SELF-TUNED Q-LEARNING AND SARSA: POLICY CONTROL WITHOUT WORRYING ABOUT STEP-SIZE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Q-learning and SARSA are foundational reinforcement learning algorithms whose practical success depends critically on step-size calibration: step-sizes that are too large cause numerical instability, while step-sizes that are too small yield slow progress. We propose self-tuned variants of both algorithms that recast the standard iterative updates as fixed-point equations, yielding a data-adaptive step-size adjustment. Our non-asymptotic analysis shows that the proposed methods remain stable over substantially broader step-size ranges. Under mild conditions, they admit arbitrarily large initial step sizes, greatly reducing the influence of initialization while matching the convergence rates of the standard algorithms. Empirical validation across environments with both discrete and continuous state spaces shows that self-tuned Q-learning and SARSA are markedly less sensitive to step-size selection, remaining stable at step-sizes that cause the standard methods to suffer numerical instability.

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful framework for sequential decision-making, enabling agents to learn optimal behaviors through interaction with their environment (Sutton & Barto, 1998). A key component of many successful RL algorithms is the action-value function, which quantifies how much total reward an agent can expect to accumulate after taking a particular action from a given state. Among the most fundamental and widely studied action-value based approaches in RL are Q-learning (Watkins, 1989; Watkins & Dayan, 1992) and State-Action-Reward-State-Action (SARSA) (Rummery & Niranjan, 1994), which have demonstrated remarkable success across diverse applications, including game playing (Mnih et al., 2015), robotics (Kober et al., 2013), and autonomous systems (Sallab et al., 2017). A persistent challenge with such methods is their sensitivity to step-size selection, which can critically affect numerical stability. Although tabular methods with decreasing step-sizes under a suitable behavioral policy enjoy well-established asymptotic convergence guarantees (Watkins & Dayan, 1992; Jaakkola et al., 1993; Tsitsiklis, 1994; Singh et al., 2000), asymptotic convergence alone does not preclude numerical instability in the early phase of training, and existing finite-time bounds typically impose restrictive conditions on step-sizes (Zou et al., 2019b;a; Zhang & Xie, 2024; Liu et al., 2025). This sensitivity is further compounded by function approximation, which not only introduces well-known instability issues such as divergence or failure to converge in Q-learning and SARSA with linear function approximation (Baird, 1995; Gordon, 1996), but also amplifies step-size sensitivity, making careful step-size selection a persistent practical challenge. Overly large step-sizes induce oscillations and potential divergence, while conservative step-sizes result in prohibitively slow learning. Such sensitivity makes it difficult to identify a step-size schedule that achieves both stability and efficiency (Dabney & Barto, 2012; Sutton & Barto, 1998; Szepesvári, 2022; Chen et al., 2019; Zou et al., 2019a). We seek a principled approach to reduce step-size sensitivity with theoretical guarantees.

1.1 RELATED WORKS

Convergence Analysis. For tabular settings, asymptotic convergence of Q-learning was established under appropriate step-size conditions (Watkins & Dayan, 1992; Jaakkola et al., 1993; Tsitsiklis,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1994), and convergence of tabular SARSA was similarly established under suitable assumptions on the behavioral policy (Singh et al., 2000). However, with linear function approximation, the convergence properties become substantially more complex. Baird (1995) provided an illustrative counter-example demonstrating the divergence of Q-learning with linear function approximation. Recent work has shown that linear Q-learning with an ϵ-softmax behavioral policy and adaptive temperature yields iterates that remain within a bounded set (Meyn, 2024; Liu et al., 2025) under a restrictive step-size condition. For SARSA with linear function approximation, convergence depends critically on the policy improvement operator. Gordon (1996; 2000) demonstrated that SARSA with an ϵ-greedy policy can exhibit chattering and fail to converge. Melo et al. (2008) showed that linear SARSA converges to a unique fixed point when the policy improvement operator is Lipschitz continuous with a sufficiently small constant. Zou et al. (2019a) later established finite-time error bounds for projected linear SARSA through novel techniques for handling time-inhomogeneous Markov chains.

Implicit Updates. Implicit update schemes have emerged as a powerful approach to enhance the stability of stochastic approximation algorithms. Works on implicit stochastic gradient descent (SGD) (Toulis & Airoldi, 2015; 2017; Chee et al., 2023) demonstrated that evaluating a gradient at a future iterate rather than the current one provides adaptive shrinkage. In the context of policy evaluation, Kim et al. (2025b) formalized implicit temporal difference (TD) methods and analyzed their finite-time and asymptotic behavior, showing that robustness to step-size selection is substantially improved. These advances demonstrate that implicit updates provide a principled mechanism for step-size adaptation, motivating their extension to control algorithms such as Q-learning and SARSA.

1.2 Contributions

Our main contributions are as follows.

- **Self-tuned Q-learning and SARSA algorithms:** We propose variants of Q-learning and SARSA, termed self-tuned Q-learning (ST-Q-learning) and SARSA (ST-SARSA), whose updates are formulated as fixed-point equations. These admit closed-form iterative updates with data-adaptive step-sizes that scale inversely with feature norms, providing automatic step-size adjustment.
- **Finite-time analysis of self-tuned Q-learning:** We establish non-asymptotic error bounds for projected linear Q-learning and its self-tuned version under Markovian data. We consider decreasing step-sizes of the form $\beta_t = \beta_0/(t+1)^s$ with $s \in (0.5, 1)$. Our analysis reveals a sharp contrast between the two methods. Standard Q-learning exhibits critical sensitivity to step-size selection: large initial step-sizes cause instability and require careful calibration. In contrast, self-tuned Q-learning is remarkably robust: its error bound holds for any initial step-size, eliminating the need for tuning. Moreover, larger step-sizes are desirable, as they reduce the effect of initialization.
- **Finite-time analysis of self-tuned SARSA:** Under time-varying Lipschitz behavioral policies, we establish non-asymptotic error bounds for projected linear SARSA with decreasing step-sizes, improving upon those of Zou et al. (2019a), as well as corresponding bounds for self-tuned SARSA. The established bound holds for any initial step-size, including values for which standard SARSA may become unstable, thereby eliminating the need for careful step-size tuning and allowing larger initial step-sizes to reduce the effect of initialization.
- **Environment agnostic step-size:** Importantly, our all-time bounds for the last iterate are established under a universal step-size schedule that requires no prior knowledge of the environment. Moreover, our results can be interpreted as last-iterate error bounds under data-adaptive step sizes, for which finite-time guarantees remain relatively scarce in the existing literature.
- **Empirical validation:** Across benchmarks with discrete and continuous state spaces, we demonstrate the effectiveness of self-tuned Q-learning and SARSA, which remain stable at step-sizes where standard methods fail, validating our theory. Code is available at `https://anonymous.4open.science/r/implicit_control-BEDF`.

2 BACKGROUND

Markov Decision Processes. Consider a discounted MDP $(\mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma)$ composed of a continuous state space $\mathcal{S} \subset \mathbb{R}^n$, a finite action space $\mathcal{A}$, a time-homogeneous transition kernel $\mathsf{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, a bounded reward $r : \mathcal{S} \times \mathcal{A} \to [0, b]$ for some $b > 0$ and a discount factor $\gamma \in (0, 1)$. Suppose we are given a stationary policy π, i.e., policy does not change over time. A policy π is deterministic if it is a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$, and stochastic if it is a mapping from the state space $\mathcal{S}$ to the simplex $\Delta(\mathcal{A})$, the space of probability distributions over $\mathcal{A}$. At each time step $t \geq 0$, an agent is at state $\boldsymbol{S}_t^\pi$ and takes an action A_t^π according to π. The agent then receives a reward $R_t^\pi := r(\boldsymbol{S}_t^\pi, A_t^\pi)$ and transitions to the next state $\boldsymbol{S}_{t+1}^\pi$ with a probability $\mathsf{P}(\boldsymbol{S}_t^\pi, \boldsymbol{S}_{t+1}^\pi, A_t^\pi)$.

Given a policy π, the corresponding value function is given by $V^\pi(\boldsymbol{s}) = \mathbb{E}^\pi\left(\sum_{t=0}^\infty \gamma^t R_t^\pi \mid \boldsymbol{S}_0^\pi = \boldsymbol{s}\right)$ and the corresponding action-value function is defined as $Q^\pi(\boldsymbol{s}, a) = \mathbb{E}^\pi\left(\sum_{t=0}^\infty \gamma^t R_t^\pi | \boldsymbol{S}_0^\pi = \boldsymbol{s}, A_0^\pi = a\right)$. Note that the expectation $\mathbb{E}^\pi$ is over trajectories $\{\boldsymbol{S}_t^\pi, A_t^\pi, R_t^\pi\}_{t\geq 0}$ in which π is followed at each time step. The optimal value function is defined as $V^\star(\boldsymbol{s}) := \sup_\pi V^\pi(\boldsymbol{s})$ for all $\boldsymbol{s} \in \mathcal{S}$ and the optimal action-value function is defined as $Q^\star(\boldsymbol{s}, a) := \sup_\pi Q^\pi(\boldsymbol{s}, a)$ for all $(\boldsymbol{s}, a) \in \mathcal{S} \times \mathcal{A}$. An optimal policy $\pi^\star$ can be obtained from the optimal action-value function $Q^\star$ by selecting $\pi^\star(\boldsymbol{s}) \in \arg\max_{a\in\mathcal{A}} Q^\star(\boldsymbol{s}, a)$ (Puterman, 2014).

