker can bias the reward seen by the learner. A fundamental lower bound of $\Omega(\eta T)$ on the regret is also established in Kapoor et al. (2019). While our reward contamination model is directly inspired by the above line of work, **we emphasize that the stochastic approximation setting we study here fundamentally differs from the bandit problem**. As such, our algorithms and proof techniques are also different from the bandit literature.

3. **Robust Statistics.** The study of computing different statistics (e.g., mean, variance, etc.) of a data set in the presence of outliers was pioneered by Huber (Huber, 1992, 2004). Since then, the field of robust statistics has significantly advanced, with more recent work focusing on computationally tractable algorithms in the high-dimensional setting (Lai et al., 2016; Chen et al., 2015; Minsker, 2018; Cheng et al., 2019; Lugosi and Mendelson, 2021; Dalalyan and Minasyan, 2022). Our paper builds on this rich line of work and uses it in the context of RL. As mentioned earlier, the standing assumption in the existing robust statistics papers is that the inliers are generated in an i.i.d. manner. We relax this assumption for robust univariate mean estimation, and show how bounds can be obtained with correlated data generated from a Markov chain.

B Useful Results

In this section, we will compile some known results and facts that will play a key role in our subsequent analysis. In what follows, unless otherwise stated, we will use $\|\cdot\|$ to refer to the standard Euclidean norm.

To proceed, we remind the reader that $\bar{A} = \Phi^\top D\left(\gamma P_\mu - I\right)\Phi$ and $\bar{b} = \Phi^\top D R_\mu$ are the steady-state versions of A_t and b_t, respectively. Recall that the mean-path/steady-state TD(0) update direction is as follows:

$$\bar{g}(\theta) = \bar{A}\theta + \bar{b}.$$

The next result from Bhandari et al. (2018) tells us that the steady-state direction $\bar{g}(\theta)$ drives the iterates towards the optimal parameter θ^*.

Lemma 2. *The following holds $\forall \theta \in \mathbb{R}^K$:*

$$\langle \theta^* - \theta, \bar{g}(\theta)\rangle \geq \omega(1-\gamma)\|\theta^* - \theta\|^2,$$

where ω is the smallest eigenvalue of the matrix $\Sigma = \Phi^\top D \Phi$.

Under the assumptions on the feature matrix Φ in Section 2, and Assumption 1, it is easy to see that Σ is positive definite with $\omega \in (0,1)$. Next, thanks to feature-normalization, it is easy to establish bounds on the norms of $\bar{A}_t$ and $\bar{A}$ (Bhandari et al., 2018; Srikant and Ying, 2019):

$$\|A_t\| \leq 2, \forall t \in \mathbb{N},\ \|\bar{A}\| \leq 2. \tag{10}$$

We will use the above fact at several points in our analysis. In addition to the above results, we will require a couple of standard concentration tools that we list here to keep the paper self-contained. For reference, see Boucheron (2013); Chung and Lu (2006).

Lemma 3. ***(Hoeffding's Inequality)*** *If $X_1, X_2, \ldots, X_N$ are independent random variables with $\mathbb{P}(a \leq X_i \leq b) = 1$ and common mean μ, then for any $\epsilon > 0$:*

$$\mathbb{P}(|\bar{X}_N - \mu| > \epsilon) \leq 2\exp\left\{\frac{-2N\epsilon^2}{(b-a)^2}\right\}, \tag{11}$$

where $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$.

Lemma 4. ***(Bernstein's Inequality)*** *If $X_1, X_2, \ldots, X_N$ are independent random variables with $\mathbb{P}(|X_i| \leq c) = 1$ and common mean μ, then for any $\epsilon > 0$:*

$$\mathbb{P}(|\bar{X}_N - \mu| > \epsilon) \leq 2\exp\left\{-\frac{N\epsilon^2}{2\sigma^2 + \frac{2c\epsilon}{3}}\right\}, \tag{12}$$

where $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\sigma^2 = \frac{1}{N}\sum_{i=1}^N Var(X_i)$.

C Proof of Theorem 1

In this section, we provide a proof for Theorem 1. First, suppose the corruption fraction $\epsilon = 0$. In this case, under the conditions in Theorem 1, it is well known (Tsitsiklis and Van Roy, 1997) that TD(0) converges with probability one to $\theta^* = -\bar{A}^{-1}\bar{b}$, where $\bar{A} = \Phi^\top D\left(\gamma P_\mu - I\right)\Phi$, and $\bar{b} = \Phi^\top D R_\mu$. Notice that the rewards only affect the term $\bar{b}$. Now under the Huber-contaminated reward model described in Section 3, for each state $s \in \mathcal{S}$, whenever in s, with probability $(1-\epsilon)$, the learner observes the true mean reward $R_\mu(s)$, and with probability ϵ, a bounded, deterministic signal $C(s)$. Thus, effectively, the mean of the observed reward random variable $\tilde{r}(s)$ in state s is $\tilde{R}(s) \triangleq (1-\epsilon)R_\mu(s) + \epsilon C(s)$. Stacking up the individual components $\tilde{R}(s)$ into a perturbed reward vector $\tilde{R}$, let us define $\tilde{b} \triangleq \Phi^\top D\tilde{R}$. Since everything else remains the same, TD(0) will now converge with probability one to

$$\begin{aligned}\tilde{\theta}^* &= -\bar{A}^{-1}\tilde{b} = -\bar{A}^{-1}\Phi^\top D\tilde{R} \\ &= \underbrace{(1-\epsilon)(-\bar{A}^{-1}\Phi^\top D R_\mu)}_{(1-\epsilon)\theta^*} + \epsilon(-\bar{A}^{-1}\Phi^\top D C),\end{aligned} \tag{13}$$

where we used the expression for θ^*. This establishes the first part of the theorem.

For the second part, suppose the adversary wishes $\tilde{\theta}^*$ to be some point $w \in \mathbb{R}^K$. We will show that this is possible by explicitly designing an appropriate corrupted reward vector C_w. In particular, let

$$C_w = \frac{1}{\epsilon} D^{-1} \Phi \left(\Phi^\top \Phi\right)^{-1} \bar{A} \left((1-\epsilon)\theta^* - w\right).$$

We note here that in light of Assumption 1, $\pi(s) > 0, \forall s \in \mathcal{S}$. Hence, D^{-1} exists. Moreover, the fact that Φ is full column rank ensures the existence of $\left(\Phi^\top \Phi\right)^{-1}$. Plugging in our choice of C_w into Eq. (13), it is easy to see that $\tilde{\theta}^* = w$. This completes the proof.

D Performance Analysis for RUMEM: Proof of Theorem 2

In this section, we will prove Theorem 2. Let us recall the setting quickly. First, $X_1, \ldots, X_N$ is a sequence of N samples drawn from a stationary Markov chain with stationary distribution π. Thus, each X_i is distributed as per π. Next, for each $i \in [N]$, with probability $(1-\epsilon)$, the learner observes a random variable $f(X_i)$ with mean $F(X_i)$ and variance at most ρ^2; and with probability ϵ, an arbitrary object chosen by the adversary. The resulting sample is denoted $\tilde{f}_i$.

Our analysis will rely on a coupling argument that is inspired by the recent work of Dorfman and Levy (2022). To apply this argument, consider the sub-sampled sequence $X_1, X_{\tau+1}, \ldots, X_{(n-1)\tau+1}$, where n and τ are as in Section 4. We couple this sequence with its i.i.d. counterpart $(X_{\mathtt{I},1}, X_{\mathtt{I},\tau+1}, \ldots, X_{\mathtt{I},(n-1)\tau+1}) \sim \pi^{\otimes n}$. Here, and henceforth throughout the proof, we will use the subscript $\mathtt{I}$ to denote the i.i.d. counterpart of a Markov sample. The next result bounds the probability of the Markovian sub-sampled sequence being different from its i.i.d. counterpart.

Lemma 5. *(Nagaraj et al., 2020, Lemma 3) Let $\mathcal{E}_1$ be an event where $(X_1, X_{\tau+1}, \ldots, X_{(n-1)\tau+1}) = (X_{I,1}, X_{I,\tau+1}, \ldots, X_{I,(n-1)\tau+1})$. Then,*

$$\mathbb{P}(\mathcal{E}_1^c) \leq (n-1) d_{mix}(\tau),$$

where $d_{mix}(\cdot)$ is as defined as

$$d_{mix}(t) \triangleq \sup_{X \in \mathcal{X}} D_{TV}\left(\mathbb{P}(X_{t+1} \in \cdot | X_1 = X), \pi\right).$$

We will call upon the above lemma at a later point in our analysis. For now, we split the proof into multiple steps.

Step 1. Bounding the number of corrupted samples. The size of the sub-sampled set used in the RUMEM algorithm is n. Our first step is to control the number of corrupted samples in this sub-sampled set. To that end, consider an event $\mathcal{E}_2$, where the maximum number of corrupted samples in the sub-sampled set is $\frac{3\epsilon' n}{2}$; recall here that

$$\epsilon' = \epsilon + \frac{32}{3n} \log\left(\frac{24}{\delta}\right).$$

Our immediate goal is to provide an upper bound on the complementary event $\mathcal{E}_2^c$. With this in mind, let W_i be an indicator random variable, such that $W_i = 1$ if the i^{th} sub-sample is corrupted, and 0 otherwise. Based on the Huber attack model, $\mathbb{E}[W_i] = \epsilon, \forall i \in [n]$. Furthermore, $\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(W_i) \leq \epsilon$. Now observe:

$$\begin{aligned}
\mathcal{E}_2^c &= \left\{\sum_{i=1}^n W_i \geq \frac{3\epsilon' n}{2}\right\} \\
&= \left\{\frac{1}{n}\sum_{i=1}^n W_i - \epsilon \geq \frac{3\epsilon'}{2} - \epsilon\right\} \\
&\implies \left\{\frac{1}{n}\sum_{i=1}^n W_i - \epsilon \geq \frac{\epsilon'}{2}\right\},
\end{aligned} \tag{14}$$

where in the last step, we used $\epsilon' > \epsilon$. Invoking Bernstein's inequality (Lemma 4) then yields

$$\mathbb{P}\left(\mathcal{E}_2^c\right) \leq 2e^{-\frac{3\epsilon' n}{32}}. \tag{15}$$

This completes step 1.

