FINITE-TIME BOUNDS FOR DISTRIBUTIONALLY ROBUST TD LEARNING WITH LINEAR FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributionally robust reinforcement learning (DRRL) focuses on designing policies that achieve good performance under model uncertainties. In particular, we are interested in maximizing the worst-case long-term discounted reward, where the data for RL comes from a nominal model while the deployed environment can deviate from the nominal model within a prescribed uncertainty set. Existing convergence guarantees for robust temporal-difference (TD) learning for policy evaluation are limited to tabular MDPs or are dependent on restrictive discount-factor assumptions when function approximation is used. We present the first robust TD learning with linear function approximation, where robustness is measured with respect to the total-variation distance uncertainty set. Additionally, our algorithm is both model-free and does not require generative access to the MDP. Our algorithm combines a two-time-scale stochastic-approximation update with an outer-loop target-network update. We establish an $\tilde{O}(1/\epsilon^2)$ sample complexity to obtain an ϵ-accurate value estimate. Our results close a key gap between the empirical success of robust RL algorithms and the non-asymptotic guarantees enjoyed by their non-robust counterparts. The key ideas in the paper also extend in a relatively straightforward fashion to robust Q-learning with function approximation.

1 INTRODUCTION

Reinforcement learning (RL) aims to learn policies that maximize long-term reward. Standard RL methods learn the optimal strategy from trajectories generated by a simulator or the real environment, implicitly assuming that training and deployment environments share the same dynamics. Many applications face two issues: simulation–reality gaps and distribution shift between training and deployment. These call for policies that are robust to perturbations in the environment. Distributionally robust RL (DRRL) tackles this by assuming the true environment lies in an uncertainty set around a nominal model. It then learns a policy that maximizes the worst-case cumulative reward over that set, using data from trajectories corresponding to the nominal model. In this work, we focus on model-free DRRL with linear function approximation for the value function to deal with large state spaces.

In contrast to our model-free approach, model-based DRRL often proceeds by fitting an empirical transition model, defining an uncertainty set from it, and then optimizing for a robust policy (Shi & Chi, 2024; Wang & Zou, 2021; Xu et al., 2023; Panaganti & Kalathil, 2022; Yang et al., 2022; Zhou et al., 2021). In some model-based papers, access to a generative-model is assumed, which is not realistic in many cases (Wang & Zou, 2021; Xu et al., 2023). Whether one assumes generative access or not, the number of parameters that need to be estimated in a model-based approach grows with the cardinality of the state and action spaces, unless one makes additional structural assumptions on the model.

Another line of work focuses on model-free learning of robust policies, that is, learning without constructing an empirical transition matrix. In the tabular setting, Liang et al. (2023) analyzes Cressie–Read f-divergence–based uncertainty sets and establishes asymptotic convergence guarantees for robust temporal-difference (TD) learning. A complementary tabular result, Li et al. (2022), studies the R-contamination uncertainty set and exploits a distinctive property: the robust Bellman operator

in this model admits an unbiased stochastic estimator. The techniques developed there extend to any uncertainty set that likewise permits an unbiased estimator of the robust Bellman operator, enabling unbiased policy evaluation and, consequently, policy improvement in a model-free manner. However, these papers do not consider function approximation, which is essential to deal with large stae spaces.

When function approximation is introduced to represent the robust value function, the literature typically proceeds along two directions with different limitations. One line of research constructs the uncertainty set expressly so that the robust Bellman operator admits an unbiased estimator (Zhou et al., 2023), allowing standard stochastic approximation arguments to go through or restrict to R-contamination uncertainty set (Wang & Zou, 2021). For R-contamination uncertainty set, Wang & Zou (2021) investigates the TD-C algorithm under function approximation and provides finite-time bounds for convergence to a stationary point of the associated objective, offering non-asymptotic guarantees in a setting where the objective is nonconvex and only stationarity is generally attainable. The other direction assumes extremely small discount factors to induce a contraction mapping for the robust Bellman operator, which restores fixed-point uniqueness and enables convergence proofs Zhou et al. (2023); Badrinath & Kalathil (2021); Tamar et al. (2014). Both approaches trade generality for tractability: the first restricts attention to uncertainty sets with unbiased estimators and focuses only on local optimality, while the second relies on unrealistically small discounting to guarantee contraction.

Another line of work (Tang et al., 2024; Ma et al., 2022) for model-free DRRL considers linear Markov decision process (MDP) for DRRL where the transition matrix of the underlying MDP has a lower-dimensional structure. This reduces the complexity associated with large state spaces. In this paper, we do not make such a modeling assumption.

In summary, most existing results on model-free robust RL are limited in at least one crucial way: they prove only local or asymptotic convergence; focus on narrow uncertainty models (e.g., Liang et al. (2023) observe on FrozenLake that R-contamination–based methods can mirror non-robust baselines and even underperform due to over-conservatism); restricted to tabular settings; assume generative access; or require extremely small discount factors. In particular, there are no finite-time guarantees for robust TD with function approximation from a single trajectory under broad, practically motivated uncertainty classes—such as those induced by total variation or Wasserstein-ℓ distances. At the same time, practice-oriented deep-RL pipelines often use ad-hoc "robust TD" heuristics, leaving a sizable gap between theory and deployment. This work closes a portion of that gap by establishing finite-time guarantees for robust TD learning with function approximation under commonly used uncertainty sets, without relying on generative sampling, vanishing discount factors, or purely asymptotic arguments.

Contributions. Our main contributions are summarized below.

1. **General conditions and finite-time guarantees.** We identify a set of mild structural conditions—satisfied by widely used uncertainty metrics such as total variation and Wasserstein-ℓ—under which distributionally robust policy evaluation considered in the paper with linear function approximation admits non-asymptotic guarantees from a single trajectory. For any uncertainty model obeying these conditions, our robust TD method achieves an ϵ-accurate value estimate with sample complexity $\tilde{O}(1/\epsilon^2)$.
2. **Overcoming projection mismatch via target networks.** While the robust Bellman operator is a contraction in ℓ_∞ (Iyengar, 2005), function approximation induces a projected fixed-point equation that breaks direct contraction arguments. Prior approaches either remain tabular or require unrealistically small discount factors. We resolve this by incorporating a target-network mechanism—conceptually related to Munos & Szepesvári (2008) and, in the non-robust setting, Chen et al. (2023)—and prove stable, finite-time convergence of the resulting projected robust TD updates without restrictive discount-factor assumptions.
3. **Function approximation in the dual space.** Standard DRRL solvers compute the worst-case distribution at each step of an RL algorithm by using a dual formulation Iyengar (2005). However, this requires estimating a dual variable for each (state, action) pair, which is infeasible for large state spaces. To overcome this problem, we provide the first analysis of function approximation in the dual space.
4. **Robust Q-Learning.** The main technical contributions of the paper are in the proof of convergence and sample complexity bounds for robust TD learning with function approximation.

It is straightforward to use these ideas to obtain finite-time bounds for robust Q-learning with function approximation, which, to the best of our knowledge, has not been studied in the literature. We refer the reader to the short argument in the supplemental material (Section E).

Since our paper focuses on discounted-reward robust RL, we have not made an exhaustive comparison of our work with work on average-reward robust RL; see, for example, Xu et al. (2025); Roch et al. (2025); Chen et al. (2025). However, to the best of our knowledge, it is worth noting that there are no performance guarantees even in the average-reward literature when function approximation is used.

2 Model and Preliminaries

Model We consider finite state and finite action infinite horizon discounted MDPs denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(\cdot \mid s, a) \in \Delta_{\mathcal{S}}$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the bounded instantaneous reward, and $\gamma \in (0, 1)$ the discount factor. A (stochastic) policy π maps states to distributions over actions: $\pi(a \mid s) \in \Delta_{\mathcal{A}}$. For any policy π and transition model P, the (policy-dependent) state–action value is

$$Q_P^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \;\middle|\; S_0 = s, A_0 = a,\ A_t \sim \pi(\cdot \mid S_t),\ S_{t+1} \sim P(\cdot \mid S_t, A_t)\right].$$

Robust MDPs (RMDPs) and uncertainty sets. Distributionally robust RL (DRRL) models transition uncertainty via an *uncertainty set* around a nominal kernel P_0. We adopt the standard (s, a)-rectangular model:

$$\mathcal{P}_s^a \;=\; \Big\{ q \in \Delta_{\mathcal{S}} :\; D\big(q,\, P_0(\cdot \mid s, a)\big) \le \delta \Big\}, \qquad \mathcal{P} \;=\; \bigotimes_{(s,a)} \mathcal{P}_s^a, \tag{1}$$

where $D(\cdot, \cdot)$ is a probability distance or divergence (e.g., total variation or Wasserstein-ℓ), and $\delta > 0$ is the radius. An RMDP is then the tuple

$$\mathcal{M}_r = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma).$$

Robust value functions (fixed policy). Given a fixed policy π, the *robust* state–action value function is the worst-case value over $\mathcal{P}$ (subscript r stands for "robust"):

$$Q_r^\pi(s,a) \;:=\; \min_{P \in \mathcal{P}} Q_P^\pi(s,a), \qquad V_r^\pi(s) \;:=\; \sum_a \pi(a \mid s)\, Q_r^\pi(s,a). \tag{2}$$

It satisfies the *robust Bellman equation*:

$$Q_r^\pi(s,a) \;=\; r(s,a) \;+\; \gamma \min_{q \in \mathcal{P}_s^a} \sum_{s'} q(s' \mid s,a) \underbrace{\sum_{a'} \pi(a' \mid s')\, Q_r^\pi(s', a')}_{=:\, V_r^\pi(s')}. \tag{3}$$

Equivalently, defining the robust Bellman operator $(\mathcal{T}_r^\pi Q)(s,a) := r(s,a) + \gamma\, \sigma_{\mathcal{P}_s^a}(V_r^\pi)$ with

$$\sigma_{\mathcal{P}_s^a}(V) \;:=\; \min_{q \in \mathcal{P}_s^a} \sum_{s'} q(s' \mid s,a)\, V(s'), \qquad V_r^\pi(s') := \sum_{a'} \pi(a' \mid s')\, Q(s', a'), \tag{4}$$

the fixed point relation is $Q_r^\pi = \mathcal{T}_r^\pi Q_r^\pi$. We can write from the definitions,

$$|V_r^\pi(s)| \le \frac{1}{1-\gamma}, \forall s; \qquad |Q_r^\pi(s,a)| \le \frac{1}{1-\gamma}, \forall (s,a).$$

For a fixed π, evaluating Q_r^π reduces to solving for Equation (3), which at each (s, a) requires solving the inner problem Equation (4).

2.1 Robust Temporal-Difference Learning: Challenges

Function approximation. Fix a policy π. We approximate the robust state–action value function by a linear function class

$$Q^{\pi}_{r,\theta}(s,a) = \phi(s,a)^{\top}\theta, \qquad \|\phi(s,a)\|_2 \le 1, \forall (s,a)$$

with feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$. Let $d^{\pi}(s,a)$ be the stationary distribution of (S_t, A_t) under π, and define $D^{\pi} := \mathrm{diag}\big(\{d^{\pi}(s,a)\}_{(s,a)}\big)$. Assume the weighted feature covariance is well-conditioned:

$$\Phi^{\top} D^{\pi} \Phi \succeq \mu I_d \quad \text{for some } \mu > 0.$$

Let $\mathcal{W} := \{\Phi\theta : \theta \in \mathbb{R}^d\}$ and denote by $\Pi : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathcal{W}$ the D^{π}-orthogonal projection,

$$\Pi f \;=\; \Phi(\Phi^{\top} D^{\pi} \Phi)^{-1} \Phi^{\top} D^{\pi} f.$$

We define the function approximation error for approximating the robust Q-function as:

$$\epsilon_{approx} := \sup_{Q = Clip\left(\Phi\theta, -\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right), \theta \in \mathbb{R}^d} \left\| Clip\left(\Pi \mathcal{T}^{\pi}_r(Q), -\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right) - \mathcal{T}^{\pi}_r(Q) \right\|_{\infty}, \quad (5)$$

where $\mathrm{Clip}(f, a, b)$ denotes the element-wise clipping of a vector f to the interval $[a, b]$.

Key challenges in robust policy evaluation and our approach. Model-free robust policy evaluation on a single trajectory typically hinges on a data-driven unbiased estimate $\hat{\sigma}_{\mathcal{P}^a_s}(V)$ of the inner-optimization objective defined in Equation (4). Except for special uncertainty sets (e.g., R-contamination), there is no direct plug-in *unbiased* single-sample estimator of this inner minimum, which creates a bias in standard TD updates. To overcome this challenge, we use a *two-time-scale* stochastic-approximation scheme in the inner loop of the algorithm: a fast time scale solves for the inner optimization problem defined in Equation (4) in its equivalent dual form, while the slow loop performs TD learning updates on θ using the estimate of the inner-optimization objective of the fast time scale. Our two-time scale algorithm is motivated by the algorithm in Liang et al. (2023), but the key difference here is the use of function approximation which necessitates a different analysis.

While $\mathcal{T}^{\pi}_r$ is a γ-contraction in ℓ_∞-norm (Iyengar, 2005), function approximation introduces the *projected* operator $\Pi\mathcal{T}^{\pi}_r$, which is *not* known to be a contraction in any norm for typical $\gamma \in (0,1)$. Prior work Zhou et al. (2023) circumvents this by imposing restrictive assumptions on γ which we do not adopt. We address the non-contraction of $\Pi\mathcal{T}^{\pi}_r$ via a *target-network* mechanism prevalent in deep RL, analyzed by Munos & Szepesvári (2008) and later used in the non-robust setting by Chen et al. (2023), for Q-learning to overcome the contraction issue with the projected robust Bellman operator. At outer iteration t, we freeze a target parameter $\hat{\theta}_t$ and solve

$$\Phi\theta \;=\; \Pi\, \mathcal{T}^{\pi}_r(\Phi\hat{\theta}_t)$$

in the inner loop, then update the target in the outer loop. This decoupling stabilizes the projected robust updates and enables our finite-time analysis under linear function approximation.

3 Robust TD Learning with Linear Function Approximation

Before presenting the robust policy evaluation algorithm, we discuss a few assumptions on the uncertainty sets considered.

3.1 Uncertainty Sets

We outline a few properties of uncertainty sets in the following assumption, satisfied by common uncertainty sets defined by distance metrices D in Equation (1) as the total variation distance $D_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$ and the Wasserstein-ℓ distance (discussed in detail later). We provide theoretical convergence guarantee of the robust policy evaluation Algorithm 1 under Assumption 1. In Section 5, we discuss how our algorithm can be trivially modified to satisfy a similar convergence guarantee for the R-contamination uncertainty set and Cressie-Read family of f-divergences considered in Liang et al. (2023).