Linear Function Approximation. When the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite and relatively low-dimensional, storing the action-value function Q^π in tabular form for all possible state-action pairs is manageable. However, when either space becomes high-dimensional or continuous, this tabular approach is no longer feasible. A computationally tractable alternative is to approximate the action-value function using a basis function representation. To this end, we introduce feature basis functions $\{\phi_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_{i=1}^M$ satisfying $\|\boldsymbol{\phi}(\boldsymbol{s}, a)\|_2 \leq 1$ for all $(\boldsymbol{s}, a) \in \mathcal{S} \times \mathcal{A}$. We then approximate the action-value function Q^π via a linear combination of these basis functions, i.e., $Q^\pi(\boldsymbol{s}, a) \approx \boldsymbol{\phi}(\boldsymbol{s}, a)^\top \boldsymbol{\theta} = \sum_{i=1}^M \phi_i(\boldsymbol{s}, a)\theta_i$ where $\boldsymbol{\phi}(\boldsymbol{s}, a) := [\phi_1(\boldsymbol{s}, a), \ldots, \phi_M(\boldsymbol{s}, a)]^\top$, $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_M]^\top$. Here, $\boldsymbol{\theta} \in \mathbb{R}^M$ is the weight parameter vector characterizing the linear function approximation. The problem of learning an optimal action-value function (and thus an optimal policy) therefore reduces to learning an optimal parameter vector $\boldsymbol{\theta}^\star \in \mathbb{R}^M$.

Q-Learning. Given a behavioral policy π used to generate data $\{\boldsymbol{S}_t^\pi, A_t^\pi\}_{t\in\mathbb{N}}$, Q-learning is an iterative algorithm that approximates the optimal action-value function $Q^\star$ from data. In the tabular setting, where both state space $\mathcal{S}$ and action space $\mathcal{A}$ are finite, Q-learning maintains a look-up table $\widehat{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ which stores the estimated action-value for all state-action pairs. Given initial state $\boldsymbol{S}_0^\pi$ and initial estimate $\widehat{Q}_0$, at iteration $t \in \mathbb{N} \cup \{0\}$, the update rule yields $(t+1)^{\text{th}}$ estimate given by

$$\widehat{Q}_{t+1}(\boldsymbol{S}_t^\pi, A_t^\pi) = \widehat{Q}_t(\boldsymbol{S}_t^\pi, A_t^\pi) + \beta_t \delta_t^Q$$
$$\delta_t^Q := R_t^\pi + \gamma \max_{a\in\mathcal{A}} \widehat{Q}_t(\boldsymbol{S}_{t+1}^\pi, a) - \widehat{Q}_t(\boldsymbol{S}_t^\pi, A_t^\pi)$$

where $(\beta_t)_{t\geq 0}$ is a sequence of step-sizes. Since the update rule does not rely on the action A_{t+1}^π obtained from the behavioral policy π, Q-learning is classified as an off-policy learning algorithm. The tabular setting can be viewed as a special case of linear function approximation with one-hot encoding basis functions $\{\boldsymbol{e}(\boldsymbol{s}, a)\}_{\boldsymbol{s}\in\mathcal{S}, a\in\mathcal{A}}$, where $\boldsymbol{e}(\boldsymbol{s}, a) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ has a single entry equal to one at the position corresponding to state-action pair $(\boldsymbol{s}, a)$ and zeros elsewhere. With the representation $\widehat{Q}_t(\boldsymbol{s}, a) = \boldsymbol{e}(\boldsymbol{s}, a)^\top \widehat{\boldsymbol{\theta}}_t$ for all $(\boldsymbol{s}, a) \in \mathcal{S} \times \mathcal{A}$ and $\boldsymbol{e}_t^\pi := \boldsymbol{e}(\boldsymbol{S}_t^\pi, A_t^\pi)$, the tabular Q-learning update can be equivalently expressed as

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t \left\{ R_t^\pi + \gamma \max_{a\in\mathcal{A}} \boldsymbol{e}(\boldsymbol{S}_{t+1}^\pi, a)^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{e}_t^{\pi\top} \widehat{\boldsymbol{\theta}}_t \right\} \boldsymbol{e}_t^\pi.$$

Generalizing to a broader class of basis functions $\boldsymbol{\phi}(\boldsymbol{s}, a) = [\phi_1(\boldsymbol{s}, a), \ldots, \phi_M(\boldsymbol{s}, a)]^\top \in \mathbb{R}^M$, with $\boldsymbol{\phi}_t^\pi := \boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)$, the linear Q-learning update rule at each time step $t \in \mathbb{N}$ is given by

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t \left\{ R_t^\pi + \gamma \max_{a\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a)^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^{\pi\top} \widehat{\boldsymbol{\theta}}_t \right\} \boldsymbol{\phi}_t^\pi.$$

State-Action-Reward-State-Action (SARSA). SARSA is an on-policy algorithm that learns an optimal policy through data obtained from a sequence of evolving behavioral policies. Similar to

Q-learning, in the tabular setting, SARSA maintains a look-up table $\widehat{Q}_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ which stores the estimated action-value function for all state-action pairs. Given initial state $\boldsymbol{S}_0$, initial estimate $\widehat{Q}_0$ and initial policy $\pi_0 := \mathcal{I}(\widehat{Q}_0)$ where $\mathcal{I}$ denotes a policy improvement operator (e.g., ϵ-greedy, softmax), SARSA proceeds as follows. At iteration $t \in \mathbb{N} \cup \{0\}$, the algorithm starts from state $\boldsymbol{S}_t^{\pi_{t-2}}$ (generated under policy π_{t-2}) and selects action $A_t^{\pi_{t-1}}$ according to policy π_{t-1}. The algorithm then observes reward $R_t = r(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})$ and transitions to the next state $\boldsymbol{S}_{t+1}^{\pi_{t-1}}$. Finally, it updates the policy to $\pi_t = \mathcal{I}(\widehat{Q}_t)$ and selects the next action $A_{t+1}^{\pi_t}$ under this new policy. By convention, we use $\pi_{-1} := \pi_0$ and $\boldsymbol{S}_0^{\pi_{-2}} := \boldsymbol{S}_0$. In particular, the Q-function estimate is updated via

$$\widehat{Q}_{t+1}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}) = \widehat{Q}_t(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}) + \beta_t \delta_t^S,$$
$$\delta_t^S := R_t + \gamma \widehat{Q}_t(\boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t}) - \widehat{Q}_t(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})$$

where $(\beta_t)_{t \geq 0}$ is a sequence of step-sizes. Crucially, the data $(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}, R_t, \boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t})$ are generated by the evolving policy sequence. The key difference from Q-learning is that the update uses the action $A_{t+1}^{\pi_t}$ selected under the behavioral policy π_t rather than the greedy action $\arg\max_{a \in \mathcal{A}} \widehat{Q}_t(\boldsymbol{S}_{t+1}^{\pi_{t-1}}, a)$. This makes SARSA an on-policy algorithm, as the Q-function estimate corresponds to the policy currently generating the data.

As with Q-learning, the tabular setting can be viewed as a special case of linear function approximation with one-hot encoding basis functions $\{\boldsymbol{e}(\boldsymbol{s}, a) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}_{\boldsymbol{s} \in \mathcal{S}, a \in \mathcal{A}}$. Let $\boldsymbol{e}_t := \boldsymbol{e}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})$, then tabular SARSA update admits the following update rule

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t (R_t + \gamma \boldsymbol{e}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{e}_t^\top \widehat{\boldsymbol{\theta}}_t) \boldsymbol{e}_t.$$

Generalizing to a broader class of basis functions $\boldsymbol{\phi}(\boldsymbol{s}, a) = [\phi_1(\boldsymbol{s}, a), \ldots, \phi_M(\boldsymbol{s}, a)]^\top \in \mathbb{R}^M$ with $\boldsymbol{\phi}_t := \boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})$, the linear SARSA update rule at each time step $t \in \mathbb{N}$ is given by

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t (R_t + \gamma \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t) \boldsymbol{\phi}_t$$

where the data $(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}, R_t, \boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t})$ are generated by a sequence of time-varying policies $\pi_k := \mathcal{I}\{\boldsymbol{\phi}(\boldsymbol{s}, a)^\top \widehat{\boldsymbol{\theta}}_k\}$ for $k \in \{t-2, t-1, t\}$. After each update, we set $\pi_{t+1} = \mathcal{I}\{\boldsymbol{\phi}(\boldsymbol{s}, a)^\top \widehat{\boldsymbol{\theta}}_{t+1}\}$.

Projection. In practice, Q-learning and SARSA are often implemented with a projection step to ensure numerical stability and theoretical tractability (Bhandari et al., 2018; Zou et al., 2019a; Zhang et al., 2023). We incorporate such a projection to ensure the iterates $(\widehat{\boldsymbol{\theta}}_n)_{n \in \mathbb{N}}$ remain in a Euclidean ball of radius $r > 0$. Specifically, the projected Q-learning update becomes

$$\widehat{\boldsymbol{\theta}}_{t+1} = \Pi_r[\widehat{\boldsymbol{\theta}}_t + \beta_t \{R_t^\pi + \gamma \max_{a \in \mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a)^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\} \boldsymbol{\phi}_t]$$

where the projection operator is defined as $\Pi_r(\boldsymbol{\theta}) := \begin{cases} r\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2 & \text{if } \|\boldsymbol{\theta}\|_2 > r \\ \boldsymbol{\theta} & \text{otherwise} \end{cases}$. Similarly, the projected linear SARSA updates according to

$$\widehat{\boldsymbol{\theta}}_{t+1} = \Pi_r[\widehat{\boldsymbol{\theta}}_t + \beta_t \{R_t + \gamma \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\} \boldsymbol{\phi}_t].$$

Projection is widely used in RL and optimization as a computationally cheap way to ensure boundedness while preserving convergence (Nemirovski et al., 2009; Bubeck, 2015; Bhandari et al., 2018; Zou et al., 2019a; Xu et al., 2019; Xu & Gu, 2020; Wu et al., 2020). We therefore focus on projected linear Q-learning and SARSA.