Step 2: Statistics of each non-contaminated sample. Suppose a particular sample i is non-contaminated, i.e., $\tilde{f}_i = f(X_i)$. We then have

$$\begin{aligned}
&\mathbb{E}_{X_i\sim\pi, f(X_i)\sim\mathcal{D}(\cdot|X_i)}[f(X_i)] \\
&= \sum_{x\in\mathcal{X}} \mathbb{E}_{f(X_i)\sim\mathcal{D}(\cdot|X_i)}[f(X_i)|X_i = x]\pi(x) \\
&= \sum_{x\in\mathcal{X}} \mathbb{E}_{f(x)\sim\mathcal{D}(\cdot|x)}[f(x)]\pi(x) \\
&= \sum_{x\in\mathcal{X}} F(x)\pi(x) \\
&= \bar{f}.
\end{aligned} \tag{16}$$

Now given that $|F(x)| \leq \psi, \forall x \in \mathcal{X}$, we have $|\bar{f}| \leq \sum_{x\in\mathcal{X}} |F(x)|\pi(x) \leq \psi$. Using this, we can bound the variance of a non-contaminated sample as follows:

$$\begin{aligned}
&\mathbb{E}_{X_i\sim\pi, f(X_i)\sim\mathcal{D}(\cdot|X_i)}[f^2(X_i)] \\
&= \sum_{x\in\mathcal{X}} \mathbb{E}_{f(X_i)\sim\mathcal{D}(\cdot|X_i)}[f^2(X_i)|X_i = x]\pi(x) \\
&= \sum_{x\in\mathcal{X}} \mathbb{E}_{f(x)\sim\mathcal{D}(\cdot|x)}[f^2(x)]\pi(x) \\
&\overset{(a)}{\leq} \sum_{x\in\mathcal{X}} (\bar{f}^2 + \rho^2)\pi(x) \\
&\overset{(b)}{\leq} \sum_{x\in\mathcal{X}} (\psi^2 + \rho^2)\pi(x) \\
&\leq \bar{\sigma}^2,
\end{aligned} \tag{17}$$

where $\bar{\sigma}^2 \triangleq 2(\max\{\psi, \rho\})^2$. For (a), we used the basic identity: $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$; and for (b), we used the fact that $\bar{f} \leq \psi$ and that the variance of $f(x)$ is at most ρ^2.

In light of Lemma 5, we will first focus on assessing the performance of `RUMEM` on the i.i.d. sub-sampled sequence $(X_{\mathtt{I},1}, X_{\mathtt{I},\tau+1}, \ldots, X_{\mathtt{I},(n-1)\tau+1})$. For this data set, let the mean of the i-th bucket $\mathcal{B}_i$ be denoted $\hat{\mu}_{\mathtt{I},i}$, and the final `MoM` estimate be denoted $\hat{\mu}_{\mathtt{I}} = \texttt{Median}\{\hat{\mu}_{\mathtt{I},1}, \hat{\mu}_{\mathtt{I},2}, \ldots, \hat{\mu}_{\mathtt{I},L}\}$.

Step 3. Bound on performance for each non-contaminated bucket under i.i.d. data. Consider a bucket $\mathcal{B}_i$ that only contains non-contaminated i.i.d. samples. Based on Step 2, each sample has mean $\bar{f}$ and variance at most $\bar{\sigma}^2$. Using $M = \lfloor n/L \rfloor$ to denote the number of samples in each bucket, we obtain the following bound using Chebyshev's inequality $\forall d > 0$:

$$\mathbb{P}\left(\hat{\mu}_{\mathtt{I},i} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right) \leq \frac{\frac{\bar{\sigma}^2}{M}}{\frac{d^2\bar{\sigma}^2}{\frac{N}{\tau}}} = \frac{N}{M\tau d^2}. \tag{18}$$

To proceed, we will require a lower bound on the number of samples M in each bucket. To that end, we start by noting that

$$n = \lfloor \frac{N-1}{\tau} \rfloor + 1 \geq \frac{N-1}{\tau} \geq \frac{N}{2\tau},$$

where in the last step, we used $N \geq 2$. Next, using $N \geq 4L\tau$ - as required in the statement of Theorem 2 - we obtain

$$M = \lfloor n/L \rfloor \geq \frac{n}{L} - 1 \geq \frac{N}{2L\tau} - 1 \geq \frac{N}{4L\tau}.$$

Plugging the above bound back in Eq. (18), and setting $d = 4\sqrt{L}$, we obtain

$$p \triangleq \mathbb{P}\left(\hat{\mu}_{\mathtt{I},i} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right) \leq \frac{4L}{d^2} \leq \frac{1}{4}. \tag{19}$$

Step 4. Bounding the performance of RUMEM under i.i.d. data: Similar to Step 3, consider again the scenario when the sub-sampled sequence is generated in an i.i.d. manner. Now with each non-contaminated bucket $\mathcal{B}_i$, let us associate an indicator random variable Y_i, such that $Y_i = 1$ if $\hat{\mu}_{\mathtt{I},i} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}$, and 0 otherwise. From Step 3, we know that $\mathbb{E}[Y_i] = p \leq 1/4$. To proceed, we will find it useful to define a couple of events. By $\mathcal{C}$, we define an event where $\exists\ \frac{L}{2}$ buckets $\mathcal{B}_i, i \in \{1, L\}$, such that the corresponding means of those buckets satisfy $\hat{\mu}_{\mathtt{I},i} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}$. We also define an event $\mathcal{D}$ where $\exists \left(\frac{L}{2} - \frac{3\epsilon' n}{2}\right)$ non-contaminated buckets, such that each such bucket satisfies the same property as above. Noting that on the event $\mathcal{E}_2$, at most $\frac{3\epsilon' n}{2}$ buckets can be corrupted, we then have:

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_{\mathtt{I}} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right) &\leq \mathbb{P}(\mathcal{C}) \\ &\leq \mathbb{P}\left(\{\mathcal{C} \cap \mathcal{E}_2\}\right) + \mathbb{P}\left(\mathcal{E}_2^c\right) \\ &\leq \mathbb{P}(\mathcal{D}) + \mathbb{P}\left(\mathcal{E}_2^c\right).\end{aligned} \tag{20}$$

Our next goal is to establish an upper bound on $\mathbb{P}(\mathcal{D})$. To that end, let $\mathcal{J}$ denote the set of indices corresponding to the non-contaminated buckets, and let $\tilde{N} = |\mathcal{J}|$. We then have:

$$\begin{aligned}\mathbb{P}(\mathcal{D}) &= \mathbb{P}\left(\frac{1}{\tilde{N}}\sum_{i\in\mathcal{J}} Y_i \geq \frac{\frac{L}{2} - \frac{3\epsilon' n}{2}}{\tilde{N}}\right) \\ &\overset{(a)}{\leq} \mathbb{P}\left(\frac{1}{\tilde{N}}\sum_{i\in\mathcal{J}} Y_i \geq \frac{\frac{L}{2} - \frac{3\epsilon' n}{2}}{L}\right) \\ &\leq \mathbb{P}\left(\frac{1}{\tilde{N}}\sum_{i\in\mathcal{J}} Y_i - p \geq \frac{1}{2} - \frac{3\epsilon' n}{2L} - p\right) \\ &\overset{(b)}{\leq} \mathbb{P}\left(\frac{1}{\tilde{N}}\sum_{i\in\mathcal{J}} Y_i - p \geq \frac{1}{8}\right) \\ &= \mathbb{P}(\mathcal{F}),\end{aligned} \tag{21}$$

where

$$\mathcal{F} = \left\{\frac{1}{\tilde{N}}\sum_{i\in\mathcal{J}} Y_i - p \geq \frac{1}{8}\right\}.$$

In the above steps, for (a), we used $\tilde{N} \leq L$, and for (b), we used $p \leq \frac{1}{4}$ and $L \geq 12\epsilon' n$. At this stage, one might be tempted to use a Hoeffding bound to control $\mathbb{P}(\mathcal{F})$. However, care needs to be taken here since $\tilde{N}$ is random. As such, some more work is needed before one can apply Hoeffding's inequality. We proceed by using the law of total probability to obtain:

$$\mathbb{P}(\mathcal{F}) = \mathbb{P}(\mathcal{F} \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{F} \cap \mathcal{E}_2^c) \leq \mathbb{P}(\mathcal{F} \cap \mathcal{E}_2) + \mathbb{P}(\mathcal{E}_2^c). \tag{22}$$

Considering the definition of the event $\mathcal{E}_2$, and using $L \geq 12\epsilon' n$, we have the following bound on $\tilde{N}$ on the event $\mathcal{E}_2$: $\tilde{N} \geq L - \frac{3\epsilon' n}{2} \geq \frac{7L}{8}$. This implies:

$$\mathbb{P}(\mathcal{F} \cap \mathcal{E}_2) = \sum_{j=\frac{7L}{8}}^{L} \mathbb{P}(\mathcal{F} \cap \mathcal{E}_2 \cap \{\tilde{N} = j\}). \tag{23}$$

Combining the above bound with those in equations (21) and (22), we obtain:

$$\begin{aligned}\mathbb{P}(\mathcal{D}) &\leq \sum_{j=\frac{7L}{8}}^{L} \mathbb{P}(\mathcal{F} \cap \mathcal{E}_2 \cap \{\tilde{N} = j\}) + \mathbb{P}(\mathcal{E}_2^c) \\ &\leq \sum_{j=\frac{7L}{8}}^{L} \mathbb{P}(\mathcal{F} \cap \{\tilde{N} = j\}) + \mathbb{P}(\mathcal{E}_2^c) \\ &\leq \sum_{j=\frac{7L}{8}}^{L} \mathbb{P}\left(\frac{1}{j}\sum_{i\in\mathcal{J},|\mathcal{J}|=j} Y_i - p \geq \frac{1}{8}\right) + \mathbb{P}(\mathcal{E}_2^c).\end{aligned} \tag{24}$$

We can now use Hoeffding's inequality (Lemma 3) to bound the R.H.S. of Eq. (24) since j is deterministic. This yields:

$$\begin{aligned}\mathbb{P}(\mathcal{D}) &\leq \sum_{j=\frac{7L}{8}}^{L} e^{-\frac{j}{32}} + \mathbb{P}(\mathcal{E}_2^c). \\ &\leq \frac{L}{8}e^{-\frac{7L}{256}} + \mathbb{P}(\mathcal{E}_2^c).\end{aligned} \tag{25}$$

Finally, we combine the above bound with Eq. (15) and Eq. (20) to obtain

$$\mathbb{P}\left(\hat{\mu}_{\mathtt{I}} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right) \leq \frac{L}{8}e^{-\frac{7L}{256}} + 2\mathbb{P}(\mathcal{E}_2^c) = \frac{L}{8}e^{-\frac{7L}{256}} + 4e^{-\frac{3\epsilon' n}{32}}. \tag{26}$$

This completes the analysis of `RUMEM` on i.i.d. data.