Assumption 1 *The optimization problem $\sigma_{\mathcal{P}_s^a}(V)$ for a generic value function V as defined in Equation (4) has an equivalent dual optimization problem corresponding to a dual variable λ_s^a :*

$$\sigma_{\mathcal{P}_s^a}(V) \equiv \sup_{\lambda_s^a \geq 0} \left(F(\lambda_s^a; V, P_0(\cdot|s,a))\right)$$

where $F(\lambda_s^a; V, P_0(\cdot|s,a))$ is a λ_s^a-concave function with the following properties:

1. *There exists at least one maximizer in the compact set $|\lambda_s^a| \leq \lambda_M$ for some $\lambda_M < \infty$.*
2. *Let $G(\lambda_s^a; V; P_0(\cdot|s,a))$ be a supergradient of the concave function $F(\lambda_s^a; P_0(\cdot|s,a))$. There exists an unbiased bounded estimator of $G(\lambda_s^a; V, P_0(\cdot|s,a))$ as $g(\lambda_s^a; S', V)$ from a sample of the next state as $S' \sim P_0(\cdot|s,a)$, that is, $\mathbb{E}_{S'\sim P_0(\cdot|s,a)}[g(\lambda_s^a; S', V)] = g(\lambda_s^a; V; P_0(\cdot|s,a))$ satisfying $|g(\lambda_s^a, S'; V)| \leq g_M < \infty$ for all $|\lambda_s^a| \leq \lambda_M$.*
3. *There exists an unbiased estimator of the dual objective $F(\lambda_s^a; V, P_0(\cdot|s,a))$ denoted as $\sigma(\lambda_s^a; S', V)$ from a sample of next state as S' satisfying, $\mathbb{E}_{S'\sim P_0(\cdot|s,a)}\sigma(\lambda_s^a; S', V) = F(\lambda_s^a; V, P_0(\cdot|s,a))$ and $|\sigma(\lambda_s^a; S', V)| \leq \sigma_M$ for some $\sigma_M < \infty$ for all $|\lambda_s^a| \leq \lambda_M$.*

3.1.1 Uncertainty Sets Satisfying Assumption 1

Total Variation Uncertainty Metric: The total variation uncertainty set is defined as: for each (s,a), $\mathcal{P}_s^{aTV} = \{q \in \Delta(\mathcal{S}) : \frac{1}{2}\|q - P_0(\cdot|s,a)\|_1 \leq \delta\}$.

Simplications (see: Appendix B) on the dual formulation originally given by Iyengar (2005) for the Total Variation uncertainty set. We get the following equivalent dual optimization:

$$\sigma_{\mathcal{P}_s^a}(V) \equiv \max_{\lambda_s^a \in [-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]} \{\mathbb{E}_{P_0(\cdot|s,a)}[\min\left(V(X), \lambda_s^a\right)] - \delta\lambda_s^a\}$$

In Appendix B, we prove that the Total Variation uncertainty set satisfies Assumption 1 with $\lambda_M = \frac{1}{1-\gamma}$, $g_M = \max(\delta, 1-\delta), \forall \lambda_s^a : |\lambda_s^a| \leq \lambda_M$ and $\sigma_M = \frac{1}{1-\gamma} + \delta, \forall \lambda_s^a : |\lambda_s^a| \leq \lambda_M$.

Wasserstein-ℓ uncertainty Set: The uncertainty metric is defined as: for each (s,a) as: $\mathcal{P}_s^{aW_\ell} = \{q \in \Delta(|\mathcal{S}|) : W_\ell(P_0(\cdot|s,a), q) \leq \delta\}$, where $\delta > 0$ is the uncertainty radius and $W_\ell(P_0(\cdot|s,a), q)$ is the Wasserstein-ℓ distance defined in detail in Appendix B.2.

In Appendix B, we prove that the Wasserstein-ℓ uncertainty set satisfies Assumption 1 with $\lambda_M = \frac{span(V)}{\delta^\ell}$, $g_M \coloneqq 1 + \delta^\ell, \forall \lambda_s^a \in [0, \lambda_M]$ and $\sigma_M \coloneqq (\delta^\ell + 1)\lambda_M^{W_\ell} + \frac{1}{1-\gamma}, \forall \lambda_s^a \in [0, \lambda_M]$.

3.2 Algorithm and Main Results

In this subsection, we present our robust policy evaluation algorithm and the main results of the paper.

3.2.1 Robust Policy Evaluation Algorithm

Our robust TD learning algorithm is presented in Algorithm 1. In the rest of this section, we describe the algorithm and explain the notation used in the algorithm. In the outer loop (indexed by $t = 0, ..., T-1$), we freeze a *target parameter* $\hat{\theta}_t$; at the end of the inner loop we set $\hat{\theta}_{t+1}$ to the inner loop's final iterate. In the inner loop (indexed by $k = 0, .., K-1$) we approximately solve for θ satisfying:

$$\Phi\theta \;=\; \Pi\, \mathcal{T}_r^\pi(\Phi\hat{\theta}_t),$$

using a two-time-scale stochastic approximation: a fast loop for the dual variables corresponding to the inner optimization problem 4, and a slow loop for the TD parameters. For a fixed outer loop t, the inner loop iterates are $\theta_{t,k}$ for $k \in [0, K]$.

Each inner loop iteration k, in a fast time scale, we approximately solve the equivalent dual optimization problem in (4). Instead of maintaining a separate dual variable λ_s^a for each (s,a) (which would be tabular), we parameterize the dual variables λ_s^a with the learnable parameter vector $\nu \in \mathbb{R}^{\lambda_d}$ as

$$\lambda_s^a \approx \psi(s,a)^\top \nu, \qquad \|\psi(s,a)\|_2 \leq 1, \forall (s,a)$$

with feature matrix $\Psi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d_\lambda}$. We define the function approximation error for approximating the dual variables next. For compactness of notation, denote for a value function V, for each (s,a) , $F^{*,V}_{s,a} \coloneqq \sup_{\lambda^a_s} F(\lambda^a_s; V, P_0(\cdot|s,a))$ and $F^V(\nu)_{s,a} \coloneqq F(\psi(s,a)^\top \nu; V, P_0(\cdot|s,a))$. Define

$$\epsilon^{dual}_{approx} \coloneqq \sup_{V:Q=Clip(\Phi\theta,-\frac{1}{1-\gamma},\frac{1}{1-\gamma});\theta\in\mathbb{R}^d;\pi\in\Pi} \inf_{\nu\in\mathcal{M}_\nu} \|F^{*,V} - F^V(\nu)\|_\infty \tag{6}$$

Denote the value function estimate $V_{\hat{\theta}_t}$ evaluated at the target parameter $\hat{\theta}_t$ as

$$V_{\hat{\theta}_t}(s) = \sum_a \pi(a|s) Clip\left(\phi(s,a)^\top \hat{\theta}_t, -\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right), \forall s \in \mathcal{S}. \tag{7}$$

The quantity $V_{\hat{\theta}_t}$ can be computed exactly for any fixed target parameter $\hat{\theta}_t$. In the case of the TV distance uncertainty set, it suffices to compute $V_{\hat{\theta}_t}(s)$ *only for the state visited in each inner-loop iteration*, rather than for all states. We update $\nu_{t,k}$ with step size β_k using a projected *super-gradient ascent* on the dual objective with a super-gradient evaluated at the fresh data sample $(S^t_k, A^t_k, S^t_{k+1})$. The projection $Proj_{\mathcal{M}_\nu}$ enforces projection of dual parameter vector into the set $\mathcal{M}_\nu \coloneqq \{\nu \in \mathbb{R}^{d_\lambda} : \|\nu\|_2 \le B_\nu\}$ to keep the iterates bounded.

In the algorithm, $\bar{\nu}_{t,k}$ denotes the half-tail iterate-average of the dual parameter vector, i.e.,

$$\bar{\nu}_{t,k} = \lceil 2/k \rceil \sum_{l=\lfloor k/2 \rfloor}^{k-1} \nu_{t,l} \tag{8}$$

which can be calculated easily by keeping track of the following two quantities: $\sum_{l=0}^{k-1} \nu_{t,l}$ and $\sum_{l=\lfloor k/2 \rfloor}^{k-1} \nu_{t,l}$. While many elements of our algorithm have been used in implementations of robust TD learning, to the best of our knowledge, such an averaging of the dual variables has not been used previously. The averaging turns out to be crucial in obtaining finite-time bounds, since it allows us to control the variance of the dual objective.

In the slow time scale of the inner loop, θ^t_k is updated using asynchronous stochastic approximation with a step size denoted by α_k with a robust TD-target $TD_{t,k-1}$. The two-time-scale scheme ensures that, at the slow scale, the dual variables appear near their sample-path equilibrium, yielding an (asymptotically) unbiased robust TD target.

3.2.2 Main Result

We make the following assumption on the policy π.

Assumption 2 *The policy π induces an irreducible and aperiodic Markov chain under the nominal transition kernel P_0.*

Let $\mu^{t,\tau}_k(\cdot) := \mathbb{P}(S^t_k, A^t_k \in \cdot \mid S^t_0, A^t_0, S^t_1, A^t_1, ..., S^t_{k-\tau}, A^t_{k-\tau})$ for some $\tau \le k$. Note that the above assumption ensures that the Markov chain is geometrically mixing :

$$\exists C_{mix} < \infty : \|\mu^t_k - d^\pi\|_{\mathrm{TV}} \;\le\; C_{\mathrm{mix}}\, \rho^\tau \qquad (0<\rho<1), \qquad \forall t \in [0, T-1].$$

In Theorem 1, we present our main result, which establishes the convergence of $\hat{Q}_T$ to the robust value function Q^π_r, up to terms arising from function-approximation error.

Theorem 1 *Let $\hat{Q}_t \coloneqq \Phi\hat{\theta}_t$ be the estimate of Q^π_r returned from Algorithm 1 at iteration t. Under Assumptions 1 and 2, the following guarantee holds for the algorithm:*

$$\mathbb{E}\|\hat{Q}_T - Q^\pi_r\|_\infty \le \gamma^T \|\Phi\theta_0 - Q^\pi_r\|_\infty + \frac{A}{1-\gamma} + \frac{\epsilon_{approx}}{1-\gamma} + \frac{2\epsilon^{dual}_{approx}}{\mu(1-\gamma)}, \tag{9}$$

where the constant A depends on the chosen schedule.

Algorithm 1 Robust TD learning with Function Approximation

Input: Integers T, K. Initial $\nu_0 \in \mathbb{R}^{d_\lambda}$, $\theta_0 :=$ zero vector, fast time-scale step-sizes $\beta_k = \frac{\beta_0}{\sqrt{k+1}}$, slow time-scale step-sizes $\alpha_k = \frac{c}{(k+1)^\omega}$ for some $\omega \in (0.5, 1]$; $\hat{\theta}_0 = \theta_0, \theta_{0,0} = \theta_0$, candidate policy π, Reward function $r : (\mathcal{S} \times \mathcal{A}) \mapsto [0, 1]$, initial state S_0^0.
for $t = 0, 1, ..., T-1$ **do**
  for $k = 0, 1, \ldots, K-1$ **do**
    Take action A_k^t according to policy π and Sample S_{k+1}^t ($S_{k+1}^t \sim P_0(\cdot|S_k^t, A_k^t)$)
  Fast scale (β_k)
    Compute $g(\psi(S_k^t, A_k^t)^\top \nu_{t,k}; S_{k+1}^t, V_{\hat{\theta}_t})$ from Equation (19) for TV distance and Equation (22) for Wasserstein-ℓ uncertainty
    $\nu_{t,k+1} = Proj_{\mathcal{M}_\nu}(\nu_{t,k} + \beta_k[g(\psi(S_k^t, A_k^t)^\top \nu_{t,k}; S_{k+1}^t, V_{\hat{\theta}_t})\psi(S_k^t, A_k^t)])$
  Slow scale (α_k)
    Compute $\bar{\nu}_{t,k}$ from Equation (8)
    Compute $\sigma(\psi(S_k^t, A_k^t)^\top \bar{\nu}_{t,k}; S_{k+1}^t, V_{\hat{\theta}_t})$ from Equation (20) for TV distance and Equation (23) for Wasserstein-ℓ uncertainty
    $TD_{t,k+1} = r(S_k^t, A_k^t) + \gamma\sigma(\psi(S_k^t, A_k^t)^\top \bar{\nu}_{t,k}; S_{k+1}^t, V_{\hat{\theta}_t}) - \phi(S_k^t, A_k^t)^\top \theta_{t,k}$
    $\theta_{t,k+1} = \theta_{t,k} + \alpha_k TD_{t,k+1}\phi(S_k^t, A_k^t)$
  end for
  $\hat{\theta}_{t+1} = \theta_{t,K}, S_0^{t+1} = S_K^t, \theta_{t,0} = \theta_0, \nu_{t,0} = \nu_0$
end for
Output: $\hat{\theta}_T$

(A) Polynomially decaying step size: *For the slow time scale step-size* $\alpha_k = \frac{c}{(1+k)^\omega}; \omega \in (1/2, 1)$,

$$A := \sqrt{\frac{C_*}{(K+1)^\omega}}, \tag{10}$$

where, with the notation $k_0 := \lceil \frac{2C_e c}{\mu} \rceil$

$$C_* = \max\left\{ (k_0+1)^\omega H_1 \left(\frac{1}{(1-\gamma)^2\mu} + \frac{2\sigma_M^2}{\mu^2} + c^2 C_1 \left(\frac{2\omega}{2\omega - 1} \right) \right), \frac{8c^2 C_1}{c\mu} \right\},$$

$$C_1 = \frac{\max(C_{mix}, 1)^2}{\mu^2(1-\gamma)^2} \left(2^9 \beta_0 g_M^4 + \frac{B_\nu^4}{\beta_0^2} + 2^{10} B_\nu^2 g_M^2 + 2^8 \gamma \sigma_M + 2^{10} + 5\gamma^2 \sigma_M^2 \right)$$

$$C_2 = 1 + 4C_{mix} \ + \ 16\big(1 + C_{mix}^2 c^2\big); \quad H_1 = \prod_{k=0}^{k_0-1} \left(1 - \frac{c\mu}{(k+1)^\omega} + \frac{C_2 c^2}{(k+1)^{2\omega}} \right)$$

(B) Harmonic step size : *For the slow time scale step size rule* $\alpha_k = \frac{c}{k+1}$,

$$A := \sqrt{H_2 \left(\frac{1}{(1-\gamma)^2\mu} + (C_1 c^2 + 2c\sigma_M^2) \ln(k_0) \right) \left(\frac{k_0+1}{K+1} \right)^{\frac{c\mu}{2}} + I}$$

$$I := \begin{cases} \frac{c^2 C_1}{\frac{c\mu}{2}-1} \frac{1}{K+1} + \frac{c^2 C_1}{\frac{c\mu}{2}-1} \frac{(k_0+2)^{\frac{c\mu}{2}-1}}{(K+1)^{\frac{c\mu}{2}}} & , \frac{c\mu}{2} > 1 \\ \frac{c^2 C_1}{K+1} \left(1 + \ln\left(\frac{K+1}{k_0+2} \right) \right) & , \frac{c\mu}{2} = 1 \\ \frac{c^2 C_1}{1-\frac{c\mu}{2}} \frac{(k_0+2)^{\frac{c\mu}{2}-1}}{(K+1)^{\frac{c\mu}{2}}} & , 0 < \frac{c\mu}{2} < 1 \end{cases}$$

where,

$$H_2 = \prod_{k=0}^{k_0-1} \left(1 - \frac{c\mu}{(k+1)} + \frac{C_2 c^2}{(k+1)^2} \right), \quad k_0 := \lceil (2C_e c)/\mu \rceil$$

Corollary 1 (Sample Complexity) *The following sample complexity results hold:*

- *If the step size rule $\alpha_k = \frac{c}{(1+k)^\omega}$, $\omega \in (1/2, 1)$, is used, then with $T = O\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ and $K = O\left(\frac{1}{\epsilon^{\frac{2}{\omega}}}\right)$, Algorithm 1 achieves an element-wise ϵ-accurate estimate of Q_r^π up to the function approximation error. Thus, to achieve this approximation error, the sample complexity is*

$$O\left(\ln\left(\frac{1}{\epsilon}\right)\frac{1}{\epsilon^{\frac{2}{\omega}}}\right). \tag{11}$$

- *If the step size rule $\alpha_k = \frac{c}{1+k}$ is used, then the sample complexity is given by $O\left(\ln\left(\frac{1}{\epsilon}\right)\frac{1}{\epsilon^2}\right)$ if $c\mu \geq 2$.*

We note that the step size rule $c/(1+k)$ achieves the best sample complexity, but it requires c to be chosen sufficiently large. This is consistent with similar results in the non-robust RL literature; see, for example, Chen et al. (2023).