Sensitivity to Step-Size. Step-size selection remains a fundamental challenge in reinforcement learning, with its importance well-documented in the literature (Sutton & Barto, 1998; Bertsekas et al., 2004; Szepesvári, 2022; Powell, 2022). Figures 1 and 2 illustrate the numerical stability and learning efficiency of the aforementioned algorithms on two standard benchmark environments. Both Q-learning and SARSA maintain stable performance below certain thresholds, approximately $\beta \approx 1.0$ for Cliff Walking and $\beta \approx 0.8$ for Taxi, beyond which performance degrades sharply. However, conservative step-size choices incur substantial learning inefficiency. Figure 2 further shows that, at fixed iteration budgets, increasing the step-size from conservative values toward the stability threshold substantially improves cumulative reward, whereas excessively large step-sizes

lead to sharp performance degradation. This presents a fundamental trade-off: conservative step-sizes guarantee algorithmic stability at the expense of learning efficiency, while aggressive step-sizes enable rapid progress but introduce the risk of divergence. Consequently, practitioners must engage in careful hyperparameter tuning, a process frequently neglected in practice (Powell, 2022). We address this limitation by proposing variants of Q-learning and SARSA that exhibit substantially improved robustness to step-size selection, permitting larger step-sizes without performance degradation while maintaining convergence guarantees.

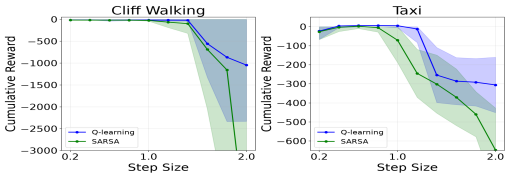

Figure 1: Sensitivity of Q-learning and SARSA to step-size selection.

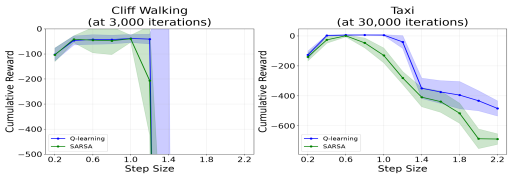

Figure 2: Learning efficiency of Q-learning and SARSA to step-size selection.

3 Policy Control via Implicit Procedures

Implicit update schemes serve as a powerful approach to enhance the numerical stability of stochastic approximation algorithms. The core principle involves reformulating the standard recursive update as a fixed-point equation where the future iterate appears on both the left- and right-hand sides of the algorithm update rule. In the context of SGD with an objective function f that one wishes to minimize, the implicit SGD iterates admit the following update rule $\boldsymbol{\theta}_{t+1}^{\text{im}} := \boldsymbol{\theta}_t^{\text{im}} + \beta_t \nabla f(\boldsymbol{\theta}_{t+1}^{\text{im}}; \xi_t)$ for a step-size β_t with a stochastic component ξ_t. It evaluates the gradient at the future iterate rather than the current one, as proposed and analyzed by (Toulis & Airoldi, 2015; Toulis et al., 2016; 2014; Toulis & Airoldi, 2017; Chee et al., 2023). This seemingly subtle modification yields significant stability improvements by providing adaptive shrinkage that automatically adjusts the effective step-size based on gradient information. These implicit procedures have also demonstrated substantial practical benefits in resolving numerical instabilities inherent in policy evaluation (Kim et al., 2025b;a), motivating their adoption in policy control algorithms.

Self-Tuned Q-Learning. Inspired by developments in the aforementioned implicit procedures, we propose a numerically stable variant of Q-learning, which we term self-tuned Q-learning. Recall, in the standard linear Q-learning, t^{th} update direction is $\{R_t^\pi + \gamma \max_{a' \in \mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\}\boldsymbol{\phi}_t$. Instead, we propose using the future iterate $\widehat{\boldsymbol{\theta}}_{t+1}$ in place of $\widehat{\boldsymbol{\theta}}_t$ when computing the current state-action value, i.e., $\{R_t^\pi + \gamma \max_{a' \in \mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_{t+1}\}\boldsymbol{\phi}_t$. This reformulation transforms the recursive update into a fixed-point equation, where the future iterate appears on both sides of the update rule. Combining terms involving the future iterate from both sides of the standard Q-learning update, we obtain the self-tuned Q-learning update rule:

$$(I + \beta_t \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top)\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t \{R_t^\pi + \gamma \max_{a' \in \mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t\}\boldsymbol{\phi}_t. \tag{1}$$

Self-Tuned SARSA. Following the same principle of implicit updates, we propose self-tuned SARSA, a numerically stable variant of SARSA in which the current state-action value is evaluated at the next iterate. In standard SARSA, the tth update direction is $(R_t + \gamma \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t)\boldsymbol{\phi}_t$, while we propose moving along $(R_t + \gamma \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_{t+1})\boldsymbol{\phi}_t$. This modification similarly transforms the SARSA update into a fixed-point equation, which leads to the self-tuned SARSA update rule:

$$(I + \beta_t \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top)\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t (R_t + \gamma \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t)\boldsymbol{\phi}_t. \tag{2}$$

Both self-tuned Q-learning and SARSA share the structure of incorporating the future iterate, differing only in action selection: Q-learning uses greedy maximization while SARSA follows the policy $\pi_t = \mathcal{I}\{\boldsymbol{\phi}(\boldsymbol{s}, a)^\top \widehat{\boldsymbol{\theta}}_t\}$. This formulation enhances stability through adaptive, feature-dependent step-size adjustment. Applying the Sherman-Morrison formula to (1) and (2) yields an explicit update

with a data-adaptive step-size that scales inversely with the squared feature norm, providing automatic shrinkage and improved stability over standard Q-learning and SARSA, as formally stated in the proposition below.

Proposition 3.1. *For all $t \in \mathbb{N}$, the self-tuned Q-learning update rule is equivalent to*

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t\{R_t^\pi + \gamma \max_{a' \in \mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\}\boldsymbol{\phi}_t/(1 + \beta_t\|\boldsymbol{\phi}_t\|_2^2)$$

and the self-tuned SARSA update rule is equivalent to

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t)\boldsymbol{\phi}_t/(1 + \beta_t\|\boldsymbol{\phi}_t\|_2^2).$$

4 Non-Asymptotic Analysis of Self-tuned Q-learning and SARSA

In this section, we provide finite-time error bounds for both the standard algorithms and their self-tuned counterparts under decreasing step-size sequences. Our theoretical analysis reveals a unifying message: self-tuned variants admit substantially greater flexibility in step-size selection while maintaining finite-time error guarantees without sacrificing learning efficiency, thereby resolving the numerical instability issues of standard algorithms described in Section 2.

4.1 Finite-time Analysis of Q-learning

To facilitate the theoretical analysis of linear Q-learning, we first assume uniform geometric ergodicity of the underlying Markov chain $\{\boldsymbol{S}_t^\pi\}_{t\geq 0}$ induced by a policy π.

Assumption 4.1 (Uniform geometric ergodicity)**.** Given a behavioral policy π, the Markov chain $\{\boldsymbol{S}_t^\pi\}_{t\geq 0}$ is uniformly geometrically ergodic with a stationary distribution μ^π, i.e., there exist constants $m > 0$ and $\rho \in (0, 1)$ such that $\sup_{\boldsymbol{s}\in\mathcal{S}} d_{TV}\{\mathbb{P}(\boldsymbol{S}_t^\pi \in \cdot \mid \boldsymbol{S}_0^\pi = \boldsymbol{s}), \mu^\pi\} \leq m\rho^t, \forall t \geq 0$ where $d_{TV}(P, Q)$ denotes the total variation distance between the probability distributions P and Q.

Assumption 4.2 (Solvability)**.** There exists $w_q > 0$ such that

$$\boldsymbol{\Sigma}^{\mu^\pi} - \gamma^2\boldsymbol{\Sigma}^\star(\boldsymbol{\theta}) \succeq w_q\mathbf{I} \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^M,$$

where

$$\boldsymbol{\Sigma}^{\mu^\pi} := \mathbb{E}^{\mu^\pi}\{\boldsymbol{\phi}(\boldsymbol{S}, A)\boldsymbol{\phi}(\boldsymbol{S}, A)^\top\}, \qquad \boldsymbol{\Sigma}^\star(\boldsymbol{\theta}) := \mathbb{E}^{\mu^\pi}\{\boldsymbol{\phi}(\boldsymbol{S}, A_{\boldsymbol{S}}^{\boldsymbol{\theta}})\boldsymbol{\phi}(\boldsymbol{S}, A_{\boldsymbol{S}}^{\boldsymbol{\theta}})^\top\},$$

and $A_{\boldsymbol{S}}^{\boldsymbol{\theta}} := \arg\max_{a\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}, a)^\top\boldsymbol{\theta}$.

Here, $\mathbb{E}^{\mu^\pi}$ denotes the expectation where $\boldsymbol{S}$ follows the stationary distribution μ^π, A is sampled according to the behavioral policy $\pi(\cdot|\boldsymbol{S})$, and transition to $\boldsymbol{S}'$ according to P. Under Assumptions 4.1–4.2, Melo et al. (2008) established asymptotic convergence of Q-learning iterates to the limit point $\boldsymbol{\theta}^\star$, and Chen et al. (2019) provided finite-time error bounds under a restrictive step-size condition. The limit point $\boldsymbol{\theta}^\star$ of Q-learning satisfies $\mathbb{E}^{\mu^\pi}[\phi(\boldsymbol{S}, A)\{r(\boldsymbol{S}, A) + \gamma \max_{a\in\mathcal{A}} \phi(\boldsymbol{S}', a)^\top\boldsymbol{\theta}^\star - \phi(\boldsymbol{S}, A)^\top\boldsymbol{\theta}^\star\}] = 0$. We present finite-time error bounds for projected linear Q-learning and its self-tuned variant under decreasing step-sizes, demonstrating self-tuned Q-learning's superior robustness to step-size selection.