Step 5. Bounding the performance of `RUMEM` under Markov data: In this last step, we will extend our bound for `RUMEM` with i.i.d. data to the Markov setting by appealing to Lemma 5. To see how this can be done, recall that $\hat{\mu}$ is the final `MoM` estimate under Markov data, and observe

$$\begin{aligned}\mathbb{P}\left(\hat{\mu} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right) &\leq \mathbb{P}\left(\left\{\hat{\mu} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right\} \cap \mathcal{E}_1\right) + \mathbb{P}\left(\mathcal{E}_1^c\right) \\ &\leq \mathbb{P}\left(\hat{\mu}_{\mathtt{I}} \geq \bar{f} + \frac{d\bar{\sigma}}{\sqrt{\frac{N}{\tau}}}\right) + \mathbb{P}\left(\mathcal{E}_1^c\right) \\ &\overset{(a)}{\leq} \frac{L}{8}e^{-\frac{7L}{256}} + 4e^{-\frac{3\epsilon' n}{32}} + \mathbb{P}\left(\mathcal{E}_1^c\right) \\ &\overset{(b)}{\leq} \frac{L}{8}e^{-\frac{7L}{256}} + 4e^{-\frac{3\epsilon' n}{32}} + (n-1)d_{mix}(\tau).\end{aligned} \tag{27}$$

For (a), we used Eq. (26), and for (b), we invoked Lemma 5. Our goal is to now ensure that each term appearing on the R.H.S. of the above display is bounded from above by $\delta/6$. Let us start with the term $(n-1)d_{mix}(\tau)$. From Dorfman and Levy (2022), we know that for any positive integer $\ell \in \mathbb{N}$, if $\tau = \ell\tau_{mix}$, then $d_{mix}(\tau) \leq 2^{-\ell}$. Thus, picking $\tau = \lceil \log_2(6N/\delta)\,\tau_{mix} \rceil$, we obtain

$$(n-1)d_{mix}(\tau) \leq N \cdot 2^{-\log_2(6N/\delta)} \leq \frac{\delta}{6}.$$

Next, given our choice of $\epsilon' = \epsilon + \frac{32}{3n}\log(\frac{24}{\delta})$, straightforward calculations reveal that

$$4e^{-\frac{3\epsilon' n}{32}} \leq \frac{\delta}{6}.$$

Finally, given that $\frac{L}{8}e^{-\frac{7L}{256}} \leq \frac{N}{8}e^{-\frac{7L}{256}}$, it is easy to verify that by picking L to satisfy

$$L \geq \frac{256}{7}\log\left(\frac{N}{\delta}\right),$$

one can ensure that the first term in the R.H.S. of Eq. (27) is at most $\delta/6$. Combining the prior requirement $L \geq 12\epsilon' n$ on L from step 4 with the one above, it suffices to set $L = \lceil 12\epsilon' n + \frac{256}{7}\log(\frac{N}{\delta})\rceil$. We conclude that the R.H.S of Eq. (27) is at most $\delta/2$. Using a symmetric argument for the lower tail, we have that with probability at least $1-\delta$, the following is true:

$$|\hat{\mu} - \bar{f}| \leq d\bar{\sigma}\sqrt{\frac{\tau}{N}}. \tag{28}$$

Recalling $d = 4\sqrt{L}$ from Step 2, and using the expression for L, we further have that with probability at least $1-\delta$,

$$\begin{aligned}
|\hat{\mu} - \bar{f}| &\leq O\left(\bar{\sigma}\sqrt{\frac{\tau}{N}}\right)\sqrt{12\epsilon' n + \frac{256}{7}\log\left(\frac{N}{\delta}\right)} \\
&\overset{(a)}{\leq} O\left(\bar{\sigma}\sqrt{\frac{\tau}{N}}\right)\left(\sqrt{\epsilon' n} + \sqrt{\log\left(\frac{N}{\delta}\right)}\right) \\
&\overset{(b)}{\leq} O\left(\bar{\sigma}\sqrt{\frac{\tau}{N}}\right)\left(\sqrt{\epsilon n} + \sqrt{\log\left(\frac{24}{\delta}\right)} + \sqrt{\log\left(\frac{N}{\delta}\right)}\right) \\
&\overset{(c)}{\leq} O(\bar{\sigma})\left(\sqrt{\epsilon} + \sqrt{\frac{\tau}{N}\log\left(\frac{N}{\delta}\right)}\right),
\end{aligned} \tag{29}$$

where for (a) and (b), we used the fact that for positive, real scalars α, β, the following is true: $\sqrt{\alpha+\beta} \leq \sqrt{\alpha}+\sqrt{\beta}$. Finally, for (c), we used $\sqrt{n} \leq 2\sqrt{\frac{N}{\tau}}$. This completes the proof.

E Main Convergence Analysis for `Robust-TD`: Proof of Theorem 3

In this section, we will prove our main convergence result for `Robust-TD`, namely Theorem 3. The key new technical ingredient that we need in our analysis is a robust estimate of the object $\bar{b}$ that features in the mean-path TD update direction $\bar{g}(\theta) = \bar{A}\theta + \bar{b}$. This is achieved in the following lemma.

Lemma 6. *(**Adversarial Perturbation Bound**) Suppose Assumption 1 holds, and the initial distribution of s_0 is the steady-state distribution π. There exists a universal constant $c > 0$, such that if the burn-in time $\bar{T}$ satisfies the requirement in Theorem 3, then the following is true for all $t \geq \bar{T}$:*

$$\begin{aligned}
&\mathbb{E}[\|\hat{b}_t - \bar{b}\|^2] \leq c\left(\epsilon + \frac{\log^2(KT)\tau_{mix}}{t}\right)K\sigma_1^2, \\
&\text{where } \sigma_1 = \max\{1, \bar{r}, \rho\}.
\end{aligned} \tag{30}$$

Proof. The proof comprises three steps. In the first step, we use the guarantees from the `RUMEM` estimator in Theorem 2 to establish a high-probability bound on the error $\|\hat{b}_t - \bar{b}\|$. In the second step, we use the result from step 1 to argue that with high probability, the resetting operating in line 8 of Algorithm 1 gets bypassed, and $\hat{b}_t$ corresponds to the output of the robust estimation procedure in line 6 of Algorithm 1. Finally, in the last step, we establish a mean-square error bound by leveraging the thresholding operation in line 8 of `Robust-TD`. We now proceed to provide the details for each of these steps.

Step 1: A high-probability estimate on $\|\hat{b}_t - \bar{b}\|$. Let us start by fixing a component $i \in [K]$ of $\bar{b}$ and $\hat{b}_t$, and a time-step $t \geq \bar{T}$. Now consider the estimation process in line 6 of Algorithm 1: $[\hat{b}_t]_i = \texttt{RUMEM}\left(\{y_{i,k}\}_{0\leq k\leq t}; \delta = 1/(KT^2)\right)$, and $y_{i,k} = [\phi(s_k)]_i\tilde{r}_k$. We wish to relate this estimation step to the robust mean estimation set up in Section 4. To that end, let us make the following observations by considering the data set $\mathcal{S} = \{y_{i,k}\}_{0\leq k\leq t}$. First, note that since the initial distribution of s_0 is the stationary distribution π, the resulting Markov chain $\{s_0, s_1, \ldots\}$ induced by the policy μ is stationary. Thus, $s_t \sim \pi, \forall t$. Next, let us

consider the statistics of a sample $y_{i,k}$ that is not corrupted, i.e., a sample for which $\tilde{r}_k = r(s_k) \sim \mathcal{D}_\mu(\cdot|s_k)$. For such a sample, we have:

$$\begin{aligned}
&\mathbb{E}_{s_k\sim\pi, r(s_k)\sim\mathcal{D}_\mu(\cdot|s_k)}[y_{i,k}] \\
&= \sum_{s\in\mathcal{S}} \mathbb{E}_{r(s_k)\sim\mathcal{D}_\mu(\cdot|s_k)}[[\phi(s_k)]_i r(s_k)|s_k = s]\pi(s) \\
&= \sum_{s\in\mathcal{S}} [\phi(s)]_i \mathbb{E}_{r(s)\sim\mathcal{D}_\mu(\cdot|s)}[r(s)]\pi(s) \\
&= \sum_{s\in\mathcal{S}} [\phi(s)]_i R_\mu(s)\pi(s) \\
&= [\bar{b}]_i.
\end{aligned} \tag{31}$$