4 Key Ideas and Proof Outline

While the detailed proof of Theorem 1 is presented in Appendix C, we provide the key ideas behind the proof in this section.

Fix an outer loop iteration t. Recall the definition $F_{s,a}^{*,V} := \sup_{\lambda_s^a} F(\lambda_s^a; V, P_0(\cdot|s,a))$. Define the inner loop error for outer iteration index t as $e_k^t := \theta_{t,k} - \theta^{*,t}$ with

$$\theta^{*,t} := (\Phi^\top D^\pi \Phi)^{-1}\Phi^\top D^\pi \left[r + \gamma\, F^{*,V_{\hat{\theta}_t}}\right] \tag{12}$$

The next lemma bounds the expected estimation error at the final outer-loop iterate in terms of the inner-loop error terms.

Lemma 1 *Under Assumptions 1 and 2, Algorithm 1 guarantees*

$$\mathbb{E}\|\hat{Q}_T - Q_r^\pi\|_\infty \leq \gamma^T \|\Phi\theta_0 - Q_r^\pi\|_\infty + \underbrace{\sum_{t=1}^{T}\gamma^{T-t-1}\mathbb{E}\left[\|e_k^t\|_\infty\right]}_{\textit{Inner loop convergence error}} + \frac{\epsilon_{approx}}{1-\gamma}.$$

The proof of Lemma 1 is provided in Appendix C.1 and is inspired by the analysis in Chen et al. (2023) for non-robust Q-learning.

In the analysis that follows, we establish that the inner loop error remains small (up to function approximation error terms) in ℓ_∞-norm for sufficiently large k. We decompose the slow time-scale update at inner loop k in Algorithm 1 into mean drift, noise and bias terms as

$$\theta_{t,k+1} = \theta_{t,k} + \alpha_k\left[G(\theta_{t,k}) \;+\; b_{t,k}^\theta \;+ n_{t,k+1}^\theta\right],$$

where

$$G(\theta_{t,k}) := \Phi^\top D^\pi\left[r + \gamma F^{*,V_{\hat{\theta}_t}} - \Phi\theta_{t,k}\right] \underbrace{=}_{\text{from Equation (12)}} \Phi^\top D^\pi \Phi(\theta^{*,t} - \theta_{t,k}),$$

$$b_{k,t}^\theta := \Phi^\top D^\pi\left[F^{V_{\hat{\theta}_t}}(\bar{\nu}_{t,k}) - F^{*,V_{\hat{\theta}_t}}\right],$$

with the noise term n_{k+1}^θ collects all remaining terms.

Idealized recursion (without noise and bias). The mean drift term corresponds to the deterministic recursion:

$$\theta_{t,k+1} = \Phi^\top D^\pi \Phi(\theta^{*,t} - \theta_{t,k})$$

his recursion admits $\theta^{*,t}$ as its unique fixed point. Since the matrix $\Phi^\top D^\pi \Phi$ is symmetric and positive definite with minimum eigenvalue $\mu > 0$, in the absence of bias and noise terms, the iterates satisfy

$$\|\theta_{t,k+1} - \theta^{*,t}\|_2 \;\leq\; (1 - \alpha_k\mu)\,\|\theta_{t,k} - \theta^{*,t}\|_2, \tag{13}$$

which implies geometric convergence of $\theta_{t,k}$ to $\theta^{*,t}$ at a rate governed by μ.

Bias term analysis. Recall that the bias term is given by

$$b^{\theta}_{t,k} \;:=\; \Phi^{\top} D^{\pi}\Big[F^{V_{\hat{\theta}_t}}(\bar{\nu}_{t,k}) - F^{*,V_{\hat{\theta}_t}}\Big].$$

We show that this term becomes small for large k, up to a function approximation error $\epsilon^{\text{dual}}_{\text{approx}}$.

In the fast time scale analysis, we prove that the stochastic update on ν performs a supergradient ascent on the concave objective

$$L^t(\nu) \;:=\; \sum_{s,a} d^{\pi}(s,a)\, F\big(\psi(s,a)^{\top}\nu;\, V_{\hat{\theta}_t},\, P_0(\cdot|s,a)\big),$$

which has bounded supergradients. By a standard Lyapunov argument for stochastic approximation under a mixing Markov chain, we obtain

$$\mathbb{E}\Big[\max_{\nu\in M_\nu} L^t(\nu) - L^t(\bar{\nu}_k)\Big] \;\le\; \frac{C_{\text{fast}}}{\sqrt{k}}, \tag{14}$$

where the constant C_{fast} is given in equation 26.

Using $\|\phi(s,a)\|_2 \le 1$ for all (s,a), we can write

$$\|b^{\theta}_{t,k}\|_2 \le \gamma \sum_{s,a} d^{\pi}(s,a)\left|F^{V_{\hat{\theta}_t}}(\bar{\nu}_{t,k})_{s,a} - F^{*,V_{\hat{\theta}_t}}_{s,a}\right| = \gamma \sum_{s,a} d^{\pi}(s,a)\left(F^{*,V_{\hat{\theta}_t}}_{s,a} - F^{V_{\hat{\theta}_t}}(\bar{\nu}_{t,k})_{s,a}\right)$$

$$\le \gamma \underbrace{\inf_{\nu\in M_\nu} \sum_{s,a} d^{\pi}(s,a)\left(F^{*,V_{\hat{\theta}_t}}_{s,a} - F^{V_{\hat{\theta}_t}}(\nu)_{s,a}\right)}_{\le \epsilon^{\text{dual}}_{\text{approx}}} + \gamma \underbrace{\Big[\sup_{\nu\in M_\nu} L^t(\nu) - L^t(\bar{\nu}_k)\Big]}_{\text{fast-scale objective gap}}.$$

Handling the noise term. Finally, to handle the noise terms $n^{\theta}_{t,k+1}$, we employ the approach in Srikant & Ying (2019), where a bound is obtained on expectation of the error $\|\theta_{t,k} - \theta^{*,t}\|_2^2$ conditioned on a lagged filtration over the set $(S^t_0, A^t_0, S^t_1, A^t_1, ..., S^t_{k-\tau}, A^t_{k-\tau})$. By choosing a lag τ such that the underlying Markov chain has mixed sufficiently, the effect of noise can be controlled.

5 Discussion

As mentioned in the introduced, we provide the first proof of convergence and finite-time bounds for robust TD learning with function approximation without making any assumptions on the underlying model or making very restrictive assumptions on the discount factor. Some immediate extensions and open problems are identified below:

1. The algorithm and the results can be extended to other families of distances between probability distributions, such as the Cressie-Read family of f-divergences considered in Liang et al. (2023), which admit duality representations that allow one to obtain unbaised estimators of the quantities of interest. For the Cressie-Read family, this would require the addition of one more time-scale but the rest of the analysis would be similar. Our results also apply to the R-contamination set, but the algorithm is even simpler in that case due to the fact that the dual problem has a closed-form solution Xu et al. (2025).
2. While the computational complexity of the algorithm is quite small for TV distance uncertainty set, the super-gradient computation for the Wasserstein-ℓ distance can be quite prohibitive for large state-spaces. This is due to the fact that the super-gradient computation in Equation (22), in the supplemental material, requires a minimization over all states. It is an interesting open question whether this computational complexity can be mitigated for Wassertein-ℓ distances.
3. Although the results in the main body of the paper have been presented for robust TD learning, they can be easily extended to robust Q-learning with function approximation to obtain optimal policies; see the supplemental material.

REFERENCES

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

Zaiwei Chen, John-Paul Clarke, and Siva Theja Maguluri. Target network and truncation overcome the deadly triad in-learning. *SIAM Journal on Mathematics of Data Science*, 5(4):1078–1101, 2023.

Zijun Chen, Shengbo Wang, and Nian Si. Sample complexity of distributionally robust average-reward reinforcement learning. *arXiv preprint arXiv:2505.10007*, 2025.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Yan Li, Guanghui Lan, and Tuo Zhao. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.

Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

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

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Zachary Roch, Chi Zhang, George Atia, and Yue Wang. A finite-sample analysis of distributionally robust average-reward reinforcement learning. *arXiv preprint arXiv:2505.12462*, 2025.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on learning theory*, pp. 2803–2830. PMLR, 2019.

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *International conference on machine learning*, pp. 181–189. PMLR, 2014.

Cheng Tang, Zhishuai Liu, and Pan Xu. Robust offline reinforcement learning with linearly structured f-divergence regularization. *arXiv preprint arXiv:2411.18612*, 2024.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Yang Xu, Washim Uddin Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. *arXiv preprint arXiv:2502.16816*, 2025.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022.

Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36:97–133, 2023.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

A CONTENTS

The contents of the Appendix are as follows:

1. In Section B, we analyze the TV distance and Wasserstein-ℓ uncertainty sets in detail. We prove that both of them satisfy Assumption 1.
2. Section C proves the main result of the paper, that is, Theorem 1 in detail.
3. In Section E, we present the robust Q learning algorithm (Algorithm 2) and discuss how the theoretical analysis for robust TD learning can be extended to the robust Q learning straightforwardly.

B DETAILED ANALYSIS ON UNCERTAINTY METRICES CONSIDERED IN THIS PAPER

In this section, we discuss in detail the uncertainty sets considered in this paper, namely, TV distance uncertainty and Wasserstein-ℓ uncertainty sets. For each uncertainty set,

1. We define the uncertainty set first. Then, we discuss and analyze the equivalent dual optimization that corresponds to the inner-optimization problem defined in Equation 4.
2. We show the uncertainty set satisfies Assumption 1 and provide data-driven, unbiased estimates of the dual objective and the corresponding super-gradient.

B.1 TOTAL VARIATION DISTANCE UNCERTAINTY SET

The total variation uncertainty set is defined for each (s, a) pair as,

$$\mathcal{P}_s^{a\,TV} = \{q \in \Delta(\mathcal{S}) : \frac{1}{2}\|q - P_0(\cdot|s,a)\|_1 \leq \delta\}.$$

Next, we show that the optimization problem given in Equation 4 in the main body of the paper with $\mathcal{P}_s^{a\,TV}$ as the uncertainty set satisfies assumption 1. Let us rewrite the optimization problem here for the TV distance uncertainty set.

$$\sigma_{\mathcal{P}_s^{a\,TV}}(V) = \min_{q \in \mathcal{P}_s^{a\,W_\ell}} q^\top V$$

From Lemma 4.3 in Iyengar (2005), we know that the above optimization problem can be solved under the dual formulation :

$$\sigma_{\mathcal{P}_s^{a\,TV}}(V) = \max_{f \geq 0} \left(\mathbb{E}_{P_0(\cdot|s,a)}[V - f] - \delta span(V - f)\right) \tag{15}$$

Next, we prove that the above dual optimization problem is equivalent to a scalar optimization problem.

Lemma 2 *The optimization problem given in Equation (15) is equivalent to the following optimization problem: Let us say, $m = \min_s V(s)$ and $M = \max_s V(s)$.*

$$\sigma_{\mathcal{P}_s^{a,TV}}(V) \equiv \max_{\lambda_s^a \in [m,M]} \{\mathbb{E}_{P_0(\cdot|s,a)}[\min V(X), \lambda_s^a] - \delta\lambda_s^a\} \tag{16}$$

Proof 1 ***From the μ-vector dual to a 1–D cut off problem:*** *The optimization problem in Equation (15) can be written as*

$$\sigma_{\mathcal{P}_a^{a,TV}}(V) \;=\; \max_{f \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|}} \Big\{\mathbb{E}_{P_0(\cdot|s,a)}\big[V - f\big] \;-\; \delta\left[\max_{s'}(V - f) - \min_{s'}(V - f)\right]\Big\}. \tag{17}$$

Step 1 – restrict to "cut–off" vectors: *For any scalar $a \in [m, M]$, with $m := \min_s V(s),\ M := \max_s V(s)$, define*

$$f_a(s) \;:=\; \big[V(s) - a\big]_+ \;=\; \max\big\{0,\ V(s) - a\big\}.$$

Replacing an arbitrary feasible f by the corresponding $f_{a:=\max_s(V-f)}$ cannot decrease the objective in equation 17, so an optimizer always has the form f_{a^*}.

Step 2 – plug f_a into the objective. *Because $V(s) - f_a(s) = \min\{V(s), a\}$,*

$$\max_s\big(V - f_a\big) = a, \qquad \min_s\big(V - f_a\big) = m,$$

and

$$\mathbb{E}_{P_0}[V - f_a] \;=\; \mathbb{E}_{P_0}\big[\min\{V(X), a\}\big].$$

Substituting these identities into equation 17 yields the scalar *optimization*

$$\sigma_{\mathcal{P}_s^{a,TV}}(V) \;=\; \max_{a\in[m,M]}\Big\{\mathbb{E}_{P_0(\cdot|s,a)}\big[\min\{V(X), a\}\big] \;-\; \delta\, a\Big\}. \tag{18}$$

As we are dealing with V functions for which $V(s) \in \{-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\}$, the optimum dual variable lies in: $\lambda_s^a \in \{-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\}$ and we can equivalently write,

$$\sigma_{\mathcal{P}_s^{a,TV}}(V) = \max_{\lambda_s^a \in \{-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\}} \{\mathbb{E}_{P_0(\cdot|s,a)}[\min V(X), \lambda_s^a] - \delta\lambda_s^a\}$$

It is easy to verify that the concave objective has a super-gradient:

$$G^{TV}(\lambda_s^a; V) = \mathbb{P}_{X\sim P_0(\cdot|s,a)}[V(X) \geq \lambda_s^a] - \delta$$

An unbiased estimate of the supergradient for a value of λ_s^a and the value function V from a next state $S' \sim P_0(\cdot|s,a)$ can be given as:

$$g^{TV}(\lambda_s^a; S', V) = \mathbf{1}_{V(S')\geq\lambda_s^a} - \delta \tag{19}$$

An unbiased estimate of the dual objective for a value of λ_s^a and the value function V from a next state $S' \sim P_0(\cdot|s,a)$ can be given as

$$\sigma^{TV}(\lambda_s^a; S', V) = \min(V(S'), \lambda_s^a) - \delta \tag{20}$$

If we assume $|V(s)| \leq \frac{1}{1-\gamma}, \forall s$, its easy to see that,

$$|g^{TV}(\lambda_s^a; S', V)| \leq g_M^{TV} \coloneqq \max(\delta, 1-\delta), \forall \lambda_s^a \in [0, \lambda_M^{W_\ell}]$$

and,

$$|\sigma^{TV}(\lambda_s^a; S', V)| \leq \sigma_M^{TV} \coloneqq \delta + \frac{1}{1-\gamma}, \forall \lambda_s^a \in [0, \lambda_M^{W_\ell}]$$

B.2 Wasserstein-ℓ uncertainty Set

We define the Wasserstein-ℓ uncertainty set for each (s,a) pair as:

$$\mathcal{P}_s^{aW_\ell} = \{q \in \Delta(|\mathcal{S}|) : W_\ell(P_0(\cdot|s,a), q) \leq \delta\}$$

where $\delta > 0$ is the uncertainty radius and $W_\ell(P_0(\cdot|s,a), q)$ is the Wasserstein-ℓ distance defined next. Consider the generic metric space $(\mathcal{S}, d)$ by defining some distance metric d. For some parameter $\ell \in [1, \infty)$, and two distributions $p, q \in \Delta(\mathcal{S})$, define the Wasserstein-ℓ distance between them as $W_\ell(q,p) = \inf_{N\in\Gamma(p,q)} \|d\|_{\mu,\ell}$, where $\Gamma(p,q)$ denotes the distribution over $\mathcal{S}\times\mathcal{S}$ with marginal distributions p, q and $\|d\|_{N,\ell} = (\mathbb{E}_{(X,Y)\sim N}[d(X,Y)^\ell])^{1/\ell}$. Let us use the distance matrix with normalization, ensuring $|d(s,s')| \leq 1, \forall(s,s')$.