Theorem 4.3 (Finite-time bounds for Q-learning under decreasing step-size)**.** *Let $r > 0$ satisfy $\|\boldsymbol{\theta}^\star\|_2 \leq r$ and $\{\beta_t\}_{t\geq 0}$ be a decreasing step-size sequence of the form $\beta_t = \frac{\beta_0}{(t+1)^s}$, where $s \in (0.5, 1)$. Under Assumptions 4.1 and 4.2, if $\beta_0 \in (0, 1/w_q)$, projected linear Q-learning satisfies*

$$\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \lesssim \exp\left[\frac{-\beta_0 w_q}{(1-s)}\{(t+1)^{1-s} - 1\}\right]\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + \tilde{\mathcal{O}}\left(t^{-s}\right),$$

where $\tilde{\mathcal{O}}$ suppresses polylogarithmic factors. In comparison, the same bound holds for projected self-tuned Q-learning with any $\beta_0 > 0$.

Remark 4.4 (Advantage of self-tuned Q-learning). Standard Q-learning requires $\beta_0 \in (0, 1/w_q)$, whereas the bound for self-tuned Q-learning holds for any $\beta_0 > 0$. Consequently, self-tuned Q-learning accommodates arbitrarily large initial step-sizes, avoiding the early-stage numerical

instability that can arise in standard Q-learning. This flexibility is particularly useful during the transient phase: larger β_0 accelerates the decay of the initialization error, while the remaining error term stays of order $\widetilde{\mathcal{O}}(t^{-s})$, matching that of standard Q-learning. Thus, the primary advantage of self-tuned Q-learning is not a faster asymptotic rate, but the removal of the initial step-size restriction, allowing aggressive initialization without sacrificing the convergence rate.

Remark 4.5 (Convergence to a bounded set). While our analysis focuses on the standard Q-learning setting, one can also consider a self-tuning mechanism with an ϵ-softmax behavioral policy and adaptive temperature. Following (Meyn, 2024; Liu et al., 2025), convergence of the iterates to a bounded set can be established without stringent conditions on the step-size or Assumption 4.2.

4.2 Finite-time Analysis of SARSA

We now establish finite-time error bounds for the self-tuned SARSA algorithm under decreasing step-sizes. The analysis is more intricate than that of Q-learning as the time-varying behavioral policy induces a non-stationary Markov chain, precluding direct application of Assumption 4.1 and 4.2. Instead, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, we assume a uniform geometric ergodicity of the Markov chain $\{\boldsymbol{S}_t^{\pi_{\boldsymbol{\theta}}}\}_{t\geq 0}$ induced by the policy $\pi_{\boldsymbol{\theta}}$.

Assumption 4.6 (Uniform geometric ergodicity)**.** For any $\boldsymbol{\theta} \in \mathbb{R}^M$, a Markov chain $\{\boldsymbol{S}_t^{\pi_{\boldsymbol{\theta}}}\}_{t\geq 0}$ induced from the policy $\pi_{\boldsymbol{\theta}}$ is uniformly geometrically ergodic with a stationary distribution $\mu^{\boldsymbol{\theta}}$, i.e., there exist constants $m > 0$ and $\rho \in (0, 1)$ such that $\sup_{\boldsymbol{s}\in\mathcal{S}} d_{TV}\left\{\mathbb{P}(\boldsymbol{S}_t^{\pi_{\boldsymbol{\theta}}} \in \cdot \mid \boldsymbol{S}_0^{\pi_{\boldsymbol{\theta}}} = \boldsymbol{s}), \mu^{\boldsymbol{\theta}}\right\} \leq m\rho^t, \forall t \geq 0$ where $d_{TV}(P, Q)$ denotes the total variation distance between the probability distributions P and Q.

Let $\boldsymbol{\theta}^\star$ be the limit point of the SARSA and $\boldsymbol{A}^{\boldsymbol{\theta}^\star} := \mathbb{E}^{\mu^{\boldsymbol{\theta}^\star}}[\boldsymbol{\phi}(\boldsymbol{S}, A)\{\gamma\boldsymbol{\phi}(\boldsymbol{S}', A')^\top - \boldsymbol{\phi}(\boldsymbol{S}, A)^\top\}]$ where the expectation is taken with respect to the distribution induced as follows: the current state $\boldsymbol{S}$ is drawn from the stationary distribution $\mu^{\boldsymbol{\theta}^\star}$, the current action A is sampled according to the behavioral policy $\pi_{\boldsymbol{\theta}^\star}(\cdot \mid \boldsymbol{S})$, the next state $\boldsymbol{S}'$ is drawn according to the transition kernel $\mathsf{P}(\cdot \mid \boldsymbol{S}, A)$, and the next action A' is sampled from $\pi_{\boldsymbol{\theta}^\star}(\cdot \mid \boldsymbol{S}')$. Following the convergence analysis of SARSA in (Melo et al., 2008; Zou et al., 2019a), we restrict our attention to the class of Lipschitz policies.

Assumption 4.7 (Lipschitz policy and solvability)**.** For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^M$ and $(\boldsymbol{s}, a) \in \mathcal{S} \times \mathcal{A}$, a behavioral policy $\pi_{\boldsymbol{\theta}}(a|\boldsymbol{s})$ satisfies $|\pi_{\boldsymbol{\theta}_1}(a|\boldsymbol{s}) - \pi_{\boldsymbol{\theta}_2}(a|\boldsymbol{s})| \leq C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ for some $C > 0$. Furthermore, the Lipschitz constant $C > 0$ is small enough that the matrix $\boldsymbol{A}^{\boldsymbol{\theta}^\star} + C\lambda\boldsymbol{I}$ is negative definite, where $\lambda := (b + 2r)|\mathcal{A}|(2 + \lceil\log_\rho \frac{1}{m}\rceil + \frac{1}{1-\rho})$.

We next present finite-time error bounds for projected linear SARSA and its self-tuned variant under decreasing step-sizes.

Theorem 4.8 (Finite-time bounds for SARSA under decreasing step-size)**.** *Let $r > 0$ satisfy $\|\boldsymbol{\theta}^\star\|_2 \leq r$ and $\{\beta_t\}_{t\geq 0}$ be a decreasing step-size sequence of the form $\beta_t = \frac{\beta_0}{(t+1)^s}$, where $s \in (0.5, 1)$. Under Assumptions 4.6 and 4.7, if $\beta_0 \in (0, \frac{1}{2w_s})$, projected linear SARSA yields*

$$\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \lesssim \exp\left[\frac{-2\beta_0 w_s}{(1-s)}\{(t+1)^{1-s} - 1\}\right]\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + \tilde{\mathcal{O}}\left(t^{-s}\right).$$

In comparison, if at least four basis functions are used, the same bound holds for projected self-tuned SARSA for any $\beta_0 > 0$.

Remark 4.9 (Advantage of self-tuned SARSA). Standard SARSA requires $\beta_0 \in (0, \frac{1}{2w_s})$, whereas self-tuned SARSA requires only that at least four basis functions be used, a mild condition satisfied by virtually any practical feature representation. Moreover, this condition is sufficient but not necessary; see Lemma C.7. Under this condition, the error bound holds for any $\beta_0 > 0$. Self-tuned SARSA therefore accommodates arbitrarily large initial step-sizes, avoiding the early-stage numerical instability that large step-sizes can induce in standard SARSA. This flexibility motivates the use of large initial step-sizes without meticulous tuning: increasing β_0 accelerates the decay of the initialization error, while the remaining error term stays of order $\widetilde{\mathcal{O}}(t^{-s})$, matching that of standard SARSA. Although logarithmic factors are suppressed by the $\widetilde{\mathcal{O}}$ notation, our bound improves upon that of Zou et al. (2019a) by a logarithmic factor. Thus, the primary advantage of self-tuned SARSA is not a faster asymptotic rate, but the removal of the initial step-size restriction, enabling aggressive initialization without sacrificing the convergence rate.

Remark 4.10 (Comparison to other implicit frameworks). Compared with (Toulis & Airoldi, 2015; 2017), our analysis must account for the additional dependence induced by Markovian observations. We handle this dependence using information-theoretic techniques in the spirit of Bhandari et al. (2018). The policy-control setting introduces a further complication: the behavioral policy evolves over time, resulting in a time-inhomogeneous Markov structure that does not arise in the policy-evaluation setting considered by Kim et al. (2025b). To address this issue, we leverage the auxiliary Markov-chain technique developed by Zou et al. (2019a). Consequently, our analysis extends the implicit framework to a substantially more general reinforcement-learning setting involving both Markovian dependence and time-varying behavioral policies.

5 Numerical Experiments

In this section, we evaluate self-tuned Q-learning and self-tuned SARSA on ten benchmark environments, several of which are drawn from the open-source `gymnasium` library in Python (Towers et al., 2024). Throughout all experiments, Q-learning employs an ϵ-greedy behavioral policy: with probability ϵ, a random action is selected; otherwise, the greedy action $\arg\max_a \widehat{Q}(s,a)$ is chosen. SARSA uses an ϵ-softmax policy: $\pi(a|s) = \epsilon/|\mathcal{A}| + (1-\epsilon)\exp\{\widehat{Q}(s,a)/\iota\}/\sum_{a'\in\mathcal{A}}\exp\{\widehat{Q}(s,a')/\iota\}$, where $\iota = 0.05$ is the temperature parameter controlling the softness of the action distribution. For each experimental setting, we conducted 50 independent runs on an Apple M2 Mac mini with 8 GB of RAM. Detailed experimental settings are provided in the appendix, along with additional experiments.

5.1 Discrete State Environments

We evaluate performance on six discrete state environments with finite state-action spaces represented using one-hot feature vectors. A constant step-size is used for Cliff Walking and Taxi, while a decreasing step-size of the form $\beta_t = \beta_0/(t+1)^{2/3}$ is used for Blackjack, Frozen Lake, Gridworld, and Dynamic Pricing. Figures 3 and 4 demonstrate that our proposed self-tuned variants resolve the stability-efficiency trade-off. Compared to Q-learning (blue) and SARSA (green), which exhibit sharp performance degradation beyond their thresholds, self-tuned Q-learning (orange) and self-tuned SARSA (red) maintain stable performance across the entire range of step-sizes tested, remaining robust even at $\beta = 2.0$ where standard methods fail catastrophically. Figure 4 further shows that at fixed iteration counts, self-tuned variants achieve strong performance at aggressive step-sizes where standard methods collapse, enabling practitioners to leverage large step-sizes to accelerate learning. Figure 5 additionally demonstrates that self-tuned variants consistently outperform their standard counterparts across a wide range of initial step-sizes $\beta_0 \in (0, 300]$. While Q-learning (blue) and SARSA (green) exhibit high sensitivity to the choice of β_0, with performance deteriorating significantly as it increases, self-tuned Q-learning (orange) and self-tuned SARSA (red) maintain stable and superior final cumulative rewards across all environments.