In other words, the mean of an uncorrupted sample corresponds exactly to the i-th component of $\bar{b}$. Proceeding as above, we have:

$$\begin{aligned}
&\mathbb{E}_{s_k\sim\pi, r(s_k)\sim\mathcal{D}_\mu(\cdot|s_k)}[y_{i,k}^2] \\
&= \sum_{s\in\mathcal{S}} \mathbb{E}_{r(s_k)\sim\mathcal{D}_\mu(\cdot|s_k)}[([\phi(s_k)]_i r(s_k))^2|s_k = s]\pi(s) \\
&= \sum_{s\in\mathcal{S}} ([\phi(s)]_i)^2 \mathbb{E}_{r(s)\sim\mathcal{D}_\mu(\cdot|s)}[r^2(s)]\pi(s) \\
&\overset{(a)}{\leq} \sum_{s\in\mathcal{S}} \mathbb{E}_{r(s)\sim\mathcal{D}_\mu(\cdot|s)}[r^2(s)]\pi(s) \\
&\overset{(b)}{\leq} \sum_{s\in\mathcal{S}} (R_\mu^2(s) + \rho^2)\pi(s) \\
&\overset{(c)}{\leq} \sum_{s\in\mathcal{S}} (\bar{r}^2 + \rho^2)\pi(s) \\
&\overset{(d)}{\leq} 2\sigma_1^2.
\end{aligned} \tag{32}$$

In the above steps, for (a), we used the fact that $\|\phi(s)\|^2 \leq 1, \forall s \in \mathcal{S}$. For (b), we used that the variance of the random variable $r(s)$ is upper-bounded by ρ^2. To arrive at (c), we used the uniform upper bound on the means of the rewards: $|R_\mu(s)| \leq \bar{r}, \forall s \in \mathcal{S}$. Finally, for (d), we used $\sigma_1 = \max\{1, \bar{r}, \rho\}$. We conclude that for each sample in the data set $\mathcal{S}$, with probability $1-\epsilon$, we observe a "clean" random variable with mean $[\bar{b}]_i$, and variance at most $2\sigma_1^2$.

Given that the `RUMEM` sub-routine is invoked in line 6 of Algorithm 1 with $\delta = 1/(KT^2)$, and number of samples $N = (t+1)$, we have from Theorem 2 that with probability at least $1 - 1/(KT^2)$,

$$\begin{aligned}
&\left|[\hat{b}_t]_i - [\bar{b}]_i\right| \leq C\sigma_1 \left(\sqrt{\epsilon} + \sqrt{\frac{\tau}{N}\log\left(\frac{N}{\delta}\right)}\right) \\
&\leq C\sigma_1 \left(\sqrt{\epsilon} + \sqrt{\frac{2\tau_{mix}\log_2(6NKT^2)\log\left(NKT^2\right)}{t}}\right) \\
&\leq C\sigma_1 \left(\sqrt{\epsilon} + 2\log(12KT^3)\sqrt{\frac{\tau_{mix}}{t}}\right),
\end{aligned} \tag{33}$$

where we used the expression for τ in Eq. (5), and the fact that $N = t+1 \leq 2T$. Union-bounding over each component $i \in [K]$, and over all time-steps $t \geq \bar{T}$, we have that with probability at least $1 - 1/T$,

$$\begin{aligned}
\left|[\hat{b}_t]_i - [\bar{b}]_i\right| &\leq C\sigma_1 \left(\sqrt{\epsilon} + 2\log(12KT^3)\sqrt{\frac{\tau_{mix}}{t}}\right), \\
&\forall i \in [K], \forall t \geq \bar{T}.
\end{aligned} \tag{34}$$

Let us call the event on which the above inequalities hold $\mathcal{E}$. It then follows that on the event $\mathcal{E}$, the following is true:

$$\|\hat{b}_t - \bar{b}\| \leq C\sqrt{K}\sigma_1 \left(\sqrt{\epsilon} + 2\log(12KT^3)\sqrt{\frac{\tau_{mix}}{t}}\right), \forall t \geq \bar{T}. \tag{35}$$

To complete step 1, we note that for us to be able to invoke Theorem 2 and arrive at the bound in Eq. (33), we need the number of samples in $\mathcal{S}$, namely $N = t+1$, to satisfy the requirement $N \geq 4L\tau$. Here, recall from the

description of RUMEM that L is the number of buckets, and τ is the sub-sampling gap. Using the expressions for τ and L in Eq. (5), along with $N = t+1 \leq 2T$ and $\delta = 1/(KT^2)$, we have that

$$4L\tau \leq 96\epsilon N + c'\tau_{mix}\log^2(KT),$$

where c' is some suitably large universal constant. So the requirement that $N \geq 4L\tau$ is met if $\epsilon \leq 1/(192)$ and $t \geq 2c'\tau_{mix}\log^2(KT)$ - the latter requirement is taken care of by the choice of the burn-in time $\bar{T}$ in Theorem 3. This concludes step 1.

Step 2. Next, we claim that on the good event $\mathcal{E}$, line 8 of Algorithm 1 will always get bypassed, and $\hat{b}_t$ will be the output of the estimation scheme in line 6 of Algorithm 1. To see this, we start by noting that

$$\bar{b} = \sum_{s\in\mathcal{S}} \phi(s)R_\mu(s)\pi(s).$$

Thus,

$$\begin{aligned}\|\bar{b}\| &\leq \sum_{s\in\mathcal{S}} \|\phi(s)\||R_\mu(s)|\pi(s) \\ &\leq \sum_{s\in\mathcal{S}} \bar{r}\pi(s) \\ &\leq \sigma_1,\end{aligned} \tag{36}$$

where for second inequality, we used $\|\phi(s)\| \leq 1, \forall s \in \mathcal{S}$, and for the third inequality, we used $|R_\mu(s)| \leq \bar{r} \leq \sigma_1$. Combining the above observation with Eq. (35), we conclude that on the event $\mathcal{E}$, the following is true:

$$\|\hat{b}_t\| \leq \underbrace{C\sqrt{K}\sigma_1\left(\sqrt{\epsilon} + 2\log(12KT^3)\sqrt{\frac{\tau_{mix}}{t}}\right)}_{G_t} + \sigma_1, \forall t \geq \bar{T}.$$

This immediately leads to the claim that line 8 of Algorithm 1 always gets bypassed on event $\mathcal{E}$.

Step 3. Bound on the expected value of $\|\hat{b}_t - \bar{b}\|^2$. Let us start by noting that if $\|\hat{b}_t\| > G_t + \sigma_1$, then as per the thresholding operation in line 8 of Algorithm 1, $\hat{b}_t$ gets reset to 0. In this case, $\|\hat{b}_t - \bar{b}\| = \|\bar{b}\| \leq \sigma_1$, where in the last step, we used Eq. (36). We conclude that thanks to the thresholding operation, the following is always true deterministically:

$$\|\hat{b}_t - \bar{b}\| \leq G_t + 2\sigma_1, \forall t \geq \bar{T}. \tag{37}$$

Furthermore, from the requirement on $\bar{T}$ in Theorem 3, we have that $\bar{T} \geq \tau_{mix}(2\log(12KT^3)^2$. This tells us that for $t \geq \bar{T}, G_t \leq 2\sqrt{K}C\sigma_1$. We then have that

$$\|\hat{b}_t - \bar{b}\| \leq G_t + 2\sigma_1 \leq 2(\sqrt{K}C+1)\sigma_1, \forall t \geq \bar{T}. \tag{38}$$

We are now in a position to bound $\mathbb{E}[\|\hat{b}_t - \bar{b}\|^2]$. Using $\mathbf{1}_{\mathcal{V}}$ as an indicator for any event $\mathcal{V}$, we have $\forall t \geq \bar{T}$:

$$\begin{aligned}\mathbb{E}[\|\hat{b}_t - \bar{b}\|^2] &= \mathbb{E}[\|\hat{b}_t - \bar{b}\|^2\mathbf{1}_{\mathcal{E}}] + \mathbb{E}[\|\hat{b}_t - \bar{b}\|^2\mathbf{1}_{\mathcal{E}^c}] \\ &\overset{(a)}{\leq} \mathbb{E}[\|\hat{b}_t - \bar{b}\|^2\mathbf{1}_{\mathcal{E}}] + \left(2(\sqrt{K}C+1)\sigma_1\right)^2\mathbb{E}[\mathbf{1}_{\mathcal{E}^c}] \\ &= \mathbb{E}[\|\hat{b}_t - \bar{b}\|^2\mathbf{1}_{\mathcal{E}}] + \left(2(\sqrt{K}C+1)\sigma_1\right)^2\mathbb{P}(\mathcal{E}^c) \\ &\overset{(b)}{\leq} \mathbb{E}[\|\hat{b}_t - \bar{b}\|^2\mathbf{1}_{\mathcal{E}}] + \frac{(\sqrt{K}C+1)^2 4\sigma_1^2}{T} \\ &\overset{(c)}{\leq} O(\sigma_1^2 K\epsilon) + O\left(\frac{\log^2(12KT^3)\sigma_1^2 K\tau_{mix}}{t}\right) + \frac{\sigma_1^2 K}{T} \\ &\leq O\left(\left(\epsilon + \frac{\log^2(KT)\tau_{mix}}{t}\right)K\sigma_1^2\right).\end{aligned} \tag{39}$$

In the above steps, for (a), we used Eq. (38) to bound $\|\hat{b}_t - \bar{b}\|$ on the event $\mathcal{E}^c$. For (b), we used $\mathbb{P}(\mathcal{E}^c) \leq 1/T$. Finally, in view of steps 1 and 2, we used Eq. (35) to bound $\|\hat{b}_t - \bar{b}\|$ on the event $\mathcal{E}$. This concludes the proof. □

Equipped with the above lemma, our next step is establishing a one-step mean-square error decomposition. Before we do so, we remind the reader here of some notation: recall that $d_t = \|\theta_t - \theta^*\|^2$, $\tilde{g}_t(\theta) = A_t\theta + \hat{b}_t$, and $\sigma = \max\{\sigma_1, \|\theta^*\|, \|\theta_0\|\}$. Given the guarantee from Lemma 6, we will also find it useful to employ the following notation:

$$B_t = c\left(\epsilon + \frac{\log^2(KT)\tau_{mix}}{t}\right)K\sigma_1^2. \tag{40}$$

We have the following result.