Next, we show that the following optimization problem with $\mathcal{P}_s^{aW_\ell}$ as the uncertainty set satisfies Assumption 1.

$$\sigma_{\mathcal{P}_s^{aW_\ell}}(V) = \min_{q\in\mathcal{P}_s^{aW_\ell}} q^\top V.$$

From Gao & Kleywegt (2023), we know that the above optimization problem can be solved under the dual formulation :

$$\sigma_{\mathcal{P}_s^a}(V) = \sup_{\lambda_s^a \geq 0} \left(-\lambda_s^a \delta^\ell + \mathbb{E}_{P_0(\cdot|s,a)}[\inf_y (V(y) + \lambda_s^a d(S,y)^\ell)] \right).$$

As the state space $\mathcal{S}$ is finite, we can replace the inner optimization $[\inf_y (V(y) + \lambda_s^a d(S,y)^\ell)]$ with $[\min_y (V(y) + \lambda_s^a d(S,y)^\ell)]$. Next, we show that the optimum dual variable of the above optimization problem lies inside a compact set $\left[0, \lambda_M^{W_\ell}\right]$ with $\lambda_M^{W_\ell} \coloneqq \frac{span(V)}{\delta^\ell}$.

As point-wise minimum of affine functions is concave, the above optimization problem is a concave optimization problem. It is easy to verify that the concave objective has a super-gradient:

$$G^{W_\ell}(\lambda_s^a; V, P_0(\cdot|s,a)) = -\delta^\ell + \mathbb{E}_{X \sim P_0(\cdot|s,a)}[d(X, y^*_{\lambda_s^a}(X))^\ell] \tag{21}$$

where,

$$y^*_{\lambda_s^a}(x) \in \arg\min_y [V(y) + \lambda_s^a d(x,y)^\ell]$$

Let us fix an $S = s$ and its minimizer $y^*_{\lambda_s^a}(x)$ for the inner optimization $[\inf_y (V(y) + \lambda_s^a d(s,y)^\ell)]$. Because the candidate $y = s$ is always feasible,

$$V(y^*_{\lambda_s^a}(s)) + \lambda d(s, y^*_{\lambda_s^a}) \leq V(s).$$

Rearrange:

$$d(s, y^*_{\lambda_s^a}) \leq \frac{V(s) - V(y^*_{\lambda_s^a})}{\lambda_s^a} \leq \frac{span(V)}{\lambda_s^a}.$$

Taking expectation in Equation 21 and using the above equation gives

$$G^{W_\ell}(\lambda_s^a; V, P_0(\cdot|s,a)) \leq \delta^l + \frac{span(V)}{\lambda_s^a}.$$

Now, for any $\lambda_s^a > \lambda_M^{W_\ell} = \frac{span(V)}{\delta^\ell}$, we have,

$$G^{W_\ell}(\lambda_s^a; V, P_0(\cdot|s,a)) \leq 0$$

Due to the concavity of the objective, a non-positive super-gradient means the function is non-increasing for all $\lambda_s^a > \lambda_M^{W_\ell}$. Combining the observation with the boundedness of the objective for bounded λ_s^a, we conclude that the supremum is attained and lies in $[0, \lambda_M^{W_\ell}]$.

An unbiased estimate of the supergradient for a value of λ_s^a and the value function V from a next state $S' \sim P_0(\cdot|s,a)$ can be given as:

$$g^{W_\ell}(\lambda_s^a; S', V) = -\delta^\ell + d(S', y^{*'})^\ell \tag{22}$$

where,

$$y^{*'} = \arg\min_y [V(y) + \lambda_s^a d(S', y)^\ell]$$

An unbiased estimate of the dual objective for a value of λ_s^a and the value function V from a next state $S' \sim P_0(\cdot|s,a)$ can be given as

$$\sigma^{W_\ell}(\lambda_s^a; S', V) = -\lambda_s^a \delta^\ell + V(y^{*'}) + \lambda_s^a d(S', y^{*'})^\ell \tag{23}$$

If we assume $|V(s)| \leq \frac{1}{1-\gamma}, \forall s$, its easy to show that,

$$|g^{W_\ell}(\lambda_s^a; S', V)| \leq g_M^{W_\ell} \coloneqq 1 + \delta^\ell, \forall \lambda_s^a \in [0, \lambda_M^{W_\ell}]$$

and,

$$|\sigma^{W_\ell}(\lambda_s^a; S', V)| \leq \sigma_M^{W_\ell} \coloneqq (\delta^\ell + 1)\lambda_M^{W_\ell} + \frac{1}{1-\gamma}, \forall \lambda_s^a \in [0, \lambda_M^{W_\ell}]$$

C Convergence Analysis of Algorithm 1 and the Proof of Theorem 1

In this section, we provide the proof of Theorem 1. The proof follows in a similar manner described in the proof sketch in the main body of the paper. We start with proving Lemma 1 which establishes the convergence of the outer loop iterates in terms of inner loop convergence error. Subsequently, we establish the convergence of the inner loop. Finally we combine them to prove Theorem 1.

C.1 Outer Loop Convergence Analysis: Proof of Lemma 1

In this subsection, we prove Lemma 1. The proof is inspired by the analysis in Chen et al. (2023) for non-robust Q-learning. The analysis of the outer loop follows from the paper (Chen et al., 2023). To write the bound for the outer loop, we have to start with a few notations as used in the mentioned paper. Recall, the function approximation error ϵ_{approx} is defined as:

$$\epsilon_{approx} \coloneqq \sup_{Q=Clip\left(\Phi\theta,-\frac{1}{1-\gamma},\frac{1}{1-\gamma}\right),\theta\in\mathbb{R}^d} \left\| Clip\left(\Pi\mathcal{T}_r^{\pi}(Q), -\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right) - \mathcal{T}_r^{\pi}(Q)\right\|_{\infty}.$$

Also, recall the definition of $\theta^{*,t}$ from Equation 12.

Recall the fact that $Q_r^{\pi} = \mathcal{T}_r(Q_r^{\pi})$.

Then, for any $t = 1, 2, ..., T$, we have,

$$\begin{aligned}
\|\hat{Q}_t - Q_r^{\pi}\|_{\infty} &= \|Clip(\Phi\hat{\theta}_t) - \mathcal{T}_r(Q_r^{\pi})\|_{\infty} \\
&= \underbrace{\|(\mathcal{T}_r(\hat{Q}_{t-1}) - \mathcal{T}_r(Q_r^{\pi}))\|_{\infty}}_{I} \\
&+ \underbrace{\|(Clip(\Phi\hat{\theta}_t) - Clip(\Pi\mathcal{T}_r(\hat{Q}_{t-1})))\|_{\infty}}_{II} \\
&+ \underbrace{\|(\mathcal{T}_r(\hat{Q}_{t-1}) - Clip(\Pi\mathcal{T}_r(\hat{Q}_{t-1})))\|_{\infty}}_{\leq \epsilon_{approx}}
\end{aligned}$$

First Term:

$$I = \|(\mathcal{T}_r(\hat{Q}_{t-1}) - \mathcal{T}_r(Q_r^{\pi}))\|_{\infty} \leq \gamma\|\hat{Q}_{t-1} - Q_r^{\pi}\|_{\infty}$$

as the robust bellman operator is a γ-contraction with respect to the ∞-norm (Iyengar, 2005).

Second Term:

$$\begin{aligned}
II &= \|(Clip(\Phi\hat{\theta}_t) - Clip(\Pi\mathcal{T}_r(\hat{Q}_{t-1})))\|_{\infty} \\
&\underbrace{\leq}_{(a)} \|\Phi\hat{\theta}_t - \Pi\mathcal{T}_r(\hat{Q}_{t-1})\|_{\infty} \\
&\underbrace{=}_{(b)} \||\Phi(\theta_{t-1,K} - \theta^{*,t-1})\|_{\infty} \\
&\leq \max_{s,a} \|\phi(s,a)\|_2 \|\theta_{t-1,K} - \theta^{*,t-1}\|_2 \\
&\underbrace{\leq}_{(c)} \|\theta_{t-1,K} - \theta^{*,t-1}\|_2
\end{aligned}$$

where (a) is using the non-expansive property of the clipping operator with respect to $\|\cdot\|_{\infty}$; for (b), recall the definition of $\theta^{*,t}$ in the inner loop in Equation 12; for (c), assume $\|\phi(s,a)\|_2 \leq 1, \forall(s,a)$.

Hence, we get:

$$\|\hat{Q}_t - Q_r^{\pi}\|_{\infty} \leq \gamma\|\hat{Q}_{t-1} - Q_r^{\pi}\|_{\infty} + \|\theta_{t-1,K} - \theta^{*,t-1}\|_2 + \epsilon_{approx}$$

Unroll the recursion and take the expectation:

$$\mathbb{E}\|\hat{Q}_T - Q_r^\pi\|_\infty \leq \gamma^T \|\hat{Q}_0 - Q_r^\pi\|_\infty + \sum_{t=0}^{T-1} \gamma^{T-t-1} \mathbb{E}[\|\theta_{t-1,K} - \theta^{*,t-1}\|_2] + \frac{\epsilon_{approx}}{1-\gamma}$$

C.2 Inner Loop Convergence Analysis

In this subsection, we show that for each outer iteration t, the inner loop parameter $\theta_{t,k}$ converges to $\theta^{*,t}$ where. Recall the definition $F^{*,V}_{s,a} \coloneqq \sup_{\lambda^a_s} F(\lambda^a_s; V, P_0(\cdot|s,a))$. We denote the inner loop error for outer iteration index t as $e^t_k := \theta_{t,k} - \theta^{*,t}$ with

$$\theta^{*,t} := (\Phi^\top D^\pi \Phi)^{-1} \Phi^\top D^\pi \left[r + \gamma F^{*,V_{\hat{\theta}_t}} \right] \tag{24}$$

Using earlier notation, the dual objective corresponding to an (s,a)-pair for a target value function $V_{\hat{\theta}_t}$ is

$$\min_{\lambda(s,a)} F(\lambda(s,a); V_{\hat{\theta}_t}, P_0(\cdot|s,a))$$

For the rest of the discussion in this subsection, let us fix an outer loop iteration t. For a given outer loop index t, for all inner loop iterations $k \geq 1$ let the filtration $\mathcal{F}^t_k$ be the sigma algebra generated by the transitions sampled till inner loop iteration index $k-1$. Formally, $\mathcal{F}^t_k = \sigma(S^t_j, A^t_j, S^t_{j+1} : 0 \leq j \leq k-1)$.

Observe that the pair process $Z^t_k \coloneqq (S^t_k, A^t_k)$ is a Markov chain. We define another filtration $\mathcal{G}^t_k = \sigma(Z^t_0, Z^t_1, ..., Z^t_k)$.

C.2.1 The Fast Time Scale:

Define the diagonal matrix $D_k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ with each diagonal element as $D_k((s,a),(s,a)) = \mathbf{1}_{(s,a)=(S_k,A_k)}$. Define the (s,a)-th component of the mean super-gradient vector as a function of the dual vector ν as

$$[\bar{g}(\nu)]_{s,a} = \mathbb{E}_{S' \sim P_0(\cdot|s,a)}[g(\psi(s,a)^\top \nu; S', V_{\hat{\theta}_t})]$$

Also, define the stochastic update vector $X_k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ defined as

$$[X^t_k] \coloneqq \mathbf{1}_{(s,a)=(S^t_k,A^t_k)} \cdot g(\psi(s,a)^\top \nu_k; S^t_{k+1}, V_{\hat{\theta}_t}) \psi(s,a)$$

The fast scale update is given as

$$\nu_{t,k+1} = Proj_{\mathcal{M}_\nu}(\nu_{t,k} + \beta_k [g(\psi(S^t_k, A^t_k)^\top \nu_k; S^t_{k+1}, V_{\hat{\theta}_t}) \psi(S^t_k, A^t_k)])$$

We split the update into stationary drift and different noise terms as:

$$\nu_{t,k+1} = Proj_{\mathcal{M}_\nu} \left(\nu_{t,k} + \beta_k \left[\Psi^\top D^\pi \bar{g}(\nu_k) + \underbrace{X^t_k - \mathbb{E}[X^t_k|\mathcal{G}^t_k]}_{m^\nu_{k+1}} + \underbrace{\mathbb{E}[X^t_k|\mathcal{G}^t_k] - \Psi^\top D^\pi \bar{g}(\nu_k)}_{\zeta^\nu_{k+1}} \right] \right)$$

We see that,

$$\zeta^\nu_{k+1} = \mathbb{E}[X^t_k|\mathcal{G}^t_k] - \Psi^\top D^\pi \bar{g}(\nu_k)$$

So the update now becomes,

$$\nu_{t,k+1} = Proj_{\mathcal{M}} \left(\nu_{t,k} + \beta_k \left[\Psi^\top D^\pi \bar{g}(\nu_k) + m^\nu_{k+1} + \zeta^\nu_{k+1} \right] \right)$$

In the above equation, m^λ_{k+1} denotes the state-innovation noise that is a martingale difference on the filtration $\mathcal{G}^t_k$.

Hence,

$$\mathbb{E}[m^\nu_{k+1}|\mathcal{G}^t_k] = 0$$

We analyze the finite time convergence of the fast time scale first. Let ν^* be one solution of the dual optimization problem in the domain $\mathcal{M}$. As the slow time scale parameter does not influence the fast time scale, we can analyse the fast time scale convergence independently.

Let $d_{(s,a)} = D^\pi((s,a),(s,a))$.

Our goal is to bound the sub-optimality gap of the dual objective for each iteration in the inner loop. We will be able to use the error in estimating the dual objective from the fast time scale as a bias in the slow time scale to get a sample complexity bound for the inner loop of the algorithm 1.