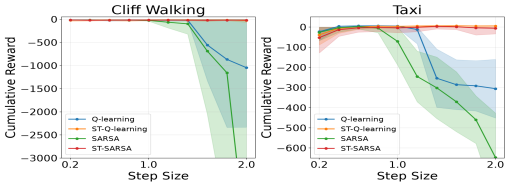

Figure 3: Numerical stability of self-tuned Q-learning/SARSA to constant step-size choice.

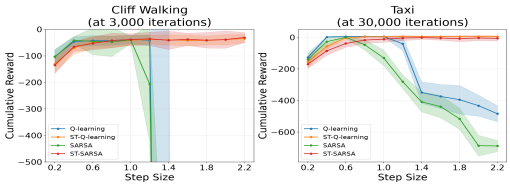

Figure 4: Learning efficiency of self-tuned Q-learning/SARSA to constant step-size choice.

5.2 Continuous State Environments

We evaluate the self-tuned algorithms on the Acrobot and Mountain Car environments, where we use linear function approximation to cope with continuous state spaces. We employ a decreasing step-size schedule $\beta_t = \beta_0/(t+1)^{2/3}$ and sweep the initial step-size over $\beta_0 \in \{0.1, 2, 4, 6, 8, 10\}$. We also include AlphaBound (Dabney & Barto, 2012), an adaptive step-size method that requires no step-size input, as a reference. Figure 6 reports the final log episode length (left two panels), i.e., the number

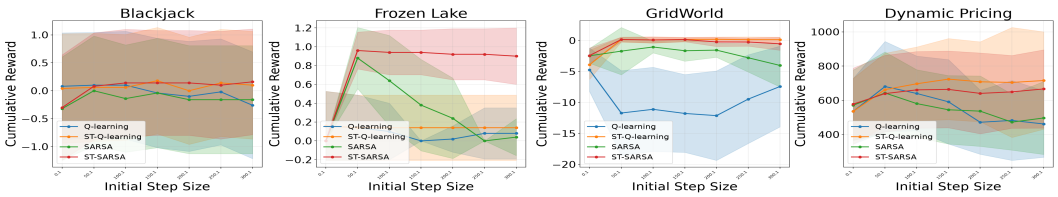

Figure 5: Sensitivity of self-tuned and standard Q-learning/SARSA to the initial step-size.

of steps needed to complete the task, so shorter is better, and the episode-length difference between each self-tuned method and its standard counterpart (right two panels), where negative values favor the self-tuned method. With a small initial step-size ($\beta_0 = 0.1$), all methods learn slowly and remain at long episode lengths, highlighting the cost of conservative step-sizes. As β_0 increases, standard SARSA fails to improve in both environments, whereas self-tuned SARSA reduces episode length by roughly 600 steps in Acrobot and 800 steps in Mountain Car for all $\beta_0 \geq 4$. Standard Q-learning performs well for moderate step-sizes but degrades as β_0 grows, sharply in Acrobot for $\beta_0 \geq 8$ and with increasing variability in Mountain Car, while self-tuned Q-learning remains stable across the entire range. Both self-tuned methods match or outperform AlphaBound once $\beta_0 \geq 4$. These results confirm that self-tuned algorithms are robust to step-size selection and can exploit large initial step-sizes for efficient policy learning.

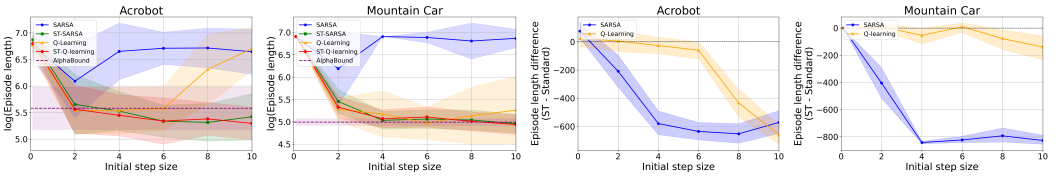

Figure 6: Performance on continuous state environments with linear function approximation.

5.3 Extensions to deep Q-learning

We next examine whether the proposed self-tuning mechanism remains effective beyond linear function approximation by applying it to deep Q-learning (Mnih et al., 2015). In the nonlinear setting, uniqueness of an optimal parameterization is generally not guaranteed; nevertheless, the self-tuned update can still stabilize aggressive learning and rapidly reach parameter regions associated with high reward. We compare standard deep Q-learning and self-tuned deep Q-learning with Q-AMSGrad (Weng et al., 2020). A more detailed description of how the self-tuned framework is adapted to the deep Q-learning setting is provided in the Appendix G. Figure 9 in Appendix G shows that the advantage of self-tuning becomes more pronounced as the initial step-size increases. In particular, for $\beta_0 \in \{2, 4\}$, self-tuned deep Q-learning improves reward more rapidly and generally attains stronger performance than both standard deep Q-learning and Q-AMSGrad. These results provide empirical evidence that the practical benefits of self-tuning extend beyond linear function approximation, suggesting that the proposed mechanism can remain effective in nonlinear deep reinforcement-learning settings.

6 Conclusion

This work addresses the persistent challenge of step-size sensitivity in Q-learning and SARSA. We reformulate these algorithms as fixed-point equations that evaluate state-action values at future iterates, which induces a data-adaptive regularization that scales inversely with the feature norm. Our theoretical analysis shows that the proposed methods remain stable under arbitrarily large initial step size, in sharp contrast to the restrictive step-size requirements of the standard algorithms. Across diverse environments, the self-tuned variants learn stably and efficiently, succeeding where standard methods fail and eliminating the need for step-size tuning.

AI USE STATEMENT

AI tools were used to assist with language editing and code debugging. Specifically, they were used to improve the clarity and presentation of the manuscript text and to help identify and correct implementation and programming issues. All methodological choices, theoretical developments, experimental design, analysis, and interpretation of results were carried out and verified by the authors.

ETHICS STATEMENT

This work focuses on methodological improvements to existing reinforcement learning algorithms and their theoretical analysis. The study does not involve human subjects, personal or sensitive data, or data collected from individuals. Accordingly, we do not identify any human-subject or privacy-related ethical concerns associated with this research.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility of both the theoretical and empirical results. The proposed self-tuned Q-learning and SARSA algorithms are fully specified in Section 3, while Section 4 states the assumptions and finite-time guarantees used in the theoretical analysis. Complete proofs, including the supporting lemmas, are provided in Appendices A, B and C. The experimental protocol is described in Section 5, with detailed environment descriptions, feature representations, step-size schedules, hyperparameters, training budgets, and numbers of independent runs provided in Appendix E; additional experimental results are reported in Appendix F. Appendix G provides the deep Q-learning implementation details and additional results. Finally, complete source code for reproducing the experiments is provided through the anonymous repository linked in the paper.

REFERENCES

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37, 1995.

Dimitri P Bertsekas, Vivek S Borkar, and Angelia Nedic. Improved temporal difference methods with linear function approximation. *Learning and Approximate Dynamic Programming*, pp. 231–255, 2004.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pp. 1691–1692. PMLR, 2018.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

Jerry Chee, Hwanwoo Kim, and Panos Toulis. "plus/minus the learning rate": Easy and scalable statistical inference with SGD. In *International Conference on Artificial Intelligence and Statistics*, pp. 2285–2309. PMLR, 2023.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, pp. 4, 2019.

William Dabney and Andrew Barto. Adaptive step-size for online temporal difference learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 872–878, 2012.

Geoffrey J Gordon. Chattering in sarsa (λ). *CMU Learning Lab Technical Report*, 1996.

Geoffrey J Gordon. Reinforcement learning with function approximation converges to a region. *Advances in neural information processing systems*, 13, 2000.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Hwanwoo Kim, Dongkyu Derek Cho, and Eric Laber. Implicit updates for average-reward temporal difference learning. *arXiv preprint arXiv:2510.06149*, 2025a.

Hwanwoo Kim, Panos Toulis, and Eric Laber. Stabilizing temporal difference learning via implicit stochastic recursion. *arXiv preprint arXiv:2505.01361*, 2025b.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Xinyu Liu, Zixuan Xie, and Shangtong Zhang. Linear Q-learning does not diverge in L^2: Convergence rates to a bounded set. In *Proceedings of the International Conference on Machine Learning*, 2025.

Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 664–671, 2008.

Sean Meyn. The projected bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 69(12):8323–8337, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Warren B Powell. *Reinforcement learning and stochastic optimization: A unified framework for sequential decisions*. John Wiley & Sons, 2022.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 29:70–76, 2017.

Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308, 2000.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*, volume 1. MIT Press, 1998.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

Panos Toulis and Edoardo M. Airoldi. Implicit stochastic gradient descent. *arXiv preprint*, 2015.

Panos Toulis and Edoardo M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Panos Toulis, Edoardo Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In *Proceedings of the International Conference on Machine Learning*, volume 32, pp. 667–675. PMLR, June 2014.

Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pp. 1290–1298. PMLR, 2016.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun K. G., et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint*, 2024.

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3): 185–202, 1994.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

C.J Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

C.J.C.H Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, Cambridge, England, 1989.

Bowen Weng, Huaqing Xiong, Yingbin Liang, and Wei Zhang. Analysis of q-learning with adaptation and momentum restart for gradient descent. *arXiv preprint arXiv:2007.07422*, 2020.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International conference on machine learning*, pp. 10555–10565. PMLR, 2020.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. *Advances in Neural Information Processing Systems*, 32, 2019.