Lemma 7. ***(Main Recursion)*** *Suppose the conditions in the statement of Theorem 3 are met. Then, the following is true $\forall t \geq \bar{T}$:*

$$\mathbb{E}[d_{t+1}] \leq (1 - \alpha\beta + 12\alpha^2)\mathbb{E}[d_t] + O(\alpha^2\sigma^2 K) + \frac{\alpha B_t}{\beta} + \mathbb{E}[M_t], \tag{41}$$

where $\beta = \omega(1-\gamma)$, $M_t = 2\alpha\langle\theta_t - \theta^*, (A_t - \bar{A})\theta_t\rangle$, and B_t is as in Eq. (40).

Proof. From the update rule of `Robust-TD` in line 10 of Algorithm 1, we have:

$$\begin{aligned}
\|\theta_{t+1} - \theta^*\|^2 &= \|\theta_t - \theta^*\|^2 + 2\alpha\langle\theta_t - \theta^*, \tilde{g}_t(\theta_t)\rangle + \alpha^2\|\tilde{g}_t(\theta_t)\|^2 \\
&= \|\theta_t - \theta^*\|^2 + 2\alpha\langle\theta_t - \theta^*, \bar{g}(\theta_t)\rangle \\
&\quad + \alpha^2\|\tilde{g}_t(\theta_t)\|^2 + 2\alpha\langle\theta_t - \theta^*, \tilde{g}_t(\theta_t) - \bar{g}(\theta_t)\rangle \\
&= \underbrace{\|\theta_t - \theta^*\|^2 + 2\alpha\langle\theta_t - \theta^*, \bar{g}(\theta_t)\rangle}_{(*)} + \underbrace{\alpha^2\|\tilde{g}_t(\theta_t)\|^2}_{(**)} \\
&\quad + \underbrace{2\alpha\langle\theta_t - \theta^*, (A_t - \bar{A})\theta_t\rangle}_{(***)} + \underbrace{2\alpha\langle\theta_t - \theta^*, \hat{b}_t - \bar{b}\rangle}_{(****)}.
\end{aligned} \tag{42}$$

Before proceeding further, it is instructive to take a moment to interpret each of the terms in the above decomposition. The term $(*)$ captures the "steady-state" behavior of `TD`(0) and is responsible for driving the iterates toward θ^*. This term can be analyzed using Lemma 2, yielding:

$$(*) \leq (1 - 2\alpha\beta)d_t. \tag{43}$$

Each of the remaining three terms can be viewed as a perturbation/disturbance to the nominal/steady-state dynamics. As for $(**)$ and $(***)$, since they feature $\hat{b}_t$, which is processed based on contaminated rewards, these terms capture the effects of adversarial perturbations. On the other hand, the term $(***)$ depends on the "closeness" between A_t and its steady-state version $\bar{A}$, which, in turn, is dictated by how quickly the underlying Markov chain is mixing. In the remainder of the proof, we bound the terms that contain adversarial effects. We start by noting that thanks to the thresholding operation, we have that $\|\hat{b}_t\| \leq G_t + \sigma_1 \leq G_t + \sigma, \forall t \geq \bar{T}$. Furthermore, in the analysis of Lemma 6, we argued that for all $t \geq \bar{T}$, it holds that $G_t \leq 2\sqrt{K}C\sigma_1 \leq 2\sqrt{K}C\sigma$, where C is some universal constant. Thus, we have

$$\|\hat{b}_t\| \leq (2\sqrt{K}C + 1)\sigma = O(\sqrt{K}\sigma), \forall t \geq \bar{T}. \tag{44}$$

To see how the above bound helps us, observe:

$$\begin{aligned}
(**) &= \alpha^2\|A_t\theta_t + \hat{b}_t\|^2 \\
&= \alpha^2\|A_t(\theta_t - \theta^*) + A_t\theta^* + \hat{b}_t\|^2 \\
&\leq 3\alpha^2\|A_t\|^2 d_t + 3\alpha^2\|A_t\|^2\|\theta^*\|^2 + 3\alpha^2\|\hat{b}_t\|^2 \\
&\leq 12\alpha^2 d_t + 12\alpha^2\sigma^2 + 3\alpha^2\|\hat{b}_t\|^2 \\
&\leq 12\alpha^2 d_t + O(\alpha^2 K\sigma^2),
\end{aligned} \tag{45}$$

where we used $\|A_t\| \leq 2$ and $\|\theta^*\| \leq \sigma$. Now for the term $(***)$, we have

$$\begin{aligned}
(***) &\leq \alpha\beta\|\theta_t - \theta^*\|^2 + \frac{\alpha}{\beta}\|\hat{b}_t - \bar{b}\|^2 \\
&= \alpha\beta d_t + \frac{\alpha}{\beta}\|\hat{b}_t - \bar{b}\|^2.
\end{aligned} \tag{46}$$

Taking expectations on both sides of the above inequality, and using Lemma 6, we have

$$\mathbb{E}[(***)] \leq \alpha\beta\mathbb{E}[d_t] + \frac{\alpha B_t}{\beta}.$$

Taking expectations on both sides of Eq. (42), and then combining the bounds on $(*), (**)$, and $(***)$ from Eq. (43), (45), and the above display, respectively, leads to the claim of the lemma. □

From Lemma 7, it is clear that in order to proceed further, we need to bound the term $\mathbb{E}[M_t]$ that corresponds to the bias introduced by Markov noise. To that end, we will require an intermediate result. Before stating this result, we remind the reader that $\tau' = \tau'_{mix}(\alpha)$ is as defined in Section 6.

Lemma 8. ***(Bounding the Drift)*** *Suppose the conditions in the statement of Theorem 3 are met. Then, the following bound holds $\forall t \geq \bar{T} + \tau'$:*

$$\|\theta_t - \theta_{t-\tau'}\|^2 \leq O(\alpha^2\tau'^2)d_t + O(\alpha^2\tau'^2 K\sigma^2). \tag{47}$$

Proof. From Eq. (44), recall that $\|\hat{b}_t\| \leq O(\sqrt{K}\sigma), \forall t \geq \bar{T}$. Using this, we obtain

$$\begin{aligned}\|\theta_{t+1}\| &\leq \|\theta_t\| + \alpha\|\tilde{g}_t(\theta_t)\| \\ &\leq \|\theta_t\| + \alpha\left(\|A_t\|\|\theta_t\| + \|\hat{b}_t\|\right) \\ &\leq (1+2\alpha)\|\theta_t\| + O(\alpha\sqrt{K}\sigma).\end{aligned} \tag{48}$$

Rolling out the above recursion, we obtain the following for any $k \in [t-\tau', t]$:

$$\|\theta_k\| \leq (1+2\alpha)^{\tau'}\|\theta_{t-\tau'}\| + O(\alpha\sqrt{K}\sigma)\sum_{\ell=0}^{\tau'}(1+2\alpha)^{\ell}.$$

Since $(1+x) \leq \exp(x), \forall x \in \mathbb{R}$, note that $(1+2\alpha)^{\tau'} \leq \exp(0.25) < 2$, for $\alpha \leq 1/(8\tau')$. Using this to simplify the above display, we have that for any $k \in [t-\tau', t]$, the following is true:

$$\|\theta_k\| \leq 2\|\theta_{t-\tau'}\| + O(\alpha\tau'\sqrt{K}\sigma) \leq 2\|\theta_{t-\tau'}\| + O(\sqrt{K}\sigma), \tag{49}$$

where in the last step, we used $\alpha\tau' \leq 1$. Let us now observe the following chain of inequalities:

$$\begin{aligned}\|\theta_t - \theta_{t-\tau'}\| &\leq \sum_{k=t-\tau'}^{t-1}\|\theta_{k+1} - \theta_k\| \\ &\leq \alpha\sum_{k=t-\tau'}^{t-1}\|\tilde{g}_k(\theta_k)\| \\ &\leq \alpha\sum_{k=t-\tau'}^{t-1}\left(\|A_k\|\|\theta_k\| + \|\hat{b}_k\|\right) \\ &\overset{(a)}{\leq} \alpha\sum_{k=t-\tau'}^{t-1}\left(2\|\theta_k\| + O(\sqrt{K}\sigma)\right) \\ &\overset{(b)}{\leq} \alpha\sum_{k=t-\tau'}^{t-1}\left(4\|\theta_{t-\tau'}\| + O(\sqrt{K}\sigma)\right) \\ &\leq 4\alpha\tau'\|\theta_{t-\tau'}\| + O(\alpha\tau'\sqrt{K}\sigma) \\ &\leq 4\alpha\tau'\left(\|\theta_t - \theta_{t-\tau'}\| + \|\theta_t\|\right) + O(\alpha\tau'\sqrt{K}\sigma) \\ &\overset{(c)}{\leq} \frac{1}{2}\|\theta_t - \theta_{t-\tau'}\| + 4\alpha\tau'\|\theta_t\| + O(\alpha\tau'\sqrt{K}\sigma).\end{aligned} \tag{50}$$

In the above steps, for (a), we used $\|A_k\| \leq 2$ and $\|\hat{b}_k\| \leq O(\sqrt{K}\sigma)$; for (b), we used Eq. (49); and for (c), we used $\alpha\tau' \leq 1/8$. Rearranging Eq. (50) and simplifying, we obtain:

$$\|\theta_t - \theta_{t-\tau'}\| \leq 8\alpha\tau'\|\theta_t\| + O(\alpha\tau'\sqrt{K}\sigma) \leq 8\alpha\tau'\|\theta_t - \theta^*\| + O(\alpha\tau'\sqrt{K}\sigma),$$

where we used $\|\theta^*\| \leq \sigma$. Squaring both sides of the above display leads to the claim of the lemma. □

With the above lemma in hand, we can now bound the term $\mathbb{E}[M_t]$.