Let us define the dual objective sub-optimality for the fast time scale as :

$$L_k \coloneqq \sum_{s,a} d_{(s,a)}[F(\psi(s,a)^\top \nu^*; V_{\hat{\theta}_t}, P_0(\cdot|s,a)) - F(\psi(s,a)^\top \nu_{t,k}; V_{\hat{\theta}_t}, P_0(\cdot|s,a))]$$

Define the Lyapunov function for the fast time scale as

$$e^\nu_{t,k} = \|\nu_{t,k} - \nu^*\|_2^2.$$

Using the non-expansiveness of projection,

$$e^\nu_{t,k+1} \leq \|\nu_{t,k} + \beta_k \Psi^\top D^\pi \bar{g}(\nu_{t,k}) + \beta_k(m^\nu_{k+1} + \zeta^\nu_{k+1}) - \nu^*\|^2_{2,D^\pi}.$$

We can write:

$$\begin{aligned} e^\nu_{t,k+1} &\leq \|(\nu_{t,k} - \nu^*)\|_2^2 + 2\beta_k(\nu_{t,k} - \nu^*)^\top \Psi^\top D^\pi \bar{g}(\nu_{t,k}) \\ &+ 2\beta_k((\nu_{t,k} - \nu^*))^\top (m^\nu_{k+1} + \zeta^\nu_{k+1}) \\ &+ \beta_k^2 \|\Psi^\top D^\pi \bar{g}(\nu_{t,k}) + m^\nu_{k+1} + \zeta^\nu_{k+1}\|_2^2 \end{aligned}$$

We simplify the term $(\nu_{t,k} - \nu^*)^\top \Psi^\top D^\pi \bar{g}(\nu_{t,k})$ first.

$$(\nu_{t,k} - \nu^*)^\top \Psi^\top D^\pi \bar{g}(\nu_{t,k}) = \sum_{s,a \in \mathcal{S}\times\mathcal{A}} d_{(s,a)}(\nu_{t,k}(s,a) - \nu^*(s,a))\bar{g}(\nu_{t,k})(s,a)\psi(s,a)$$

Now as the $L^t(\nu)$ is a concave function ,

$$(\nu_{t,k} - \nu^*)^\top \Psi^\top (D^\pi)\bar{g}(\nu_{t,k}) \leq -L_k$$

Hence, we get,

$$\begin{aligned} e^\nu_{t,k+1} \leq \|(\nu_{t,k} - \nu^*)\|_2^2 &- 2\beta_k L_k \\ &+ 2\beta_k((\nu_{t,k} - \nu^*))^\top (m^\nu_{k+1}) + 2\beta_k((\nu_{t,k} - \nu^*))^\top (\zeta^\nu_{k+1}) \\ &+ \beta_k^2 \|\Psi^\top D^\pi \bar{g}(\nu_k) + m^\nu_{k+1} + \zeta^\nu_{k+1}\|_2^2 \end{aligned}$$

Now we condition on a lagged filtration $\mathcal{G}^t_{k-\tau}$ where τ would be chosen later.

$$\begin{aligned} \mathbb{E}\left[e^\nu_{t,k+1}|\mathcal{G}^t_{k-\tau}\right] \leq \mathbb{E}\left[e^\nu_{t,k}|\mathcal{G}^t_{k-\tau}\right] &- 2\beta_k \mathbb{E}\left[L_k|\mathcal{G}^t_{k-\tau}\right] \\ &+ 2\beta_k \mathbb{E}\left[(\nu_{t,k} - \nu^*)^\top (m^\nu_{k+1})|\mathcal{G}^t_{k-\tau}\right] + 2\beta_k \mathbb{E}\left[(\nu_{t,k} - \nu^*)^\top (\zeta^\nu_{k+1})|\mathcal{G}^t_{k-\tau}\right] \\ &+ \beta_k^2 \mathbb{E}\left[\|\Psi^\top D^\pi \bar{g}(\nu_{t,k}) + m^\nu_{k+1} + \zeta^\nu_{k+1}\|_2^2|\mathcal{G}^t_{k-\tau}\right] \end{aligned}$$

Let us first bound the β_k^2 terms. Recall from Assumption 1, $\|\bar{g}(\nu_k)\|_\infty \leq g_M$. We can show that $\|\Psi^\top D^\pi \bar{g}(\nu_k)\|_2 \leq g_M$ and $\|m^\nu_{k+1}\|_2 \leq 2g_M$ and $\|\zeta^\nu_{k+1}\|_2^2\|_2 \leq 2g_M$. Hence,

$$\beta_k^2 \mathbb{E}\left[\|\Psi^\top D^\pi \bar{g}(\nu_{t,k}) + m^\nu_{k+1} + \zeta^\nu_{k+1}\|_2^2|\mathcal{G}^t_{k-\tau}\right] \leq 25\beta_k^2 g_M^2$$

Now we work on the cross terms. Let us start with $2\beta_k \mathbb{E}\left[(\nu_{t,k} - \nu^*)^\top D^\pi (m^\nu_{k+1})|\mathcal{G}^t_{k-\tau}\right]$

We write:

$$\begin{aligned}2\beta_k\mathbb{E}\left[(\nu_{t,k}-\nu^*)^\top D^\pi(m^\nu_{k+1})|\mathcal{G}^t_{k-\tau}\right] &= 2\beta_k\mathbb{E}\left[\mathbb{E}\left[(\nu_{t,k}-\nu^*)^\top D^\pi(m^\nu_{k+1})|\mathcal{G}^t_k\right]|\mathcal{G}^t_{k-\tau}\right]\\ &= 2\beta_k\mathbb{E}\left(\nu_{t,k}-\nu^*\right)^\top D^\pi[\mathbb{E}\left[(m^\nu_{k+1})|\mathcal{G}^t_k\right]|\mathcal{G}^t_{k-\tau}]\\ &= 0.\end{aligned}$$

Now we focus on the term $2\beta_k\mathbb{E}\left[(\nu_{t,k}-\nu^*)^\top(\zeta^\nu_{k+1})|\mathcal{G}^t_{k-\tau}\right]$.

Define the vector e_{Z_k}

$$e_{Z_k}(s,a) \coloneqq \mathbf{1}_{S_k,A_k=s,a}$$

Let $\mu_k(\cdot) := \mathbb{P}(Z_k \in \cdot \mid \mathcal{G}_{k-\tau})$ and assume the π–chain is geometrically β–mixing:

$$\|\mu_k - d^\pi\|_{\mathrm{TV}} \;\le\; \beta(\tau) \coloneqq C_{\mathrm{mix}}\,\rho^\tau \qquad (0<\rho<1).$$

We can write:

$$\begin{aligned}&2\beta_k\mathbb{E}\left[(\nu_{t,k}-\nu^*)^\top(\Psi^\top(e_{Z_k}-d^\pi)\odot\bar{g}(\nu_{t,k}))|\mathcal{G}^t_{k-\tau}\right]\\ &= 2\beta_k\mathbb{E}\left[(\Psi(\nu_{t,k}-\nu^*))^\top((e_{Z_k}-d^\pi)\odot\bar{g}(\nu_{t,k}))|\mathcal{G}^t_{k-\tau}\right]\\ &\le 2\beta_k\mathbb{E}\left[\|(\Psi(\nu_{t,k}-\nu^*))\|_\infty\|((e_{Z_k}-d^\pi)\odot\bar{g}(\nu_{t,k}))\|_1|\mathcal{G}^t_{k-\tau}\right]\\ &\le 2\beta_k\mathbb{E}\left[2B_\nu g_M\|(e_{Z_k}-d^\pi)\|_1|\mathcal{G}^t_{k-\tau}\right]\\ &\le 8\beta_k B_\nu g_M\beta(\tau)\end{aligned}$$

Putting it together, we have:

$$\mathbb{E}\left[e^\nu_{t,k+1}|\mathcal{G}^t_{k-\tau}\right] \le \mathbb{E}\left[e^\nu_{t,k}|\mathcal{G}^t_{k-\tau}\right] - 2\beta_k\mathbb{E}\left[L_k|\mathcal{G}^t_{k-\tau}\right] + 8\beta_k B_\nu g_M\beta(\tau) + 25\beta_k^2 g_m^2$$

If we choose $\tau = \lceil\frac{log(k+1)}{2log\left(\frac{1}{\rho}\right)}\rceil$, then, we have,

$$\mathbb{E}\left[e^\nu_{t,k+1}|\mathcal{G}^t_{k-\tau}\right] \le \mathbb{E}\left[e^\nu_{t,k}|\mathcal{G}^t_{k-\tau}\right] - 2\beta_k\mathbb{E}\left[L_k|\mathcal{G}^t_{k-\tau}\right] + (8B_\nu g_M C_{mix}\frac{1}{\beta_0} + 25g_M^2)\beta_k^2$$

Changing the conditional expectation to a filtration $\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}$, we write $\forall l \in (\lfloor k/2\rfloor), k-1$:

$$2\beta_k\mathbb{E}[L_l|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] \le \mathbb{E}[e^\nu_{t,l}|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] - \mathbb{E}[e^\nu_{t,l+1}|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] + \beta_k^2(8B_\nu g_M C_{mix}\frac{1}{\beta_0} + 25g_M^2)$$

Next, we use telescoping for iterates over the index l from $\lfloor\frac{k}{2}\rfloor$ to $k-1$.

$$2\sum_{l=\lfloor k/2\rfloor}^{k-1}\beta_l\mathbb{E}[L_l|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] \le e^\nu_{t,\lfloor k/2\rfloor} + (8B_\nu g_M C_{mix}\frac{1}{\beta_0} + 25g_M^2)\sum_{l=\lfloor k/2\rfloor}^{k-1}\beta_l^2 \tag{25}$$

Recall, the fast time scale passes the following dual variable to the slow time scale at each iterate k:

$$\bar{\nu}_{t,k} = \frac{1}{\lceil k/2\rceil}\sum_{\lfloor k/2\rfloor}^{k-1}\nu_{t,k}$$

We use the step size rule of $\beta_k = \frac{\beta_0}{\sqrt{k+1}}$.

Because of the clipping of each iterate, we know that $e^\nu_{t,\lfloor k/2\rfloor} \le 4B_\nu^2$ for any $k>0$.

Similar to the definition of L_k, let us define

$$\overline{L}_k \coloneqq \sum_{s,a} d_{(s,a)}[F(\psi(s,a)^\top \nu^*; V_{\hat{\theta}_t}, P_0(\cdot|s,a)) - F(\psi(s,a)^\top \bar{\nu}_{t,k}; V_{\hat{\theta}_t}, P_0(\cdot|s,a))]$$

Hence,

$$\begin{aligned}
\mathbb{E}[\overline{L}_k|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] &\leq \frac{1}{\lceil k/2\rceil}\sum_{l=\lfloor k/2\rfloor}^{k-1}\mathbb{E}[L_l|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] \\
&\underbrace{\leq}_{(a)} \frac{1}{\lceil k/2\rceil}\frac{\sqrt{k}}{\beta_0}\sum_{l=\lfloor k/2\rfloor}^{k-1}\beta_l\mathbb{E}[L_k|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] \leq \frac{2}{k}\frac{\sqrt{k}}{\beta_0}\sum_{\lfloor k/2\rfloor}^{k-1}\beta_l\mathbb{E}[L_k|\mathcal{G}^t_{\lfloor k/2\rfloor-\tau}] \\
&\underbrace{\leq}_{(b)} \frac{1}{\beta_0\sqrt{k}}e^{\nu}_{t,\lfloor k/2\rfloor} + \frac{(8B_\nu g_M C_{mix}\frac{1}{\beta_0}+25g_M^2)}{\beta_0\sqrt{k}}\sum_{l=\lfloor k/2\rfloor}^{k-1}\beta_l^2 \\
&\leq \frac{4B_\nu^2}{\beta_0\sqrt{k}} + \frac{(8B_\nu g_M C_{mix}\frac{1}{\beta_0}+25g_M^2)}{\beta_0\sqrt{k}}\sum_{l=\lfloor k/2\rfloor}^{k-1}\frac{\beta_0^2}{l+1} \\
&\underbrace{\leq}_{(c)} \frac{4B_\nu^2}{\beta_0\sqrt{k}} + \frac{(8B_\nu g_M C_{mix}\frac{1}{\beta_0}+25g_M^2)\left(\beta_0^2(\ln(k)-\ln(k/2))\right)}{\beta_0\sqrt{k}} \\
&\leq \frac{C_{fast}}{\sqrt{k}}
\end{aligned}$$

where

$$C_{fast} = \frac{(4B_\nu^2 + \beta_0^2(8B_\nu g_M C_{mix}\frac{1}{\beta_0}+25g_M^2)\ln(2))}{\beta_0} \tag{26}$$

. In (a), we used $\beta_k \geq \frac{\beta_0}{\sqrt{k}}, \forall k \leq k-1$. In (b), we used Equation 25. In (c), we used the following identity: $\ln(k) \leq H_k \leq 1+\ln(k)$ where H_k is the harmonic series up to an integer k.

C.2.2 Slow Time Scale Analysis

We denote the inner loop error for outer iteration index t as $e_k^t \coloneqq \theta_{t,k} - \theta^{*,t}$ with

$$\theta^{*,t} \coloneqq (\Phi^\top D^\pi \Phi)^{-1}\Phi^\top D^\pi\left[r + \gamma F^{*,V_{\hat{\theta}_t}}\right] \tag{27}$$

With a slight abuse of notation, we drop the superscript t from the variables for the remainder of this subsection, since the outer loop index t is fixed. We will make the dependence on t explicit whenever it is essential.

The slow update (no projection) is

$\theta_{k+1} = \theta_k + \alpha_k\,\delta_{k+1}\,\phi(Z_k), \quad \delta_{k+1} = r(Z_k) + \gamma\,\sigma(\psi(Z_k)^\top\bar{\nu}_k(Z_k); S_{k+1}, V_{\hat{\theta}}) - \phi(Z_k)^\top\theta_k,$
with $\bar{\nu}_k$ the suffix average produced by the fast scale.

Let us define the (s,a)-th component of the mean dual objective estimator for any ν as:

$$[\bar{\sigma}(\nu)]_{s,a} \coloneqq \mathbb{E}_{S'\sim P_0(\cdot|s,a)}[\sigma(\psi(s,a)^\top\nu, S', V_{\hat{\theta}_t})]$$

We decompose the sampled direction as

$$G(\theta_k) \;+\; b_k^\theta \;+\; \xi_{k+1}^\theta \;+\; m_{k+1}^\theta,$$

where

$$\begin{aligned}
G(\theta_k) &\coloneqq \Phi^\top D^\pi\left[r + \gamma F^{*,V_{\hat{\theta}_t}} - \Phi\theta_k\right] \underbrace{=}_{\text{from Equation (12)}} \Phi^\top D^\pi\Phi(\theta^* - \theta_k), \\
b_k^\theta &\coloneqq \Phi^\top D^\pi\left[F^{V_{\hat{\theta}_t}}(\bar{\nu}_k) - F^{*,V_{\hat{\theta}_t}}\right], \\
\xi_{k+1}^\theta &\coloneqq \Phi^\top\left(e_{Z_k} - D^\pi\right)\Big[r + \gamma\bar{\sigma}(\bar{\nu}_k) - \Phi\theta_k\Big], \\
m_{k+1}^\theta &\coloneqq \gamma\,\Phi^\top e_{Z_k}\Big(\sigma(\bar{\nu}_k(Z_k); S_{k+1}, \cdot) - \bar{\sigma}(\bar{\nu}_k)(Z_k)\Big).
\end{aligned}$$

Note $\mathbb{E}[m^\theta_{k+1} \mid \mathcal{G}_k] = 0$ (innovation MDS) and, by tower, $\mathbb{E}[e_k^\top m^\theta_{k+1} \mid \mathcal{G}_{k-\tau}] = 0$.