Shangtong Zhang, Remi Tachet Des Combes, and Romain Laroche. On the convergence of SARSA with linear function approximation. In *International Conference on Machine Learning*, pp. 41613–41646. PMLR, 2023.

Yixuan Zhang and Qiaomin Xie. Constant stepsize q-learning: Distributional convergence, bias and extrapolation. *arXiv preprint arXiv:2401.13884*, 2024.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. *Advances in Neural Information Processing Systems*, 32, 2019a.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa and q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234v1*, 2019b.

A Proofs of Proposition 3.1

Proposition 3.1 (Self-tuned Q-learning update). The self-tuned Q-learning algorithm admits the following update rule:

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)\|_2^2}\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi).$$

and the self-tuned SARSA algorithm admits the following update rule:

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})\|_2^2}\left\{R_t + \gamma\boldsymbol{\phi}(\boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t})^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}).$$

Proof. The self-tuned Q-learning update is defined as

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)^\top \widehat{\boldsymbol{\theta}}_{t+1}\right\}\boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi). \tag{3}$$

(3) is equivalent to the fixed-point equation

$$\left(I + \beta_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\right)\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}_t \tag{4}$$

where we denote $\boldsymbol{\phi}_t := \boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)$ for notational convenience. Using the Sherman-Morrison-Woodbury formula, we have

$$\left(I + \beta_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\right)^{-1} = I - \frac{\beta_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}.$$

Solving for $\widehat{\boldsymbol{\theta}}_{t+1}$, we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{t+1} &= \left(I - \frac{\beta_t \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\right)\left[\widehat{\boldsymbol{\theta}}_t + \beta_t\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}_t\right] \\
&= \widehat{\boldsymbol{\theta}}_t + \beta_t\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}_t - \frac{\beta_t \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\boldsymbol{\phi}_t \\
&\quad - \frac{\beta_t^2\|\boldsymbol{\phi}_t\|_2^2}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \beta_t\left(1 - \frac{\beta_t\|\boldsymbol{\phi}_t\|_2^2}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\right)\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}_t - \frac{\beta_t}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}(\boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t)\boldsymbol{\phi}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\left\{R_t^\pi + \gamma \max_{a'\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a')^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\right\}\boldsymbol{\phi}_t,
\end{aligned}$$

which completes the proof for the self-tuned Q-learning.

The self-tuned SARSA update is defined as

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t\left\{R_t + \gamma\boldsymbol{\phi}(\boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t})^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})^\top \widehat{\boldsymbol{\theta}}_{t+1}\right\}\boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}). \quad (5)$$

Let $\boldsymbol{\phi}_t := \boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})$ and $\boldsymbol{\phi}_{t+1} := \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t})$. Then (5) yields the following fixed-point equation

$$\left(I + \beta_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\right)\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t\left(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{\phi}_t.$$

Solving for $\widehat{\boldsymbol{\theta}}_{t+1}$ with the Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{t+1} &= \left(I - \frac{\beta_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\right)\left\{\widehat{\boldsymbol{\theta}}_t + \beta_t\left(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{\phi}_t\right\} \\
&= \widehat{\boldsymbol{\theta}}_t + \beta_t\left(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{\phi}_t - \frac{\beta_t(\boldsymbol{\phi}_t^\top\widehat{\boldsymbol{\theta}}_t)\boldsymbol{\phi}_t}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2} - \frac{\beta_t^2\|\boldsymbol{\phi}_t\|_2^2}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\left(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{\phi}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \beta_t\left(1 - \frac{\beta_t\|\boldsymbol{\phi}_t\|_2^2}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\right)\left(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{\phi}_t - \frac{\beta_t}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}(\boldsymbol{\phi}_t^\top\widehat{\boldsymbol{\theta}}_t)\boldsymbol{\phi}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{\phi}_t\|_2^2}\left(R_t + \gamma\boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{\phi}_t,
\end{aligned}$$

which completes the proof. □

B Proofs of Q-Learning Results

B.1 Preliminary Lemmas

For any $\boldsymbol{\theta} \in \mathbb{R}^M$, let us define

$$\begin{aligned}
f_t(\boldsymbol{\theta}) &= \boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)\left\{R_t^\pi + \gamma \max_{a\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^\pi, a)^\top\boldsymbol{\theta} - \boldsymbol{\phi}(\boldsymbol{S}_t^\pi, A_t^\pi)^\top\boldsymbol{\theta}\right\} \\
\bar{f}(\boldsymbol{\theta}) &= \mathbb{E}^{\mu^\pi}\left[\boldsymbol{\phi}(\boldsymbol{S}^\pi, A^\pi)\left\{r(\boldsymbol{S}^\pi, A^\pi) + \gamma \max_{a\in\mathcal{A}} \boldsymbol{\phi}(\boldsymbol{S}', a)^\top\boldsymbol{\theta} - \boldsymbol{\phi}(\boldsymbol{S}^\pi, A^\pi)^\top\boldsymbol{\theta}\right\}\right],
\end{aligned}$$

where $\mathbb{E}^{\mu^\pi}$ denotes the expectation where $\boldsymbol{S}^\pi$ follows the stationary distribution μ^π, A^π is sampled according to the behavioral policy $\pi(\cdot|\boldsymbol{S}^\pi)$, and $\boldsymbol{S}' \sim \mathsf{P}(\cdot \mid \boldsymbol{S}^\pi, A^\pi)$. Let us also define

$$\Lambda_t(\boldsymbol{\theta}) := \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star, f_t(\boldsymbol{\theta}) - \bar{f}(\boldsymbol{\theta})\rangle,$$

where $\boldsymbol{\theta}^\star$ is the solution of $\bar{f}(\boldsymbol{\theta}) = 0$.

Lemma B.1 (Lemma 11 of (Zou et al., 2019b)). *For any $\boldsymbol{\theta} \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}\|_2 \leq r$,*

$$\|f_t(\boldsymbol{\theta})\|_2 \leq G := b + 2r \text{ for all } t \geq 0.$$

Lemma B.2 (Lemma 12 of (Zou et al., 2019b)). *For any $\boldsymbol{\theta} \in \mathbb{R}^M$,*

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star, \bar{f}(\boldsymbol{\theta}) - \bar{f}(\boldsymbol{\theta}^\star) \rangle \leq -\frac{w_q \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2}{2},$$

where $w_q > 0$ is the constant in Assumption 4.2.

Lemma B.3 (Lemma 13 of (Zou et al., 2019b)). *For any $\boldsymbol{\theta} \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}\|_2 \leq r$, $|\Lambda_t(\boldsymbol{\theta})| \leq 2G^2$ for all $t \geq 0$.*

Lemma B.4 (Lemma 14 of (Zou et al., 2019b)). *For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}_1\|_2 \leq r$, $\|\boldsymbol{\theta}_2\|_2 \leq r$,*

$$|\Lambda_t(\boldsymbol{\theta}_1) - \Lambda_t(\boldsymbol{\theta}_2)| \leq 6G\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

Lemma B.5. *Suppose the step-size sequence $\{\beta_i\}_{i\in\mathbb{N}}$ is constant, i.e., $\beta_i = \beta$, $\forall i \in \mathbb{N}$. Let $\tau_\beta := \min\{n \in \mathbb{N} : m\rho^n \leq \beta\}$, then for all $i \in \mathbb{N}$*

$$\mathbb{E}^{\mu^\pi}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\right\} \leq 4G^2 m\rho^{\tau_\beta} + 6\beta G^2\tau_\beta \leq 10\beta G^2\tau_\beta \tag{6}$$

for projected Q-learning iterates $(\widehat{\boldsymbol{\theta}}_i)_{i\in\mathbb{N}}$.

Proof. For all $i \geq 0$,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{i+1} - \widehat{\boldsymbol{\theta}}_i\right\|_2 &= \left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_i + \beta f_i(\widehat{\boldsymbol{\theta}}_i)\} - \Pi_r\widehat{\boldsymbol{\theta}}_i\right\|_2 \\
&\leq \left\|\widehat{\boldsymbol{\theta}}_i + \beta f_i(\widehat{\boldsymbol{\theta}}_i) - \widehat{\boldsymbol{\theta}}_i\right\|_2 && \text{(non-expansiveness of } \Pi_r\text{)} \\
&\leq \beta G && \text{(by Lemma B.1).}
\end{aligned}$$

When $i > \tau_\beta$,

$$\begin{aligned}
\left|\Lambda_i(\widehat{\boldsymbol{\theta}}_i) - \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_\beta})\right| &\leq 6G\left\|\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i-\tau_\beta}\right\|_2 && \text{(by Lemma B.4)} \\
&\leq 6G\left(\left\|\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i-1}\right\|_2 + \cdots + \left\|\widehat{\boldsymbol{\theta}}_{i-\tau_\beta+1} - \widehat{\boldsymbol{\theta}}_{i-\tau_\beta}\right\|_2\right) \\
&\leq 6G(\beta G + \cdots + \beta G) \\
&\leq 6\beta G^2\tau_\beta.
\end{aligned}$$

Therefore,

$$\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_\beta}) + 6\beta G^2\tau_\beta. \tag{7}$$

By the same logic, when $i \leq \tau_\beta$, we have

$$\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_0) + 6\beta G^2\tau_\beta.$$