Lemma 9. ***(Markovian Bias Bound)*** *Let M_t be as defined in Lemma 7. Suppose the conditions in the statement of Theorem 3 are met. Then, the following bound holds $\forall t \geq \bar{T} + \tau'$:*

$$\mathbb{E}[M_t] \leq O(\alpha^2\tau')\mathbb{E}[d_t] + O(\alpha^2\tau')K\sigma^2.$$

Proof. Let us start by splitting the term $\langle \theta_t - \theta^*, (A_t - \bar{A})\theta_t \rangle = T_1 + T_2 + T_3 + T_4$ into the four parts shown below:

$$\begin{aligned} T_1 &= \langle \theta_t - \theta_{t-\tau'}, (A_t - \bar{A})\theta_t \rangle \\ T_2 &= \langle \theta_{t-\tau'} - \theta^*, (A_t - \bar{A})\theta_{t-\tau'} \rangle \\ T_3 &= \langle \theta_{t-\tau'} - \theta^*, A_t(\theta_t - \theta_{t-\tau'}) \rangle \\ T_4 &= \langle \theta_{t-\tau'} - \theta^*, \bar{A}(\theta_{t-\tau'} - \theta_t) \rangle. \end{aligned} \tag{51}$$

In what follows, we proceed to bound each of the four terms above.

Bounding T_1. We bound T_1 as follows:

$$\begin{aligned} T_1 &\leq \frac{1}{2\alpha\tau'}\|\theta_t - \theta_{t-\tau'}\|^2 + \frac{\alpha\tau'}{2}\|(A_t - \bar{A})\theta_t\|^2 \\ &\leq \frac{1}{2\alpha\tau'}\|\theta_t - \theta_{t-\tau'}\|^2 + \alpha\tau'\left(\|A_t\|^2 + \|\bar{A}\|^2\right)\|\theta_t\|^2 \\ &\overset{(a)}{\leq} \frac{1}{2\alpha\tau'}\|\theta_t - \theta_{t-\tau'}\|^2 + 8\alpha\tau'\|\theta_t\|^2 \\ &\leq \frac{1}{2\alpha\tau'}\|\theta_t - \theta_{t-\tau'}\|^2 + 16\alpha\tau'\|\theta_t - \theta^*\|^2 + 16\alpha\tau'\|\theta^*\|^2 \\ &\overset{(b)}{\leq} \frac{1}{2\alpha\tau'}\|\theta_t - \theta_{t-\tau'}\|^2 + 16\alpha\tau' d_t + 16\alpha\tau'\sigma^2 \\ &\overset{(c)}{\leq} O(\alpha\tau')d_t + O(\alpha\tau')K\sigma^2. \end{aligned} \tag{52}$$

Here, for (a), we used $\max\{\|A_t\|, \|\bar{A}\|\} \leq 2$; for (b), we used $\|\theta^*\| \leq \sigma$; and for (c), we used Lemma 8.

Bounding T_3. To bound T_3, we proceed as follows:

$$\begin{aligned} T_3 &\leq \|\theta_{t-\tau'} - \theta^*\|\|A_t\|\|\theta_t - \theta_{t-\tau'}\| \\ &\leq 2\|\theta_{t-\tau'} - \theta^*\|\|\theta_t - \theta_{t-\tau'}\| \\ &\leq \alpha\tau'\|\theta_{t-\tau'} - \theta^*\|^2 + \frac{1}{\alpha\tau'}\|\theta_t - \theta_{t-\tau'}\|^2 \\ &\leq 2\alpha\tau' d_t + \left(2\alpha\tau' + \frac{1}{\alpha\tau'}\right)\|\theta_t - \theta_{t-\tau'}\|^2 \\ &\leq O(\alpha\tau')d_t + O(\alpha\tau')K\sigma^2, \end{aligned} \tag{53}$$

where in the last step, we used Lemma 8 and $\alpha\tau' \leq 1$. The term T_4 can be controlled in exactly the same way as above, with the same resulting bound.

Bounding T_2. To bound T_2, we will invoke mixing properties of the underlying Markov chain as follows:

$$\begin{aligned}
\mathbb{E}\left[T_2\right] &= \mathbb{E}\left[\langle \theta_{t-\tau'} - \theta^*, (A_t - \bar{A})\theta_{t-\tau'}\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\langle \theta_{t-\tau'} - \theta^*, (A_t - \bar{A})\theta_{t-\tau'}\rangle | \theta_{t-\tau'}, X_{t-\tau'}\right]\right] \\
&= \mathbb{E}\left[\langle \theta_{t-\tau'} - \theta^*, \mathbb{E}\left[(A_t - \bar{A})\theta_{t-\tau'} | \theta_{t-\tau'}, X_{t-\tau'}\right]\rangle\right] \\
&= \mathbb{E}\left[\langle \theta_{t-\tau'} - \theta^*, \left(\mathbb{E}\left[A_t | X_{t-\tau'}\right] - \bar{A}\right)\theta_{t-\tau'}\rangle\right] \\
&\leq \mathbb{E}\left[\|\theta_{t-\tau'} - \theta^*\| \| \left(\mathbb{E}\left[A_t | X_{t-\tau'}\right] - \bar{A}\right)\theta_{t-\tau'}\|\right] \\
&\leq \mathbb{E}\left[\|\theta_{t-\tau'} - \theta^*\| \|\mathbb{E}\left[A_t | X_{t-\tau'}\right] - \bar{A}\| \|\theta_{t-\tau'}\|\right] \\
&\overset{(a)}{\leq} \alpha \mathbb{E}\left[\|\theta_{t-\tau'} - \theta^*\| \|\theta_{t-\tau'}\|\right] \\
&\leq \alpha \mathbb{E}\left[\|\theta_{t-\tau'} - \theta^*\| \left(\|\theta^*\| + \|\theta_{t-\tau'} - \theta^*\|\right)\right] \\
&\leq \alpha \mathbb{E}\left[\|\theta_{t-\tau'} - \theta^*\| \left(\sigma + \|\theta_{t-\tau'} - \theta^*\|\right)\right] \\
&\leq O(\alpha) \mathbb{E}\left[\|\theta_{t-\tau'} - \theta^*\|^2 + \sigma^2\right] \\
&\leq O(\alpha) \mathbb{E}\left[\|\theta_t - \theta_{t-\tau'}\|^2 + d_t + \sigma^2\right] \\
&\overset{(b)}{\leq} O(\alpha) d_t + O(\alpha K \sigma^2).
\end{aligned}$$

In the above steps, (a) follows from the definition of the mixing time $\tau' = \tau'_{mix}(\alpha)$ in Definition 1, and (b) follows from Lemma 8 and $\alpha\tau' \leq 1$. Combining the bounds on $T_1 - T_4$ leads to the claim of the lemma. □

We now have all the ingredients needed to complete the proof of Theorem 3.

Proof. **Proof of Theorem 3.** Combining the bound on $\mathbb{E}[M_t]$ from Lemma 9 with the one-step recursion from Lemma 7, we obtain $\forall t \geq \bar{T} + \tau'$:

$$\begin{aligned}
\mathbb{E}[d_{t+1}] &\leq (1 - \alpha\beta + 12\alpha^2)\mathbb{E}[d_t] + O(\alpha^2 K \sigma^2) + \frac{\alpha B_t}{\beta} + \mathbb{E}[M_t] \\
&\leq (1 - \alpha\beta + C_1 \alpha^2 \tau')\mathbb{E}[d_t] + O(\alpha^2 \tau' K \sigma^2) + \frac{\alpha B_t}{\beta} \\
&\leq \left(1 - \frac{\alpha\beta}{2}\right)\mathbb{E}[d_t] + O(\alpha^2 \tau' K \sigma^2) + \frac{\alpha B_t}{\beta},
\end{aligned} \tag{54}$$

where in the second inequality, C_1 is some universal constant, and the last inequality results from picking α to satisfy

$$\alpha \leq \frac{\beta}{2C_1\tau'}.$$

Recalling that

$$B_t = c\left(\epsilon + \frac{\log^2(KT)\tau_{mix}}{t}\right) K\sigma_1^2,$$

and unrolling the inequality in Eq. (54) starting from $t = \bar{T} + \tau'$ yields:

$$\begin{aligned}
\mathbb{E}[d_T] &\leq \left(1 - \frac{\alpha\beta}{2}\right)^{T-\bar{T}-\tau'} \mathbb{E}[d_{\bar{T}+\tau'}] + O(\alpha\tau_{mix}K\sigma^2)\frac{\log^2(KT)}{\beta} \sum_{k=\bar{T}+\tau'}^{T-1} \left(1 - \frac{\alpha\beta}{2}\right)^{T-1-k} \frac{1}{k} \\
&\quad + O(\alpha^2\tau' K\sigma^2) \sum_{k=0}^{\infty} \left(1 - \frac{\alpha\beta}{2}\right)^k + O\left(\frac{\alpha K \epsilon \sigma_1^2}{\beta}\right) \sum_{k=0}^{\infty} \left(1 - \frac{\alpha\beta}{2}\right)^k \\
&\leq \left(1 - \frac{\alpha\beta}{2}\right)^{T-\bar{T}-\tau'} \mathbb{E}[d_{\bar{T}+\tau'}] + O\left(\frac{\alpha\tau_{mix}K\log^2(KT)\sigma^2}{\beta}\right) \sum_{k=1}^{T-1} \frac{1}{k} + O\left(\frac{\alpha\tau' K\sigma^2}{\beta}\right) + O\left(\frac{K\epsilon\sigma_1^2}{\beta^2}\right) \\
&\leq \left(1 - \frac{\alpha\beta}{2}\right)^{T-\bar{T}-\tau'} \mathbb{E}[d_{\bar{T}+\tau'}] + O\left(\frac{\alpha\bar{\tau}_{mix}K\log^3(KT)\sigma^2}{\beta}\right) + O\left(\frac{K\epsilon\sigma_1^2}{\beta^2}\right),
\end{aligned} \tag{55}$$

where recall that $\bar{\tau}_{mix} = \max\{\tau', \tau_{mix}\}$. To arrive at the last step, we used $\sum_{k=1}^{T-1}(1/k) = O(\log(T))$. Now suppose T is chosen sufficiently large such that $\bar{T} \leq T/4$ and $\tau' \leq T/4$. Then, we have:

$$\mathbb{E}[d_T] \leq \underbrace{\left(1-\frac{\alpha\beta}{2}\right)^{T/2}\mathbb{E}[d_{\bar{T}+\tau'}]}_{(*)} + \underbrace{O\left(\frac{\alpha\bar{\tau}_{mix}K\log^3(KT)\sigma^2}{\beta}\right)}_{(**)} + O\left(\frac{K\epsilon\sigma_1^2}{\beta^2}\right). \tag{56}$$