Notation and standing constants. We assume $\|\phi(s,a)\|_2 \le L_\phi$, $\|\Phi\|_{op} \le L_\Phi$, rewards $r \in [0,1]$, and $|\bar\sigma(\cdot)| \le \sigma_M$. Further set

$$B := \frac{1}{1-\gamma}, \qquad Y_0 := 1 + \gamma\sigma_M + L_\phi\|\theta^*\|, \qquad Y_1 := L_\phi, \qquad L_F := \|\Phi^\top D^\pi \Phi\|_{op}.$$

We know that the π–chain is geometrically β–mixing:

$$\beta(h) \;\le\; C_{mix}\,\rho^h, \qquad 0<\rho<1, \qquad \sum_{h\ge 1}\beta(h) \le \frac{C_{mix}\rho}{1-\rho}.$$

Finally, from the fast time scale we will use (proved in the fast-scale subsection)

$$\mathbb{E}\big[\overline{L}_k \,\big|\, \mathcal{G}_{k_0}\big] \;\le\; \frac{C_{fast}}{\sqrt{k}}, \tag{28}$$

with $k_0 = \lfloor k/2 \rfloor - \tau$ and the explicit constant

$$C_{fast} = \frac{(4B_\nu^2 + \beta_0^2(8B_\nu g_M C_{mix}\frac{1}{\beta_0} + 25 g_M^2)\ln(2))}{\beta_0}.$$

(I) ONE-STEP LYAPUNOV EXPANSION UNDER LAG

With $x_k := \|e_k\|^2$,

$$\begin{aligned}\mathbb{E}[x_{k+1} \,|\, \mathcal{G}_{k-\tau}] &= \mathbb{E}\Big[\big\|e_k + \alpha_k(G(\theta_k) + b_k^\theta + \xi^\theta_{k+1} + m^\theta_{k+1})\big\|^2 \,\Big|\, \mathcal{G}_{k-\tau}\Big] \\ &= x_k + 2\alpha_k\,\mathbb{E}\big[e_k^\top G(\theta_k) \,\big|\, \mathcal{G}_{k-\tau}\big] + 2\alpha_k\,\mathbb{E}\big[e_k^\top b_k^\theta \,\big|\, \mathcal{G}_{k-\tau}\big] \\ &\quad + 2\alpha_k\,\mathbb{E}\big[e_k^\top \xi^\theta_{k+1} \,\big|\, \mathcal{G}_{k-\tau}\big] + \alpha_k^2\,\mathbb{E}\big[\|G(\theta_k) + b_k^\theta + \xi^\theta_{k+1} + m^\theta_{k+1}\|^2 \,\big|\, \mathcal{G}_{k-\tau}\big].\end{aligned} \tag{29}$$

(II) MAIN DRIFT

Since $e_k^\top G(\theta_k) = -e_k^\top(\Phi^\top D^\pi \Phi)e_k \le -\mu\,\|e_k\|^2$,

$$2\alpha_k\,\mathbb{E}\big[e_k^\top G(\theta_k) \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; -2\mu\,\alpha_k\,\mathbb{E}[x_k \,|\, \mathcal{G}_{k-\tau}]. \tag{30}$$

Further Notations. Let $Z_k = (S_k, A_k)$ and let $\mathcal{G}_k = \sigma(Z_0, \dots, Z_k)$. Fix a lag $\tau \ge 1$ and define

$$\mathcal{H}_k \;:=\; \sigma\big(\mathcal{G}_{k-\tau},\, \theta_k,\, \bar\lambda_k\big).$$

Conditioning on $\mathcal{H}_k$ "freezes" $e_k := \theta_k - \theta^*$ and $y_k := r + \gamma\,\bar\sigma(\bar\nu_k) - \Phi\theta_k$; only Z_k remains random. Let $\mu_k(\cdot) := \mathbb{P}(Z_k \in \cdot \mid \mathcal{G}_{k-\tau})$. We know that the π–chain is geometrically β–mixing:

$$\|\mu_k - d^\pi\|_{\mathrm{TV}} \;\le\; C_{\mathrm{mix}}\,\rho^\tau \qquad (0<\rho<1).$$

Let the notation z denote an arbitrary (s,a)-pair. We use: for any signed vector w on $\mathcal{S}\times\mathcal{A}$,

$$\big\|\Phi^\top w\big\|_2 \;=\; \Big\|\sum_z w_z\,\phi(z)\Big\|_2 \;\le\; \sum_z |w_z| \;=\; \|w\|_1, \tag{31}$$

because each row vector satisfies $\|\phi(z)\|_2 \le 1$. We also write the Markov mismatch ("mixing noise") as

$$\xi^\theta_{k+1} \;=\; \Phi^\top\big(e_{Z_k} - D^\pi\big)\,y_k.$$

Finally set $Y_0 := 1 + \gamma\sigma_M + \|\theta^*\|_2$ and $Y_1 := 1$ so that

$$\|y_k\|_\infty \;\le\; Y_0 + Y_1\,\|e_k\|_2, \tag{32}$$

using $r \in [0,1]$, $\|\bar\sigma(\cdot)\|_\infty \le \sigma_M$, and $\|\phi(z)\|_2 \le 1$.

(III) CROSS WITH FAST-BIAS

By conditional Cauchy–Schwarz and Young inequality,

$$2\alpha_k \,\mathbb{E}\big[e_k^\top b_k^\theta \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; \mu\,\alpha_k\,\mathbb{E}[x_k \,|\, \mathcal{G}_{k-\tau}] \;+\; \frac{\alpha_k}{\mu}\,\mathbb{E}\big[\|b_k^\theta\|^2 \,\big|\, \mathcal{G}_{k-\tau}\big]. \tag{33}$$

Lemma 3 (Cross with Markov mismatch) *For any $\tau \ge 1$,*

$$2\alpha_k \,\mathbb{E}\big[e_k^\top \xi_{k+1}^\theta \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; \Big(\frac{\mu}{2} + 4Y_1\beta(\tau)\Big)\,\alpha_k\,\mathbb{E}\left[\|e_k\|_2^2 |\,\mathcal{G}_{k-\tau}\right] + \frac{8\,Y_0^2}{\mu}\,\alpha_k\,\beta(\tau)^2,$$

Here $\beta(\tau) := \|\mu_k - d^\pi\|_{\mathrm{TV}}$ and $\mu > 0$ is the minimum eigenvalue of $\Phi^\top D^\pi \Phi$.

Proof 2 *By the tower property,*

$$\mathbb{E}\big[e_k^\top \xi_{k+1}^\theta \,\big|\, \mathcal{G}_{k-\tau}\big] = \mathbb{E}\big[\,\mathbb{E}\big[e_k^\top \Phi^\top (e_{Z_k} - D^\pi)\, y_k \,\big|\, \mathcal{H}_k\big] \,\big|\, \mathcal{G}_{k-\tau}\big].$$

Given $\mathcal{H}_k$, the only randomness is $Z_k \sim \mu_k$. Hence

$$\mathbb{E}\big[\Phi^\top (e_{Z_k} - D^\pi)\, y_k \,\big|\, \mathcal{H}_k\big] = \Phi^\top(\mu_k - D^\pi)\, y_k.$$

Therefore,

$$\Big|\,\mathbb{E}\big[e_k^\top \xi_{k+1}^\theta \,\big|\, \mathcal{G}_{k-\tau}\big]\,\Big| \le \mathbb{E}\big[\,\|e_k\|_2 \,\big\|\Phi^\top(\mu_k - D^\pi)\, y_k\big\|_2 \,\big|\, \mathcal{G}_{k-\tau}\big].$$

Apply Equation (31) and $\|(\mu_k - D^\pi)\, y_k\|_1 \le \|\mu_k - D^\pi\|_1 \|y_k\|_\infty = 2\,\beta(\tau)\,\|y_k\|_\infty$:

$$\Big|\,\mathbb{E}[e_k^\top \xi_{k+1}^\theta \mid \mathcal{G}_{k-\tau}]\Big| \le 2\,\beta(\tau)\,\mathbb{E}[\|e_k\|_2\,\|y_k\|_\infty \mid \mathcal{G}_{k-\tau}].$$

Multiply by $2\alpha_k$ and split $\|y_k\|_\infty$ using Equation (32):

$$2\alpha_k|\,\mathbb{E}[\cdot]\,| \le 4\alpha_k\beta(\tau)\,\mathbb{E}[\|e_k\|_2\,Y_0 \mid \mathcal{G}_{k-\tau}] + 4\alpha_k\beta(\tau)\,\mathbb{E}[\|e_k\|_2\,Y_1\|e_k\|_2 \mid \mathcal{G}_{k-\tau}].$$

For the Y_0 term use Young's inequality with parameter $\eta = \mu/2$:

$$4\alpha_k\beta(\tau)\,Y_0\,\|e_k\|_2 \le \alpha_k\left(\frac{\mu}{4}\|e_k\|_2^2 + \frac{16\,\beta(\tau)^2\,Y_0^2}{\mu}\right).$$

Combining,

$$2\alpha_k \,\mathbb{E}\big[e_k^\top \xi_{k+1}^\theta \,\big|\, \mathcal{G}_{k-\tau}\big] \le \Big(\frac{\mu}{2} + 4Y_1\beta(\tau)\Big)\,\alpha_k\,\mathbb{E}\left[\|e_k\|_2^2 |\,\mathcal{G}_{k-\tau}\right] + \frac{8\,Y_0^2}{\mu}\,\alpha_k\,\beta(\tau)^2.$$

Lemma 4 (Second moment of the Markov mismatch) *For any $\tau \ge 1$,*

$$\mathbb{E}\big[\,\|\xi_{k+1}^\theta\|_2^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; 16\,(1+\beta(\tau)^2)\,\Big(Y_0^2 + Y_1^2\,\mathbb{E}\left[\|e_k\|_2^2 |\,\mathcal{G}_{k-\tau}\right]\Big).$$

Proof 3 *Decompose at the lag:*

$$\xi_{k+1}^\theta = \underbrace{\Phi^\top\big(e_{Z_k} - \mu_k\big)\, y_k}_{\xi_{k+1}^{(0)}} \;+\; \underbrace{\Phi^\top\big(\mu_k - D^\pi\big)\, y_k}_{\xi_{k+1}^{(b)}}.$$

Then

$$\|\xi_{k+1}^\theta\|_2^2 \le 2\|\xi_{k+1}^{(0)}\|_2^2 + 2\|\xi_{k+1}^{(b)}\|_2^2.$$

Centered part. *By Equation (31),*

$$\|\xi_{k+1}^{(0)}\|_2 \le \big\|(e_{Z_k} - \mu_k)\, y_k\big\|_1.$$

Using $(a+b)^2 \le 2a^2 + 2b^2$ and conditioning on $\mathcal{H}_k$,

$$\mathbb{E}\big[\|(e_{Z_k} - \mu_k)\, y_k\|_1^2 \,\big|\, \mathcal{H}_k\big] \le 2\,\mathbb{E}\big[|y_k(Z_k)|^2 \mid \mathcal{H}_k\big] + 2\,\|\mu_k y_k\|_1^2 \le 4\,\|y_k\|_\infty^2.$$

Hence

$$\mathbb{E}\Big[\|\xi_{k+1}^{(0)}\|_2^2 \,\Big|\, \mathcal{G}_{k-\tau}\Big] \le 4\,\mathbb{E}\|y_k\|_\infty^2.$$

Bias part. *Again by equation 31,*

$$\|\xi_{k+1}^{(b)}\|_2 \le \big\|(\mu_k - D^\pi)\,y_k\big\|_1 \le 2\,\beta(\tau)\,\|y_k\|_\infty,$$

so

$$\mathbb{E}\Big[\|\xi_{k+1}^{(b)}\|_2^2 \,\Big|\, \mathcal{G}_{k-\tau}\Big] \le 4\,\beta(\tau)^2\,\mathbb{E}\|y_k\|_\infty^2.$$

Combine and unfreeze y_k. *Therefore,*

$$\mathbb{E}\big[\|\xi_{k+1}^{\theta}\|_2^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \le 2\cdot 4\,\mathbb{E}\|y_k\|_\infty^2 + 2\cdot 4\,\beta(\tau)^2\,\mathbb{E}\|y_k\|_\infty^2 = 8\,(1+\beta(\tau)^2)\,\mathbb{E}\|y_k\|_\infty^2.$$

Use $(a+b)^2 \le 2a^2 + 2b^2$ *in equation 32 to obtain*

$$\mathbb{E}\|y_k\|_\infty^2 \le 2Y_0^2 + 2Y_1^2\,\mathbb{E}\|e_k\|_2^2,$$

hence the stated bound with the factor $16(1+\beta(\tau)^2)$.

(v) QUADRATIC BLOCK

We use

$$\|G(\theta_k)\|^2 \le L_F^2\,x_k, \qquad \mathbb{E}\big[\|m_{k+1}^{\theta}\|^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \le C_m := 4\gamma^2 L_\Phi^2 \sigma_M^2.$$

For ξ_{k+1}^{θ}, a crude (mixing-free) bound yields

$$\mathbb{E}\big[\|\xi_{k+1}^{\theta}\|^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; C_{\xi,0} \;+\; C_{\xi,1}\,\mathbb{E}[x_k \,|\, \mathcal{G}_{k-\tau}], \quad C_{\xi,0} := 8L_\Phi^2 Y_0^2,\;\; C_{\xi,1} := 8L_\Phi^2 Y_1^2.$$

Therefore,

$$\alpha_k^2\,\mathbb{E}\big[\|G(\theta_k) + b_k^{\theta} + \xi_{k+1}^{\theta} + m_{k+1}^{\theta}\|^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; \alpha_k^2\Big(C_e\,\mathbb{E}[x_k \mid \mathcal{G}_{k-\tau}] + C_0 + 2\,\mathbb{E}[\|b_k^{\theta}\|^2 \mid \mathcal{G}_{k-\tau}]\Big), \tag{34}$$

with

$$C_e := L_F^2 + C_{\xi,1}, \qquad C_0 := C_{\xi,0} + C_m.$$

Lemma 5 (Bias second order at $1/k$) *Let* $b_k^{\theta} := F(\theta_k, \bar{\nu}_k) - F(\theta_k, \nu^*)$. *Then*

$$\mathbb{E}\big[\|b_k^{\theta}\|_2^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; \frac{2C_{\text{bias}}}{k} + 2(\epsilon_{approx}^{dual})^2,$$

with the explicit constant

$$C_{\text{bias}} \;:=\; C_{\text{fast}}^2 \;+\; \frac{B^2}{2} \;+\; 16\,B^2\,\frac{C_{\text{mix}}\rho}{1-\rho}, \qquad B := \frac{1}{1-\gamma},$$

and C_{fast} *as in equation 28.*

Proof 4 *Using* $\|\phi(s,a)\|_2 \le 1$ *for all* (s,a), *we can write*

$$\begin{aligned}
\|b_{t,k}^{\theta}\|_2 &\le \gamma \sum_{s,a} d^\pi(s,a)\left|F^{V_{\hat{\theta}_t}}(\bar{\nu}_{t,k})_{s,a} - F_{s,a}^{*,V_{\hat{\theta}_t}}\right| \\
&= \gamma \sum_{s,a} d^\pi(s,a)\left(F_{s,a}^{*,V_{\hat{\theta}_t}} - F^{V_{\hat{\theta}_t}}(\bar{\nu}_{t,k})_{s,a}\right) \\
&\le \gamma \underbrace{\inf_{\nu\in M_\nu} \sum_{s,a} d^\pi(s,a)\left(F_{s,a}^{*,V_{\hat{\theta}_t}} - F^{V_{\hat{\theta}_t}}(\nu)_{s,a}\right)}_{\le \epsilon_{\text{approx}}^{\text{dual}}} + \gamma \underbrace{\bar{L}_k}_{\textit{fast-scale objective gap}}.
\end{aligned}$$

$$\mathbb{E}\big[\|b_k^{\theta}\|_2^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; 2\mathbb{E}\big[(\overline{L}_k)^2 \,\big|\, \mathcal{G}_{k-\tau}\big] + 2(\epsilon_{approx}^{dual})^2.$$