Note that $\mathbb{E}^{\mu^\pi}\Lambda_i(\boldsymbol{\theta}) = 0$ for any fixed $\boldsymbol{\theta} \in \mathbb{R}^M$. Therefore, when $i \leq \tau_\beta$, $\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_0) = 0$ yields $\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq 6\beta G^2\tau_\beta$. Now, consider the other case when $i > \tau_\beta$, let $O_i := \{\boldsymbol{S}_i^\pi, A_i^\pi, \boldsymbol{S}_{i+1}^\pi\}$ and write $\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau_\beta}, O_i) = \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_\beta})$ to highlight its dependence on O_i. Consider random variables $\widehat{\boldsymbol{\theta}}'_{i-\tau_\beta}$ and O'_i drawn independently from the marginal distributions of $\widehat{\boldsymbol{\theta}}_{i-\tau_\beta}$ and O_i, respectively. Then we obtain

$$\mathbb{E}^{\mu^\pi}\left\{\Lambda(\widehat{\boldsymbol{\theta}}'_{i-\tau_\beta}, O'_i)\right\} = \mathbb{E}\left[\mathbb{E}\{\Lambda(\widehat{\boldsymbol{\theta}}'_{i-\tau_\beta}, O'_i) \mid \widehat{\boldsymbol{\theta}}'_{i-\tau_\beta}\}\right] = 0.$$

Furthermore, since

- $\widehat{\boldsymbol{\theta}}_{i-\tau_\beta} \to \boldsymbol{S}_{i-\tau_\beta} \to \boldsymbol{S}_i^\pi \to O_i$ forms a Markov chain
- $|\Lambda(\boldsymbol{\theta}, O_i)| \leq 2G^2$ for any $\|\boldsymbol{\theta}\|_2 \leq r$

Lemma 10 of (Bhandari et al., 2018) yields

$$\left|\mathbb{E}^{\mu^\pi}\{\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau_\beta}, O_i)\}\right| \leq 4G^2 m\rho^{\tau_\beta}. \tag{8}$$

Combining (7) and (8), we get

$$\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq 4G^2 m\rho^{\tau_\beta} + 6\beta G^2\tau_\beta \leq 4\beta G^2 + 6\beta G^2\tau_\beta \leq 10\beta G^2\tau_\beta.$$

□

Lemma B.6. *Suppose the step-size sequence $\{\beta_i\}_{i\in\mathbb{N}}$ is decreasing, i.e., $\beta_{i+1} < \beta_i$, $\forall i \in \mathbb{N}$. Fix $i \leq t$, and let $\tau_{\beta_t} := \min\{n \in \mathbb{N} : m\rho^n \leq \beta_t\}$, then*

$$\mathbb{E}^{\mu^\pi}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\right\} \leq 6G^2\tau_{\beta_t}\beta_0 \quad \textit{for} \quad i \leq \tau_{\beta_t}$$
$$\mathbb{E}^{\mu^\pi}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\right\} \leq 10G^2\tau_{\beta_t}\beta_{i-\tau_{\beta_t}} \quad \textit{for} \quad i > \tau_{\beta_t}$$

for projected Q-learning iterates $(\widehat{\boldsymbol{\theta}}_i)_{i\in\mathbb{N}}$.

Proof. For all $i \geq 0$,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{i+1} - \widehat{\boldsymbol{\theta}}_i\right\|_2 &= \left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_i + \beta_i f_i(\widehat{\boldsymbol{\theta}}_i)\} - \Pi_r\{\widehat{\boldsymbol{\theta}}_i\}\right\|_2 \\
&\leq \left\|\widehat{\boldsymbol{\theta}}_i + \beta_i f_i(\widehat{\boldsymbol{\theta}}_i) - \widehat{\boldsymbol{\theta}}_i\right\|_2 && \text{(non-expansiveness of } \Pi_r\text{)} \\
&\leq \beta_i G. && \text{(by Lemma B.1)}
\end{aligned}$$

When $i > \tau_{\beta_t}$,

$$\begin{aligned}
|\Lambda_i(\widehat{\boldsymbol{\theta}}_i) - \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}})| &\leq 6G\left\|\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}}\right\|_2 && \text{(by Lemma B.4)} \\
&\leq 6G\left(\left\|\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_{i-1}\right\|_2 + \cdots + \left\|\widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}+1} - \widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}}\right\|_2\right) \\
&\leq 6G(\beta_{i-1}G + \cdots + \beta_{i-\tau_{\beta_t}}G) \\
&\leq 6G^2\sum_{j=i-\tau_{\beta_t}}^{i-1}\beta_j.
\end{aligned}$$

Therefore,

$$\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}}) + 6G^2\sum_{j=i-\tau_{\beta_t}}^{i-1}\beta_j \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}}) + 6G^2\tau_{\beta_t}\beta_{i-\tau_{\beta_t}}. \tag{9}$$

When $i \leq \tau_{\beta_t}$, by the same logic,

$$\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_0) + 6G^2\sum_{j=0}^{i-1}\beta_j \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_0) + 6G^2\tau_{\beta_t}\beta_0.$$

Since $\mathbb{E}^{\mu^\pi}\Lambda_i(\boldsymbol{\theta}) = 0$ for any fixed $\boldsymbol{\theta} \in \mathbb{R}^M$, we know $\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_0) = 0$. For $i \leq \tau_{\beta_t}$, we thus obtain $\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq 6G^2\tau_{\beta_t}\beta_0$. When $i > \tau_{\beta_t}$, using the same argument as in Lemma B.5, we know $\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau_{\beta_t}}) \leq 4G^2\beta_t$. Combined with (9), we have

$$\mathbb{E}^{\mu^\pi}\Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq 4G^2\beta_t + 6G^2\tau_{\beta_t}\beta_{i-\tau_{\beta_t}} \leq 10G^2\tau_{\beta_t}\beta_{i-\tau_{\beta_t}}.$$

□

Lemma B.7. *Suppose the step-size sequence $\{\beta_i\}_{i\in\mathbb{N}}$ is constant, i.e., $\beta_i = \beta$, $\forall i \in \mathbb{N}$. Let $\tau_\beta := \min\{n \in \mathbb{N} : m\rho^n \leq \beta\}$, then for all $i \in \mathbb{N}$*

$$\mathbb{E}^{\mu^\pi}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{st})\right\} \leq 4G^2 m\rho^{\tau_\beta} + 6\beta G^2\tau_\beta \leq 10\beta G^2\tau_\beta \tag{10}$$

for projected self-tuned Q-learning iterates $(\widehat{\boldsymbol{\theta}}_i^{st})_{i\in\mathbb{N}}$.

Proof. For all $i \geq 0$, let $\tilde{\beta}_i := \beta/(1+\beta\|\boldsymbol{\phi}(\boldsymbol{S}_i^\pi, A_i^\pi)\|_2^2)$. Observe that

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{i+1}^{\text{st}} - \widehat{\boldsymbol{\theta}}_i^{\text{st}}\right\|_2 &= \left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_i^{\text{st}} + \tilde{\beta}_i f_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}})\} - \Pi_r\widehat{\boldsymbol{\theta}}_i^{\text{st}}\right\|_2 \\
&\leq \left\|\widehat{\boldsymbol{\theta}}_i^{\text{st}} + \tilde{\beta}_i f_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}}) - \widehat{\boldsymbol{\theta}}_i^{\text{st}}\right\|_2 && \text{(non-expansiveness of } \Pi_r\text{)} \\
&\leq \tilde{\beta}_i G && \text{(by Lemma B.1)} \\
&\leq \beta G && \text{since } \tilde{\beta}_i \leq \beta.
\end{aligned}$$

The remainder of the proof follows the same arguments as those used in the proof of Lemma B.5. □

Lemma B.8. *Suppose the step-size sequence $\{\beta_i\}_{i\in\mathbb{N}}$ is decreasing, i.e., $\beta_{i+1} < \beta_i$, $\forall i \in \mathbb{N}$. Fix $i \leq t$, and let $\tau_{\beta_t} := \min\{n \in \mathbb{N} : m\rho^n \leq \beta_t\}$, then*

$$\begin{aligned}
\mathbb{E}^{\mu^\pi}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{st})\right\} &\leq 6G^2\tau_{\beta_t}\beta_0 \quad \textit{for} \quad i \leq \tau_{\beta_t} \\
\mathbb{E}^{\mu^\pi}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{st})\right\} &\leq 10G^2\tau_{\beta_t}\beta_{i-\tau_{\beta_t}} \quad \textit{for} \quad i > \tau_{\beta_t}
\end{aligned}$$

for projected self-tuned Q-learning iterates $(\widehat{\boldsymbol{\theta}}_i^{st})_{i\in\mathbb{N}}$.

Proof. For all $i \geq 0$, let $\tilde{\beta}_i := \beta_i/(1+\beta_i\|\boldsymbol{\phi}(\boldsymbol{S}_i^\pi, A_i^\pi)\|_2^2)$. Observe that

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{i+1}^{\text{st}} - \widehat{\boldsymbol{\theta}}_i^{\text{st}}\right\|_2 &= \left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_i^{\text{st}} + \tilde{\beta}_i f_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}})\} - \Pi_r\widehat{\boldsymbol{\theta}}_i^{\text{st}}\right\|_2 \\
&\leq \left\|\widehat{\boldsymbol{\theta}}_i^{\text{st}} + \tilde{\beta}_i f_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}}) - \widehat{\boldsymbol{\theta}}_i^{\text{st}}\right\|_2 && \text{(non-expansiveness of } \Pi_r\text{)} \\
&\leq \tilde{\beta}_i G && \text{(by Lemma B.1)} \\
&\leq \beta_i G && \text{since } \tilde{\beta}_i \leq \beta_i.
\end{aligned}$$

The remainder of the proof follows the same arguments as those used in the proof of Lemma B.6. □

Lastly, we upper bound w_q, the infimum over $\boldsymbol{\theta} \in \mathbb{R}^M$ of the smallest eigenvalue of $\boldsymbol{\Sigma}^{\mu^\pi} - \gamma^2\boldsymbol{\Sigma}^\star(\boldsymbol{\theta})$, which governs the decay of the initialization error in our finite-time bounds.