Now let us substitute $\alpha = \frac{4}{\beta}\frac{\log(T)}{T}$ in the above bound. With this choice of α, we have

$$(**) \leq \tilde{O}\left(\frac{K\bar{\tau}_{mix}\sigma^2}{\beta^2 T}\right).$$

We claim that $(*) \leq O(\sigma^2 K/T)$. To see why, start by noting that based on our choice of α:

$$\left(1-\frac{\alpha\beta}{2}\right)^{T/2} \leq \exp\left(-\frac{\alpha\beta T}{4}\right) = \exp(-\log(T)) = \frac{1}{T}.$$

To establish the claim, it remains to argue that $\mathbb{E}[d_{\bar{T}+\tau'}] \leq O(K\sigma^2)$. To that end, using the same reasoning as we did to arrive at Eq. (49), we have

$$\begin{aligned}\|\theta_{\bar{T}+\tau'}\| &\leq 2\|\theta_{\bar{T}}\| + O(\sqrt{K}\sigma) = 2\|\theta_0\| + O(\sqrt{K}\sigma)\\ &= O(\sqrt{K}\sigma).\end{aligned} \tag{57}$$

Here, we used that for $t \leq \bar{T}, \theta_t = \theta_0$, and $\|\theta_0\| \leq \sigma$. Thus, $d_{\bar{T}+\tau'} = \|\theta_{\bar{T}+\tau'} - \theta^*\|^2 \leq 2\|\theta_{\bar{T}+\tau'}\|^2 + 2\|\theta^*\|^2 \leq O(K\sigma^2)$. Combining all the pieces together, we have

$$\begin{aligned}&\mathbb{E}[d_T] \leq \tilde{O}\left(\frac{\bar{\tau}_{mix}\sigma^2 G}{T}\right) + O\left(\epsilon\sigma_1^2 G\right),\\ &\text{where } G = \frac{K}{\omega^2(1-\gamma)^2}.\end{aligned} \tag{58}$$

To complete the proof, it remains to specify the parameters $\bar{T}$ and T. As for the burn-in time $\bar{T}$, we note from the analysis of Lemma 6 that the following choice of $\bar{T}$ suffices:

$$\bar{T} = \lceil c_1 \tau_{mix}\log^2(KT)\rceil.$$

Next, all the requirements on the step-size α needed to arrive at our final bound can be subsumed into the following requirement:

$$\alpha \leq \frac{\omega(1-\gamma)}{C'\tau'},$$

where $C' \geq 8$ is some suitably large universal constant. Now, since we have fixed α to be $\frac{4}{\beta}\frac{\log(T)}{T}$, the above criterion can be met, provided the number of iterations T satisfies:

$$T \geq \frac{4C'\tau'\log(T)}{\omega^2(1-\gamma)^2}.$$

The above requirement on T, combined with the fact that we need $T \geq \bar{T} + \tau'$, justifies the choice of T in the statement of Theorem 3. □

F Proof of Lower Bound in Theorem 4

In this section, we will establish the lower bound in Theorem 4. Let us start by explaining the high-level idea behind our proof, and then we will supply all the technical details.

The main idea is to construct two different Markov Reward Processes (MRPs) induced by the same policy, such that (i) the value functions induced by the policy differ in magnitude by $\Omega(\sqrt{\epsilon})$ in the two MRPs; and (ii) the

distribution of rewards under the Huber-contaminated observation model is identical across the two MRPs. It is then not too hard to argue that any estimator for a value function must suffer an error of $\Omega(\sqrt{\epsilon})$ on at least one of the two MRPs. We now proceed to construct our hard instance.

Step 1: Construction of the MRPs. Consider a MDP with just one state s and one action a. Trivially, a policy μ thus maps s to a, and there is no randomness in terms of state transitions. We will now construct two MRPs, MRP 1 and MRP 2, induced by the policy μ, that differ in terms of their noisy reward models. For MRP 1, the reward random variable $r_1(s)$ has support comprising two values:

$$r_1(s) = \begin{cases} \frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{\epsilon}{4(1-\epsilon)} \\ 0 & \text{with probability } 1 - \frac{\epsilon}{4(1-\epsilon)}, \end{cases} \tag{59}$$

where $\rho > 0$ is some positive constant. We call this reward distribution $\mathcal{D}_1$. For MRP 2, the reward random variable $r_2(s)$ has distribution $\mathcal{D}_2$ defined similarly as follows:

$$r_2(s) = \begin{cases} -\frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{\epsilon}{4(1-\epsilon)} \\ 0 & \text{with probability } 1 - \frac{\epsilon}{4(1-\epsilon)}. \end{cases} \tag{60}$$

Let the mean of the rewards under $\mathcal{D}_1$ and $\mathcal{D}_2$ be denoted by R_1 and R_2, respectively. It is then easy to see that

$$R_1 = \frac{\rho\sqrt{\epsilon}}{4(1-\epsilon)}, \text{ and } R_2 = -\frac{\rho\sqrt{\epsilon}}{4(1-\epsilon)}.$$

Furthermore, the variance of both $r_1(s)$ and $r_2(s)$ is given by

$$Var(r_1(s)) = Var(r_2(s)) \leq \frac{\rho^2}{\epsilon} \times \frac{\epsilon}{4(1-\epsilon)} < 0.5\rho^2,$$

where we used the fact that the corruption fraction ϵ satisfies $\epsilon < 0.5$. Thus, each reward model has a finite variance bounded above by ρ^2. It is easily seen that the value functions in the two MRPs, say V_1 and V_2, satisfy:[4]

$$V_i = \frac{R_i}{(1-\gamma)}, i \in \{1, 2\}. \tag{61}$$

Step 2: Construction of the Attack Distributions. Consider an error distribution $\mathcal{Q}_1$ associated with MRP 1 such that a random variable $Z_1 \sim \mathcal{Q}_1$ is given by

$$Z_1 = \begin{cases} -\frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{1}{4}, \\ \frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{1}{4}. \end{cases} \tag{62}$$

Now consider a random variable X drawn from the Huber-contaminated mixture model $(1-\epsilon)\mathcal{D}_1 + \epsilon\mathcal{Q}_1$. Given the distributions of $\mathcal{D}_1$ and $\mathcal{Q}_1$, one can verify (with straightforward calculations) that the distribution of X is as follows:

$$X = \begin{cases} -\frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{\epsilon}{2} \\ 0 & \text{with probability } 1-\epsilon, \\ \frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{\epsilon}{2}. \end{cases} \tag{63}$$

For MRP 2, we construct an error distribution $\mathcal{Q}_2$ such that a random variable Z_2 drawn from $\mathcal{Q}_2$ is as follows:

$$Z_2 = \begin{cases} -\frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{1}{4} \\ 0 & \text{with probability } \frac{1}{4}, \\ \frac{\rho}{\sqrt{\epsilon}} & \text{with probability } \frac{1}{2}. \end{cases} \tag{64}$$

[4] We drop the dependence of V_i on s since there is only one state.

Now consider a random variable Y drawn as per $(1-\epsilon)\mathcal{D}_2 + \epsilon\mathcal{Q}_2$. The specific nature of our construction ensures that

$$(1-\epsilon)\mathcal{D}_1 + \epsilon\mathcal{Q}_1 = (1-\epsilon)\mathcal{D}_2 + \epsilon\mathcal{Q}_2.$$

In summary, the contaminated reward random variable X in MRP 1 has the same distribution as the contaminated reward random variable Y in MRP 2. As such, these two reward models are indistinguishable to a learner. However, we also have:

$$|V_1 - V_2| = \frac{\rho\sqrt{\epsilon}}{2(1-\epsilon)(1-\gamma)} \geq \frac{\rho\sqrt{\epsilon}}{2(1-\gamma)}. \tag{65}$$

In light of the above facts, we now proceed to argue that any value-function estimator must suffer $\Omega\left(\frac{\rho\sqrt{\epsilon}}{(1-\gamma)}\right)$ error in at least one of the MRPs.