To bound the RHS at the $1/k$ scale, use the suffix average $\overline{L}_k \le \frac{1}{m}\sum_{j=k-m}^{k-1} L_j$ with $m = \lceil k/2 \rceil$ and $0 \le L_j \le B$. Write

$$\mathbb{E}\big[(\overline{L}_k)^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \le \frac{1}{m^2}\Big\{ \underbrace{\mathrm{Var}\left(\sum_{j=k-m}^{k-1} L_j \,\middle|\, \mathcal{G}_{k-\tau} \right)}_{\text{(I)}} + \underbrace{\Big(\mathbb{E} \sum_{j=k-m}^{k-1} L_j \mathcal{G}_{k-\tau} \Big)^2}_{\text{(II)}} \Big\}.$$

Term (II) (mean square). *By equation 28, $\frac{1}{m}\,\mathbb{E}\sum_{j=k-m}^{k-1} L_j \le \mathbb{E}[\overline{L}_k] \le C_{\mathrm{fast}}/\sqrt{k}$, so $\text{(II)}/m^2 \le C_{\mathrm{fast}}^2/k$.*

Term (I) (variance). *Under geometric β–mixing and $0 \le L_j \le B$, the conditional covariances obey $|\mathrm{Cov}(L_j, L_{j+h} \mid \mathcal{G}_{k-\tau})| \le 4B^2\,\beta(h) \le 4B^2 C_{\mathrm{mix}}\rho^h$. Thus*

$$\begin{aligned}\mathrm{Var}\left(\sum_{j=k-m}^{k-1} L_j \,\middle|\, \mathcal{G}_{k-\tau} \right) &= \sum_j \mathrm{Var}(L_j \mid \cdot) \;+\; 2\sum_{h=1}^{m-1} (m-h)\,\mathrm{Cov}(L_j, L_{j+h} \mid \cdot) \\ &\le m\,\frac{B^2}{4} \;+\; 8B^2\, m \sum_{h=1}^{\infty} \beta(h) \;\le\; m\Big(\frac{B^2}{4} + 8B^2\,\frac{C_{\mathrm{mix}}\rho}{1-\rho}\Big).\end{aligned}$$

Dividing by m^2 and using $m \ge k/2$ gives

$$\frac{\text{(I)}}{m^2} \;\le\; \frac{2}{k}\Big(\frac{B^2}{4} + 8B^2\,\frac{C_{\mathrm{mix}}\rho}{1-\rho}\Big).$$

Combining the two terms yields

$$\mathbb{E}\big[(\overline{L}_k)^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; \frac{C_{\mathrm{fast}}^2}{k} \;+\; \frac{2}{k}\Big(\frac{B^2}{4} + 8B^2\,\frac{C_{\mathrm{mix}}\rho}{1-\rho}\Big),$$

which is the claimed bound with the displayed C_{bias}.

FINAL RECURSION FOR THE SLOW TIME SCALE (MARKOV SAMPLING, LAGGED)

Let $e_k := \theta_k - \theta^*$ and $x_k := \mathbb{E}\|e_k\|_2^2$. Fix a lag $\tau \ge 1$ and condition on $\mathcal{G}_{k-\tau}$. Combining the drift, the two cross terms, and the quadratic block (with the second-moment bounds proved above) yields

$$\mathbb{E}\big[\|e_{k+1}\|_2^2 \,\big|\, \mathcal{G}_{k-\tau}\big] \;\le\; \Big(1 - \tfrac{\mu}{2}\,\alpha_k + C_e\,\alpha_k^2\Big)\,\mathbb{E}\big[\|e_k\|_2^2 \mathcal{G}_{k-\tau}\big] \tag{35}$$

$$+\; \Big(\tfrac{\alpha_k}{\mu} + 2\alpha_k^2\Big)\left(\frac{2C_{bias}}{k} + 2(\epsilon_{approx}^{dual})^2\right) \;+\; C_{cross}\,\alpha_k\,\beta(\tau)^2 \;+\; C_0\,\alpha_k^2, \tag{36}$$

and hence, after taking total expectation,

$$\mathbb{E}\|e_{k+1}\|_2^2 \;\le\; \Big(1 - \tfrac{\mu}{2}\,\alpha_k + C_e\,\alpha_k^2\Big)\,\mathbb{E}\|e_k\|_2^2 + \tag{37}$$

$$+\; \Big(\tfrac{\alpha_k}{\mu} + 2\alpha_k^2\Big)\left(\frac{2C_{bias}}{k} + 2(\epsilon_{approx}^{dual})^2\right) \;+\; C_{cross}\,\alpha_k\,\beta(\tau)^2 \;+\; C_0\,\alpha_k^2. \tag{38}$$

Constants (explicit). We use that D^π is a probability diagonal ($\sum_z d_z^\pi = 1$) and $\|\phi(z)\|_2 \le 1$, which implies $\|\Phi^\top D^\pi \Phi v\|_2^2 \le \|v\|_2^2$ and thus we can take

$$L_F := \|\Phi^\top D^\pi \Phi\|_{op} \;\le\; 1.$$

Define

$$B := \frac{1}{1-\gamma}, \qquad Y_0 := 1 + \gamma\sigma_M + \|\theta^*\|_2, \qquad Y_1 := 1, \qquad \beta(\tau) := \|\mathbb{P}(Z_k \in \cdot \mid \mathcal{G}_{k-\tau}) - d^\pi\|_{\mathrm{TV}},$$

and assume geometric mixing $\beta(\tau) \le C_{mix}\rho^{\tau}$ with $0 < \rho < 1$. Then:

$$C_e := 4Y_1 C_{mix} + L_F^2 \;+\; 16\big(1+\beta(\tau)^2\big)\,Y_1^2 \;=\; 1 + 4C_{mix} \;+\; 16\big(1+\beta(\tau)^2\big),$$

$$C_0 := 16\big(1+\beta(\tau)^2\big)\,Y_0^2 + 2\sigma^2 \;+\; 4\gamma^2\sigma_M^2,$$

$$C_{cross} := \frac{8\,Y_0^2}{\mu},$$

$$C_{bias} := C_{fast}^2 \;+\; \frac{B^2}{2} \;+\; 16\,B^2\,\frac{C_{mix}\rho}{1-\rho}, \qquad C_{fast} := \frac{(4B_\nu^2 + \beta_0^2(8B_\nu g_M C_{mix}\frac{1}{\beta_0} + 25g_M^2)\ln(2))}{\beta_0}.$$

Recommended lag schedule. If $\tau_k = \lceil \frac{\log((k+1)/c)}{\log(1/\rho)} \rceil$ then $\beta(\tau_k) \le C_{mix}c/k$ and the mixing penalty in equation 37 becomes

$$C_{cross}\,\alpha_k\,\beta(\tau_k)^2 \;\le\; C_{cross}\,C_{mix}^2 c^2\,\frac{\alpha_k}{(k+1)^2},$$

with *no* $\log k$ factor.

Now we derive the last iterate convergence of the error $\|e_{k+1}^{\theta}\|_2$.

Hence,

$$\mathbb{E}\|e_{k+1}\|_2^2 \;\le\; \Big(1 - \tfrac{\mu}{2}\,\alpha_k + C_e\,\alpha_k^2\Big)\,\mathbb{E}\|e_k\|_2^2 + \alpha_k^2\left(\left(\tfrac{2}{\mu}+2\right)\frac{C_{bias}}{k} + C_{cross}^2 C_{mix}^2 + C_0\right) + \frac{2\alpha_k(\epsilon_{approx}^{dual})^2}{\mu}$$

Or,

$$\mathbb{E}\|e_{k+1}\|_2^2 \;\le\; \Big(1 - \tfrac{\mu}{2}\,\alpha_k + C_e\,\alpha_k^2\Big)\,\mathbb{E}\|e_k\|_2^2 \;+\alpha_k^2 C_1 + \frac{2\alpha_k(\epsilon_{approx}^{dual})^2}{\mu} \tag{39}$$

with $C_1 = \Big(\left(\frac{2}{\mu}+2\right)\frac{C_{bias}}{k} + C_{cross}^2 C_{mix}^2 + C_0\Big)$.

Next, we derive the final iterate convergence from the above recursion for the following two step size rules:

1. $\omega \in (1.2, 1)$
2. $\omega = 1$

C.3 Last Iterate Convergence Guarantee for Step-Size Rule $\alpha_k = \frac{c}{(1+k)}$ (with dual approximation term)

Recall the recursion

$$\mathbb{E}\big[\|e_{k+1}\|_2^2\big] \;\le\; \Big(1 - \tfrac{\mu}{2}\alpha_k + C_e\,\alpha_k^2\Big)\,\mathbb{E}\big[\|e_k\|_2^2\big] \;+\; \alpha_k^2 C_1 \;+\; \tfrac{2}{\mu}\,\alpha_k\,\big(\varepsilon_{\text{approx}}^{\text{dual}}\big)^2, \qquad \alpha_k = \tfrac{c}{k+1}. \tag{40}$$

Define

$$a_k \;:=\; 1 - \tfrac{\mu}{2}\alpha_k + C_e\,\alpha_k^2, \qquad k_0 \;:=\; \left\lceil \tfrac{2C_e c}{\mu} \right\rceil,$$

so that for all $j \ge k_0$,

$$a_j \;\le\; 1 - \frac{\mu c}{2(j+1)}.$$

For $s \le u$, set

$$G_s^u \;\coloneqq\; \prod_{j=s}^{u} a_j, \qquad H_0 \;\coloneqq\; \prod_{j=0}^{k_0-1} a_j.$$

Then for any $s \le k-1$,

$$G_s^{k-1} \;=\; H_0\, G_{k_0}^{k-1}.$$

The tail product admits the standard bound

$$G_{k_0}^{k} \;\le\; \left(\frac{k_0+1}{k+2}\right)^{\rho}, \qquad \rho \;\coloneqq\; \frac{\mu c}{2}. \tag{41}$$

Define

$$H_0 \coloneqq \prod_{j=0}^{k_0-1} a_j.$$

Also define

$$U_{pre}(k_0) \coloneqq \sum_{t=0}^{k_0-1} \frac{c^2 C_1}{(t+1)^2} \left(\prod_{j=t+1}^{k_0-1} a_j\right)$$

Unrolling equation 40 yields

$$\mathbb{E}\big[\|e_k\|_2^2\big] \le \mathbb{E}\big[\|e_0\|_2^2\big]\, G_0^{k-1} \;+\; \sum_{t=0}^{k-1} \alpha_t^2 C_1\, G_{t+1}^{k-1} \;+\; \frac{2}{\mu}\big(\varepsilon_{\text{approx}}^{\text{dual}}\big)^2 \sum_{t=0}^{k-1} \alpha_t\, G_{t+1}^{k-1}.$$

Splitting the variance sum at k_0 and using equation 41 gives the bound:

$$\begin{aligned}\mathbb{E}\big[\|e_k\|_2^2\big] \le \Big(\mathbb{E}[\|e_0\|_2^2]\, H_0 + U_{\text{pre}}(k_0)\Big)\left(\frac{k_0+1}{k+1}\right)^{\rho} \;+\; \sum_{t=k_0}^{k-1} \frac{c^2 C_1}{(t+1)^2}\left(\frac{t+2}{k+1}\right)^{\rho} \\ +\; \frac{2}{\mu}\big(\varepsilon_{\text{approx}}^{\text{dual}}\big)^2 \Bigg[\underbrace{\sum_{t=0}^{k_0-1} \alpha_t \prod_{j=t+1}^{k_0-1} a_j}_{B_{\text{pre}}(k_0)} \left(\frac{k_0+1}{k+1}\right)^{\rho} \;+\; \sum_{t=k_0}^{k-1} \frac{c}{t+1}\left(\frac{t+2}{k+1}\right)^{\rho}\Bigg].\end{aligned}$$

An elementary algebraic bound on the variance tail yields the same case split in ρ:

$$\sum_{t=k_0}^{k-1} \frac{c^2 C_1}{(t+1)^2}\left(\frac{t+2}{k+1}\right)^{\rho} \;\le\; \begin{cases} \dfrac{c^2 C_1}{\rho-1}\dfrac{1}{k+1} + \dfrac{c^2 C_1}{\rho-1}\dfrac{(k_0+2)^{\rho-1}}{(k+1)^{\rho}}, & \rho > 1, \\ \dfrac{c^2 C_1}{k+1}\left(1 + \ln\dfrac{k+1}{k_0+2}\right), & \rho = 1, \\ \dfrac{c^2 C_1}{1-\rho}\dfrac{(k_0+2)^{\rho-1}}{(k+1)^{\rho}}, & 0 < \rho < 1. \end{cases}$$

For the dual-approximation contribution, the pre-burn part is a constant times $(\frac{k_0+1}{k+1})^{\rho}$:

$$B_{\text{pre}}(k_0)\left(\frac{k_0+1}{k+1}\right)^{\rho}.$$

The tail part is handled by comparing the sum to an integral:

$$\sum_{t=k_0}^{k-1} \frac{c}{t+1}\left(\frac{t+2}{k+1}\right)^{\rho} \;\le\; \frac{c}{(k+1)^{\rho}} \int_{k_0}^{k} (x+2)^{\rho-1}\, dx \;=\; \frac{c}{\rho}\left(1 - \left(\frac{k_0+2}{k+2}\right)^{\rho}\right).$$

Therefore, the total dual-approximation contribution satisfies

$$\frac{2}{\mu}\big(\varepsilon_{\text{approx}}^{\text{dual}}\big)^2 \sum_{t=0}^{k-1} \alpha_t\, G_{t+1}^{k-1} \;\le\; \underbrace{\frac{2}{\mu}\cdot\frac{c}{\rho}}_{=\,\frac{4}{\mu^2}} \big(\varepsilon_{\text{approx}}^{\text{dual}}\big)^2 \;+\; \sigma_M^2 B_{pre}(k_0)\left(\frac{k_0+1}{k+1}\right)^{\rho}, \qquad \rho = \frac{\mu c}{2}.$$

Final bound. Combining the pieces, for $\rho = \frac{\mu c}{2}$,

$$\mathbb{E}\big[\|e_k\|_2^2\big] \le \Big(\mathbb{E}[\|e_0\|_2^2]\, H_0 + U_{\text{pre}}(k_0)\Big)\left(\frac{k_0+1}{k+1}\right)^{\rho} + \begin{cases} \dfrac{c^2 C_1}{\rho-1}\dfrac{1}{k+1} + \dfrac{c^2 C_1}{\rho-1}\dfrac{(k_0+2)^{\rho-1}}{(k+1)^{\rho}}, & \rho > 1, \\ \dfrac{c^2 C_1}{k+1}\left(1 + \ln\dfrac{k+1}{k_0+2}\right), & \rho = 1, \\ \dfrac{c^2 C_1}{1-\rho}\dfrac{(k_0+2)^{\rho-1}}{(k+1)^{\rho}}, & 0 < \rho < 1, \end{cases} + \frac{4}{\mu^2}\left(\varepsilon_{\text{approx}}^{\text{dual}}\right)^2 + \sigma_M^2 B_{pre}(k_0)\left(\frac{k_0+1}{k+1}\right)^{\rho}.$$

With this final bound along with the bound on $\|\theta^{*,t}\|_2$ from subsection D.1, we arrive at the result in Theorem 1 for the step size rule $\alpha_k = \frac{c}{1+k}$

C.4 Last-iterate convergence for $\alpha_k = \frac{c}{(k+1)^{\omega}}$ with approximation error