Step 3. Lower-bounding Error of any Estimator. For $i = 1, \ldots, T$, let (X_i, Y_i) be independent pairs of random observations satisfying:

$$\mathbb{P}(X_i = Y_i = -\rho/\sqrt{\epsilon}) = \frac{\epsilon}{2},$$
$$\mathbb{P}(X_i = Y_i = 0) = 1-\epsilon\,,\, \mathbb{P}(X_i = Y_i = \rho/\sqrt{\epsilon}) = \frac{\epsilon}{2}.$$

Let us note that X_i is distributed as per $(1-\epsilon)\mathcal{D}_1 + \epsilon\mathcal{Q}_1$, and Y_i as per $(1-\epsilon)\mathcal{D}_2 + \epsilon\mathcal{Q}_2$. Clearly, the following is true: $\mathbb{P}\left(\{X_i\}_{i\in[T]} = \{Y_i\}_{i\in[T]}\right) = 1$. Now suppose $\hat{R}_T$ is any estimator for estimating the means of the rewards in the two MRPs. As we shall see, establishing a fundamental limit on the performance of $\hat{R}_T$ is sufficient to establish a limit on the performance of any value-function estimator. In what follows, for conciseness of notation, let

$$B \triangleq \frac{\rho\sqrt{\epsilon}}{4(1-\epsilon)}.$$

We then have

$$\begin{aligned}
&\max\left\{\mathbb{P}\left(|\hat{R}_T(\{X_i\}_{i\in[T]}) - R_1| > \frac{B}{2}\right), \mathbb{P}\left(|\hat{R}_T(\{Y_i\}_{i\in[T]}) - R_2| > \frac{B}{2}\right)\right\} \\
&\geq \frac{1}{2}\mathbb{P}\left(\left\{|\hat{R}_T(\{X_i\}_{i\in[T]}) - R_1| > \frac{B}{2}\right\} \bigcup \left\{|\hat{R}_T(\{Y_i\}_{i\in[T]}) - R_2| > \frac{B}{2}\right\}\right) \\
&\geq \frac{1}{2}\mathbb{P}\left(\hat{R}_T(\{X_i\}_{i\in[T]}) = \hat{R}_T(\{Y_i\}_{i\in[T]})\right) \\
&\geq \frac{1}{2}\mathbb{P}\left(\{X_i\}_{i\in[T]} = \{Y_i\}_{i\in[T]}\right) \\
&= \frac{1}{2},
\end{aligned} \tag{66}$$

where for the second inequality, we used $R_1 = B$ and $R_2 = -B$. Using $1/(1-\epsilon) > 1$, we then conclude that:

$$\max\left\{\mathbb{P}\left(\left|\hat{R}_T(\{X_i\}_{i\in[T]}) - R_1\right| > \frac{\rho\sqrt{\epsilon}}{8}\right),\ \mathbb{P}\left(\left|\hat{R}_T(\{Y_i\}_{i\in[T]}) - R_2\right| > \frac{\rho\sqrt{\epsilon}}{8}\right)\right\} \geq \frac{1}{2}. \tag{67}$$

Let $\hat{V}_T$ be any estimator for the value functions in the two MRPs. In light of Eq. (67), we claim that

$$\max\left\{\mathbb{P}\left(|\hat{V}_T(\{X_i\}_{i\in[T]}) - V_1| > \frac{\rho\sqrt{\epsilon}}{8(1-\gamma)}\right),\ \mathbb{P}\left(|\hat{V}_T(\{Y_i\}_{i\in[T]}) - V_2| > \frac{\rho\sqrt{\epsilon}}{8(1-\gamma)}\right)\right\} \geq \frac{1}{2}. \tag{68}$$

The claim essentially follows from the simple observation that if a value-function estimator $\hat{V}_T$ can accurately estimate both V_1 and V_2, then one can use such an estimator to construct accurate estimates of both R_1 and R_2, thereby violating Eq. (67). Formally, to see that Eq. (67) implies Eq. (68), suppose there exists an estimator $\hat{V}_T$ such that

$$\max\left\{\mathbb{P}\left(\left|\hat{V}_T(\{X_i\}_{i\in[T]}) - V_1\right| > \frac{\rho\sqrt{\epsilon}}{8(1-\gamma)}\right),\ \mathbb{P}\left(\left|\hat{V}_T(\{Y_i\}_{i\in[T]}) - V_2\right| > \frac{\rho\sqrt{\epsilon}}{8(1-\gamma)}\right)\right\} < \frac{1}{2}. \tag{69}$$

Using $\hat{V}_T$, construct a reward estimator $\hat{R}_T = (1-\gamma)\hat{V}_T$. From Eq. (61), we then immediately have:

$$\max\left\{\mathbb{P}\left(\left|\hat{R}_T(\{X_i\}_{i\in[T]}) - R_1\right| > \frac{\rho\sqrt{\epsilon}}{8}\right), \mathbb{P}\left(\left|\hat{R}_T(\{Y_i\}_{i\in[T]}) - R_2\right| > \frac{\rho\sqrt{\epsilon}}{8}\right)\right\} < \frac{1}{2}. \tag{70}$$

This completes the claim and the proof.

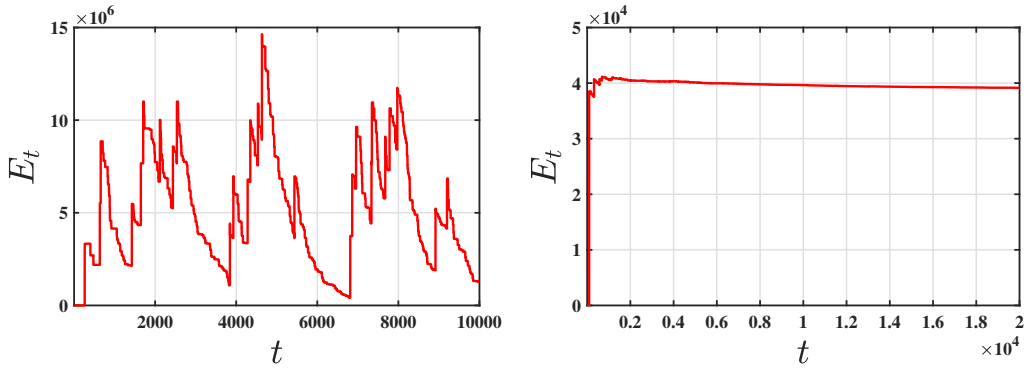

Figure 2: Plots of the mean-square error $E_t = \|\theta_T - \theta^*\|_2^2$ for TD(0) under the Huber-contaminated reward model with corruption probability $\epsilon = 0.001$, and a simple biasing attack where the attack signal is $100/\epsilon$. (**Left**) Constant step-size $\alpha = 0.1$. (**Right**) Diminishing step-size $\alpha_t = 1/t$.

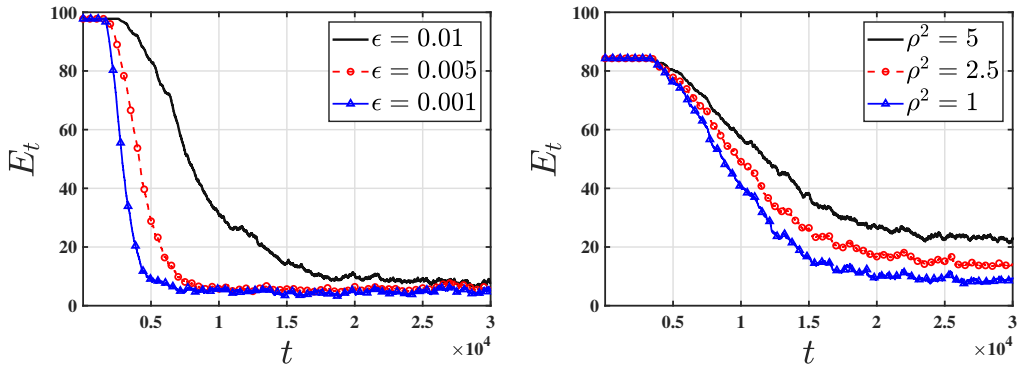

Figure 3: Plots of the mean-square error $E_t = \|\theta_T - \theta^*\|_2^2$ for `Robust-TD` under the Huber-contaminated reward model with a biasing attack, where the attack signal is $100/\epsilon$. (**Left**) Effect of varying the corruption probability ϵ. (**Right**) Effect of varying the noise variance ρ^2.

G Simulation Study

In this section, we report some synthetic experiments to support the theory developed in this paper. All simulations are performed on an HP Spectre x360 personal laptop with 11th Gen Intel(R) 4-Core Processor.

Basic Setup. We construct an MDP with 100 states, and use a feature matrix Φ with $K = 10$ independent basis vectors. Using this MDP, we generate the state transition matrix P_μ and reward vector R_μ associated with a fixed policy μ. For all our simulations, the discount factor γ is set to 0.5, and the rewards are generated uniformly at random from the interval $(0, 5)$. Unless specified, the step size α is chosen to be 0.1. We perform 10 independent trials per simulation and average the errors from each trial to report the mean-square error $E_t = \|\theta_T - \theta^*\|_2^2$. With this basic setup, we now report various experiments below.

1. **Vulnerability of TD(0).** The purpose of the first simulation is to reveal the vulnerability of the basic TD(0) algorithm to adversarial reward contamination. To that end, we consider a scenario where the corruption fraction is $\epsilon = 0.001$, and the rewards are noiseless. In each state s, with probability ϵ, the adversary injects a biasing signal of magnitude $100/\epsilon$. The outcome of this experiment is reported in Fig. 2. When the step size is held constant at $\alpha = 0.1$, convergence is to a ball centered around a perturbed parameter $\tilde{\theta}^*$, where $\tilde{\theta}^*$ is as in Theorem 1. The size of this ball scales with the effective reward magnitude that depends on the bias $100/\epsilon$. Hence, in the left display of Fig. 2, we see large oscillations that depend on the bias magnitude. With a diminishing step-size of the form $\alpha_t = 1/t$, Theorem 1 suggests exact convergence to the perturbed point $\tilde{\theta}^*$. This is reflected in the right display of Fig. 2, where the mean-square error (MSE) converges to a

steady-state value that is bounded far away from 0.

2. **Performance of** `Robust-TD`. In our next simulation, we assess the performance of `Robust-TD`. For our first experiment, the noise model comprises a zero-mean Gaussian distribution with a variance of 1. We consider the same biasing attack as before, where the attacker injects a constant bias of $100/\epsilon$. We vary the corruption probability ϵ, and report our findings in the left display of Fig. 3. For each value of ϵ, the MSE converges to a small ball around 0. In the next experiment, we fix the corruption probability to 0.01 and vary the noise variance level. We consider three different values of variance: 5, 2.5, and 1. As expected, with a constant step size, the MSE settles down to a ball around the origin, where the size of the ball depends on the noise variance. Notably, complying with Theorem 3, the MSE of `Robust-TD` is unaffected by the magnitude of the adversarial bias input.