Fix $\omega \in (\frac{1}{2}, 1)$ and constants $c, \mu > 0$. Let $a := c\mu$, and suppose the error sequence $s_k := \mathbb{E}\big[\|e_k\|_2^2\big]$ satisfies, for some $C_1 > 0$ and $C_e > 0$,

$$s_{k+1} \;\le\; \Big(1 - \frac{a}{(k+1)^{\omega}} + \frac{D}{(k+1)^{2\omega}}\Big)s_k \;+\; \frac{b}{(k+1)^{2\omega}} \;+\; \frac{\rho'}{(k+1)^{\omega}}, \qquad D := C_e c^2,\; b := c^2 C_1, \tag{42}$$

where the following term comes from the function-approximation bias on the slow time scale:

$$\frac{\rho'}{(k+1)^{\omega}} \;=\; \frac{2}{\mu}\,\alpha_k\,(\epsilon_{\text{approx}}^{\text{dual}})^2 \;=\; \frac{2c}{\mu}\cdot\frac{(\epsilon_{\text{approx}}^{\text{dual}})^2}{(k+1)^{\omega}}, \qquad \rho' := \frac{2c}{\mu}\,(\epsilon_{\text{approx}}^{\text{dual}})^2.$$

Burn-in index. Define

$$k_{\alpha^2} := \left\lceil \left(\tfrac{2D}{a}\right)^{1/\omega} \right\rceil - 1, \qquad k_{\text{res}} := \left\lceil \left(\tfrac{2\omega}{a}\right)^{1/(1-\omega)} \right\rceil - 1, \qquad k_0 := \max\{k_{\alpha^2},\, k_{\text{res}}\}.$$

Then, for all $k \ge k_0$,

$$1 - \frac{a}{(k+1)^{\omega}} + \frac{D}{(k+1)^{2\omega}} \;\le\; 1 - \frac{a}{2}\cdot\frac{1}{(k+1)^{\omega}}. \tag{43}$$

Bias removal by shifting. Introduce the shifted sequence

$$t_k \;:=\; s_k - \frac{\rho'}{a}\,.$$

Substituting $s_k = t_k + \rho'/a$ into equation 42 and simplifying (the $\rho'/(k+1)^{\omega}$ cancels with the $-a/(k+1)^{\omega}$ part) yields

$$t_{k+1} \;\le\; \Big(1 - \frac{a}{(k+1)^{\omega}} + \frac{D}{(k+1)^{2\omega}}\Big)t_k \;+\; \frac{\tilde{b}}{(k+1)^{2\omega}}, \qquad \tilde{b} \;:=\; b + \frac{\rho' D}{a}. \tag{44}$$

Rescaling and max-envelope. Let $g_k := (k+1)^{\omega} t_k$. Multiplying equation 44 by $(k+2)^{\omega}$ and using equation 43 for $k \ge k_0$, a bit of algebra gives

$$g_{k+1} \;\le\; g_k\Big(1 + \frac{\omega}{k+1} - \frac{a}{2}\cdot\frac{1}{(k+1)^{\omega}}\Big) \;+\; \frac{2\tilde{b}}{(k+1)^{\omega}}, \qquad k \ge k_0. \tag{45}$$

Let $\eta_k := \frac{a}{2}(k+1)^{-\omega}$ and $\hat{g}_k := \max_{0\le j\le k} g_j$. Then, for $k \ge k_0$,

$$g_{k+1} \;\le\; \hat{g}_k - \eta_k \hat{g}_k + \frac{2\tilde{b}}{(k+1)^{\omega}}.$$

Choose

$$G \;:=\; \max\Big\{ \underbrace{\sup_{0\le j\le k_0} g_j}_{\text{pre-burn}},\; \frac{8\tilde{b}}{a} \Big\}.$$

Since $\eta_k \cdot \frac{8\tilde{b}}{a} = \frac{4\tilde{b}}{(k+1)^{\omega}} \ge \frac{2\tilde{b}}{(k+1)^{\omega}}$, a standard induction yields $g_k \le G$ for all $k \ge k_0$.

Bounding the pre-burn maximum. Dropping the negative drift for $k \le k_0$ and iterating gives

$$\sup_{0\le j\le k_0} t_j \;\le\; \exp\Big(D\sum_{m=1}^{\infty}\tfrac{1}{m^{2\omega}}\Big)\,\Big(t_0+\tilde{b}\,\zeta(2\omega)\Big),$$

where $\zeta(s)=\sum_{m=1}^{\infty} m^{-s}$ and $\zeta(2\omega)\le 1+\frac{1}{2\omega-1}$. Hence

$$\sup_{0\le j\le k_0} g_j \;\le\; (k_0+1)^{\omega}\,\exp\Big(\tfrac{\pi^2}{6}D\Big)\,\Big(t_0+\tilde{b}\,\zeta(2\omega)\Big), \qquad t_0 = s_0-\frac{\rho'}{a} \;\le\; s_0+\frac{\rho'}{a}. \tag{46}$$

Last-iterate bound (all $k \ge 0$). Combining the pieces and recalling $t_k = s_k - \rho'/a$, we obtain

$$\boxed{\mathbb{E}\big[\|e_k\|_2^2\big] \;\le\; \frac{1}{(k+1)^{\omega}}\,\max\Big\{\underbrace{(k_0+1)^{\omega}\exp\Big(\tfrac{\pi^2}{6}D\Big)\Big(s_0+\tfrac{\rho'}{a}+\tilde{b}\,\zeta(2\omega)\Big)}_{\text{finite pre-burn constant}},\ \underbrace{\frac{8\tilde{b}}{a}}_{\text{tail constant}}\Big\} + \underbrace{\frac{\rho'}{a}}_{\text{approx. floor}}, \qquad k\ge 0.}$$
(47)

In particular, the approximation-induced floor is

$$\frac{\rho'}{a} \;=\; \frac{\frac{2c}{\mu}(\epsilon_{\text{approx}}^{\text{dual}})^2}{c\mu} \;=\; \frac{2\,(\epsilon_{\text{approx}}^{\text{dual}})^2}{\mu^2},$$

This proves the result in Theorem 1 for the step-size rule $\alpha_k = \frac{c}{(1+k)^{\omega}}$ for $\omega \in (1/2, 1)$.

D Remaining Proofs

D.1 Bound on $\theta^{*,t}$

We drop the superscript t from $\theta^{*,t}$ as t is fixed throught the discussion of this subsection. Recall,

$$\theta^* = (\Phi^\top D^\pi \Phi)^{-1}\Phi^\top D^\pi[r+\gamma\bar{\sigma}(\lambda^*)]$$

Lemma 6 (Bound on the optimal weight vector) *Let*

$$\theta^* \;=\; \big(\Phi^\top D^\pi\Phi\big)^{-1}\Phi^\top D^\pi\big[r \;+\; \gamma\,\bar{\sigma}(\lambda^*)\big],$$

where

- $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ *has full column rank and row vectors* $\phi(s,a)$ *satisfying* $\|\phi(s,a)\|_2 \le 1$*;*
- $\Phi^\top D^\pi \Phi$ *is positive definite with min-eigenvalue* ν*. The diagonal entries of the matrix* D^π *satisfy* $d_i \ge d_{\min} > 0$ *and* $\sum_i d_i = 1$*;*
- *each entry of* r *obeys* $|r_i| \le 1$*;*
- *each entry of* $\bar{\sigma}(\lambda^*)$ *obeys* $\big|\bar{\sigma}(\lambda^*)_i\big| \le 1/(1-\gamma)$.

Then

$$\boxed{\|\theta^*\|_2 \;\le\; \frac{1}{1-\gamma}\,\frac{1}{\sqrt{\mu}}.}$$

Proof 5 *Set*

$$C := \Phi^\top D^\pi \Phi, \qquad v := r+\gamma\,\bar{\sigma}(\lambda^*).$$

Because C *is invertible,* $\theta^* = C^{-1}\Phi^\top D^\pi v$ *satisfies* $C\theta^* = \Phi^\top D^\pi v$*. Multiply on the left by* $\theta^{*\top}$*:*

$$\theta^{*\top} C\theta^* \;=\; \theta^{*\top}\Phi^\top D^\pi v \;=\; (\Phi\theta^*)^\top D^\pi v.$$

Let $y := \Phi\theta^*$*. With the weighted Cauchy–Schwarz inequality and* $D^{\pi 1/2} := operatornamediag(\sqrt{d_i})$*:*

$$y^\top D^\pi v \;=\; (D^{\pi 1/2}y)^\top (D^{\pi 1/2}v) \;\le\; \|D^{\pi 1/2}y\|_2\,\|D^{\pi 1/2}v\|_2.$$

But $\|D^{\pi 1/2}y\|_2^2 = y^\top D^\pi y$, *and* $y^\top D^\pi y = y^\top D^\pi v$ *so*

$$y^\top D^\pi y \;\le\; \|D^{\pi 1/2}y\|_2 \, \|D^{\pi 1/2}v\|_2 \;=\; (y^\top D^\pi y)^{1/2} \, \|D^{\pi 1/2}v\|_2.$$

Whenever $y^\top D^\pi y > 0$ *(otherwise* $\theta^* = 0$ *and the bound is trivial), divide both sides to obtain*

$$y^\top D^\pi y \;\le\; \|D^{\pi 1/2}v\|_2^2 \;=\; v^\top D^\pi v.$$

Returning to θ^**:*

$$\theta^{*\top} C \theta^* = y^\top D^\pi y \;\le\; v^\top D^\pi v.$$

—

E ROBUST Q-LEARNING

In this section, we discuss a robust Q-learning algorithm with function approximation that finds the optimal policy for the worst-case transition kernel in the uncertainty set considered in this paper. We first define the optimal state-action value function Q_r^* as the state-action value function of the best admissible policy to maximize Q_r^π for each (s,a)-pair.

$$Q_r^*(s,a) = \max_\pi Q_r^\pi(s,a), \forall (s,a).$$

It is shown in prior literature (Iyengar, 2005) that Q_r^* satisfies the following equation, which is called the robust Bellman optimality equation

$$Q_r^*(s,a) \;=\; r(s,a) \;+\; \gamma \min_{q \in \mathcal{P}_s^a} \sum_{s'} q(s' \mid s,a) \underbrace{\max_{a'} Q_r^*(s',a')}_{=:\, V_r^*(s')}. \tag{48}$$

Equivalently, define the robust Bellman optimality operator $(\mathcal{T}_r^* Q)(s,a) := r(s,a) + \gamma \, \sigma_{\mathcal{P}_s^a}(V_r^*)$ with

$$V_r^*(s') := \max_{a'} Q(s',a'), \tag{49}$$

and $\sigma_{\mathcal{P}_s^a}(V)$ is given in Equation (4). Iyengar (2005) proved that the robust Bellman optimality operator is γ-contraction in ℓ_∞ norm.

Now, we discuss how the TD learning algorithm presented in Algorithm 1 in the main body of the paper can be extended to estimate Q_r^* in a relatively straightforward manner.Similar to the TD learning setup, assume that we can sample data corresponding to a behavioral policy π_b from the nominal model P_0. Also, assume that the policy π_b satisfies Assumption 2.

The goal here is to approximate Q_r^* by $\Phi\theta^*$ for an appropriately chosen θ^*. Our Q-learning algorithm is presented in Algorithm 2. The algorithm an estimate $\hat{\theta}_t$ of this parameter at each iteration t of the outer loop. The quantity $V_{\hat{\theta}_t}^*$ in the description of the algorithm is given by

$$V_{\hat{\theta}_t}^*(s) = \max_a Clip\left(\phi(s,a)^\top \hat{\theta}_t, -\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right), \forall s \in \mathcal{S}. \tag{50}$$

Difference between Algorithm 2 and Algorithm 1: The only difference between the robust Q-learning algorithm in Algorithm 2 and the robust TD learning algorithm in Algorithm 1 is that, we use $V_{\hat{\theta}_t}^*$ instead of $V_{\hat{\theta}_t}$ in the calculation of the dual super-gradient in line 6 and the calculation of the dual objective in line 10 in Algorithm 2.

Finite-Time Performance Bound for the Robust Q-Learning (Algorithm 2): Recall that we established a finite-time performance bound for the robust TD learning in Theorem 1. By following the steps of the proof of that theorem, it is easy to see that an analogous guarantee holds for the estimate of Q_r^* produced by Algorithm 2. The reason that the proof is identical is that the robust Bellman optimality operator is a γ-contraction in the ℓ_∞ norm as was the robust Bellman operator for a fixed policy. The only difference is that the function approximation error for approximating

the Q-function should now be defined as the error in approximating Q_r^* by the class of functions $\{\Phi\theta : \theta \in \mathbb{R}^d\}$:

$$\epsilon^*_{approx} \coloneqq \sup_{Q=Clip\left(\Phi\theta,-\frac{1}{1-\gamma},\frac{1}{1-\gamma}\right),\theta\in\mathbb{R}^d} \left\| Clip\left(\Pi\mathcal{T}_r^*(Q), -\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right) - \mathcal{T}_r^*(Q)\right\|_\infty . \quad (51)$$

Thus, the sample complexity of robust Q-learning is of the same order as that of robust TD-learning up to a function approximation error.

Algorithm 2 Robust Q-learning with Function Approximation

Input: Integers T, K. Initial $\nu_0 \in \mathbb{R}^{d_\lambda}$, $\theta_0 \coloneqq$ zero vector, fast time-scale step-sizes $\beta_k = \frac{\beta_0}{\sqrt{k+1}}$, slow time-scale step-sizes $\alpha_k = \frac{c}{(k+1)^\omega}$ for some $\omega \in (0.5, 1]$; $\hat{\theta}_0 = \theta_0, \theta_{0,0} = \theta_0$, candidate policy π, Reward function $r : (\mathcal{S} \times \mathcal{A}) \mapsto [0, 1]$, initial state S_0^0.
for $t = 0, 1, ..., T-1$ **do**
  for $k = 0, 1, \ldots, K-1$ **do**
    Take action A_k^t according to the behavioral policy π_b and Sample S_{k+1}^t ($S_{k+1}^t \sim P_0(\cdot|S_k^t, A_k^t)$)
  Fast scale (β_k)
    Compute $g(\psi(S_k^t, A_k^t)^\top \nu_{t,k}; S_{k+1}^t, V^*_{\hat{\theta}_t})$ from Equation (19) for TV distance and Equation (22) for Wasserstein-ℓ uncertainty
    $\nu_{t,k+1} = Proj_{\mathcal{M}_\nu}(\nu_{t,k} + \beta_k[g(\psi(S_k^t, A_k^t)^\top \nu_{t,k}; S_{k+1}^t, V_{\hat{\theta}_t})\psi(S_k^t, A_k^t)])$
  Slow scale (α_k)
    Compute $\bar{\nu}_{t,k}$ from Equation (8)
    Compute $\sigma(\psi(S_k^t, A_k^t)^\top \bar{\nu}_{t,k}; S_{k+1}^t, V^*_{\hat{\theta}_t})$ from Equation (20) for TV distance and Equation (23) for Wasserstein-ℓ uncertainty
    $TD_{t,k+1} = r(S_k^t, A_k^t) + \gamma\sigma(\psi(S_k^t, A_k^t)^\top \bar{\nu}_{t,k}; S_{k+1}^t, V_{\hat{\theta}_t}) - \phi(S_k^t, A_k^t)^\top \theta_{t,k}$
    $\theta_{t,k+1} = \theta_{t,k} + \alpha_k TD_{t,k+1}\phi(S_k^t, A_k^t)$
  end for
  $\hat{\theta}_{t+1} = \theta_{t,K}, S_0^{t+1} = S_K^t, \theta_{t,0} = \theta_0, \nu_{t,0} = \nu_0$
end for
Output: $\hat{\theta}_T$

USE OF LARGE LANGUAGE MODEL

The authors used large language models (e.g., ChatGPT) to polish the language in certain parts of the paper. All technical content, proofs, and conclusions are the sole work of the authors.