Lemma B.9. *Under Assumption 4.2 and the feature normalization $\|\boldsymbol{\phi}(\boldsymbol{s}, a)\|_2 \leq 1$, $w_q \in (0, 1]$.*

Proof. Positivity of w_q is Assumption 4.2. Fix $\boldsymbol{\theta} \in \mathbb{R}^M$. Since $\boldsymbol{\Sigma}^\star(\boldsymbol{\theta})$ is positive semidefinite, $\boldsymbol{\Sigma}^{\mu^\pi} - \gamma^2\boldsymbol{\Sigma}^\star(\boldsymbol{\theta}) \preceq \boldsymbol{\Sigma}^{\mu^\pi}$ and therefore

$$\lambda_{\min}\left\{\boldsymbol{\Sigma}^{\mu^\pi} - \gamma^2\boldsymbol{\Sigma}^\star(\boldsymbol{\theta})\right\} \leq \lambda_{\min}(\boldsymbol{\Sigma}^{\mu^\pi}).$$

By Assumption 4.2, $w_q \leq \lambda_{\min}(\boldsymbol{\Sigma}^{\mu^\pi})$. The trace inequality and the normalization then yield

$$\lambda_{\min}(\boldsymbol{\Sigma}^{\mu^\pi}) \leq \frac{1}{M}\operatorname{tr}\boldsymbol{\Sigma}^{\mu^\pi} = \frac{1}{M}\mathbb{E}^{\mu^\pi}\|\boldsymbol{\phi}\|_2^2 \leq \frac{1}{M}.$$

□

B.2 Proof of Theorem 4.3

Standard Q-Learning: we first obtain the finite-time error bound of the projected Q-learning iterates with a step-size sequence of the form $\beta_i = \frac{\beta_0}{(i+1)^s},\ s \in (0.5, 1),\ i \in \mathbb{N}$. Consider

$$\begin{aligned}
\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 &= \mathbb{E}^{\mu^\pi}\left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_t + \beta_t f_t(\widehat{\boldsymbol{\theta}}_t)\} - \Pi_r\boldsymbol{\theta}^\star\right\|_2^2 \\
&\leq \mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_t + \beta_t f_t(\widehat{\boldsymbol{\theta}}_t) - \boldsymbol{\theta}^\star\right\|_2^2 \\
&= \mathbb{E}^{\mu^\pi}\left\{\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2\left\|f_t(\widehat{\boldsymbol{\theta}}_t)\right\|_2^2 + 2\beta_t\langle\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star, f_t(\widehat{\boldsymbol{\theta}}_t)\rangle\right\} \\
&= \mathbb{E}^{\mu^\pi}\left\{\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2\left\|f_t(\widehat{\boldsymbol{\theta}}_t)\right\|_2^2 + 2\beta_t\langle\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star, \bar{f}(\widehat{\boldsymbol{\theta}}_t) - \bar{f}(\boldsymbol{\theta}^\star)\rangle + 2\beta_t\Lambda_t(\widehat{\boldsymbol{\theta}}_t)\right\}
\end{aligned}$$

where the inequality is thanks to the non-expansiveness of Π_r, and the last equality is due to $\bar{f}(\boldsymbol{\theta}^\star) = 0$. From Lemma B.1 and Lemma B.2, we obtain

$$\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \leq (1 - \beta_t w_q)\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2 G^2 + 2\beta_t\mathbb{E}^{\mu^\pi}\Lambda_t(\widehat{\boldsymbol{\theta}}_t).$$

Recursively applying the inequality with the assumption $\beta_0 \in (0, 1/w_q)$, we have

$$\begin{aligned}
\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 &\leq \left\{\prod_{i=0}^{t}(1 - \beta_i w_q)\right\}\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + G^2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - \beta_k w_q)\right\}\beta_i^2 \\
&\quad + 2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - \beta_k w_q)\right\}\mathbb{E}^{\mu^\pi}\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\}\beta_i \\
&\leq \left\{\prod_{i=0}^{t}(1 - \beta_i w_q)\right\}\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + G^2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - \beta_k w_q)\right\}\beta_i^2 \\
&\quad + 12G^2\tau_{\beta_t}\beta_0\sum_{i=0}^{\tau_{\beta_t}}\left\{\prod_{k=i+1}^{t}(1 - \beta_k w_q)\right\}\beta_i \\
&\quad + 20G^2\tau_{\beta_t}\sum_{i=\tau_{\beta_t}+1}^{t}\left\{\prod_{k=i+1}^{t}(1 - \beta_k w_q)\right\}\beta_{i-\tau_{\beta_t}}\beta_i
\end{aligned}$$

where the second inequality follows from Lemma B.6. Thanks to $1 - \beta_i w_q \leq \exp(-\beta_i w_q)$, we have

$$\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \leq \exp\left(-w_q\sum_{i=0}^{t}\beta_i\right)\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 \tag{11}$$

$$+ G^2\sum_{i=0}^{t}\exp\left(-w_q\sum_{k=i+1}^{t}\beta_k\right)\beta_i^2 \tag{12}$$

$$+ 12G^2\tau_{\beta_t}\beta_0\sum_{i=0}^{\tau_{\beta_t}}\exp\left(-w_q\sum_{k=i+1}^{t}\beta_k\right)\beta_i \tag{13}$$

$$+ 20G^2\tau_{\beta_t}\sum_{i=\tau_{\beta_t}+1}^{t}\exp\left(-w_q\sum_{k=i+1}^{t}\beta_k\right)\beta_{i-\tau_{\beta_t}}\beta_i. \tag{14}$$

We now provide a bound for each of the term above. For the first term, we have

$$\begin{aligned}
(11) &= \exp\left\{-w_q\beta_0\sum_{i=0}^{t}\frac{1}{(i+1)^s}\right\}\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 \\
&\lesssim \exp\left[\frac{-w_q\beta_0}{(1-s)}\left\{(t+1)^{1-s} - 1\right\}\right]\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2
\end{aligned} \tag{15}$$

For the second term, by Lemma D.1, we have

$$(12) \lesssim \exp\left[\frac{-w_q\beta_0}{2(1-s)}\{(t+1)^{1-s}-1\}\right] + \beta_t \tag{16}$$

For the third term, by Lemma D.2, we get

$$(13) \lesssim \tau_{\beta_t}\exp\left[\frac{-w_q\beta_0}{(1-s)}\{(t+1)^{1-s}-(\tau_{\beta_t}+1)^{1-s}\}\right] \tag{17}$$

For the last term, by Lemma D.2, we yield

$$(14) \lesssim \tau_{\beta_t}\left[\exp\left\{\frac{-w_q\beta_0}{2(1-s)}\{(t+1)^{1-s}-1\}\right\} + \beta_{t-\tau_{\beta_t}}\right] \tag{18}$$

Combining (15)-(18) we get,

$$\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}_{t+1}-\boldsymbol{\theta}^\star\right\|_2^2 \lesssim \exp\left[\frac{-w_q\beta_0}{(1-s)}\{(t+1)^{1-s}-1\}\right]\left\|\widehat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}^\star\right\|_2^2 + \mathcal{O}\left(\tau_{\beta_t}\beta_{t-\tau_{\beta_t}}\right).$$

Self-tuned Q-Learning: we next obtain the finite-time error bound of the projected self-tuned Q-learning iterates with a step-size sequence of the form $\beta_i = \frac{\beta_0}{(i+1)^s},\ s\in(0.5,1),\ i\in\mathbb{N}$. With a slight abuse of notation, let $\tilde{\beta}_i := \frac{\beta_i}{1+\beta_i\|\boldsymbol{\phi}(\boldsymbol{S}_i^\pi,A_i^\pi)\|_2^2}$. Consider

$$\begin{aligned}
\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_{t+1}-\boldsymbol{\theta}^\star\right\|_2^2 &= \mathbb{E}^{\mu^\pi}\left\|\Pi_r\{\widehat{\boldsymbol{\theta}}^{\text{st}}_t+\tilde{\beta}_t f_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)\}-\Pi_r\boldsymbol{\theta}^\star\right\|_2^2\\
&\le \mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_t+\tilde{\beta}_t f_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)-\boldsymbol{\theta}^\star\right\|_2^2\\
&\le \mathbb{E}^{\mu^\pi}\left\{\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star\right\|_2^2+\beta_t^2\left\|f_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)\right\|_2^2+2\tilde{\beta}_t\langle\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star,f_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)\rangle\right\}\\
&= \mathbb{E}^{\mu^\pi}\left\{\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star\right\|_2^2+\beta_t^2\left\|f_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)\right\|_2^2+2\tilde{\beta}_t\langle\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star,\bar{f}(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)-\bar{f}(\boldsymbol{\theta}^\star)\rangle+2\tilde{\beta}_t\Lambda_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)\right\}
\end{aligned}$$

where the first inequality is thanks to the non-expansiveness of Π_r, and the second inequality is due to $\tilde{\beta}_t\le\beta_t$. In the last equality, we used $\bar{f}(\boldsymbol{\theta}^\star)=0$. From Lemma B.1 and Lemma B.2, we obtain

$$\begin{aligned}
\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_{t+1}-\boldsymbol{\theta}^\star\right\|_2^2 &\le \mathbb{E}^{\mu^\pi}\left\{(1-\tilde{\beta}_t w_q)\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star\right\|_2^2\right\}+\beta_t^2G^2+2\beta_t\mathbb{E}^{\mu^\pi}\Lambda_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)+2\mathbb{E}^{\mu^\pi}\left\{(\tilde{\beta}_t-\beta_t)\Lambda_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)\right\}\\
&\le \left(1-\frac{\beta_t w_q}{1+\beta_t}\right)\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star\right\|_2^2+\beta_t^2G^2+2\beta_t\mathbb{E}^{\mu^\pi}\Lambda_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t)+4\beta_t^2G^2\\
&= \left(1-\frac{\beta_t w_q}{1+\beta_t}\right)\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_t-\boldsymbol{\theta}^\star\right\|_2^2+5\beta_t^2G^2+2\beta_t\mathbb{E}^{\mu^\pi}\Lambda_t(\widehat{\boldsymbol{\theta}}^{\text{st}}_t) \qquad (19)
\end{aligned}$$

where the second inequality follows from the almost-sure bound $\tilde{\beta}_t\ge\frac{\beta_t}{1+\beta_t}$, together with the bound $|\tilde{\beta}_t-\beta_t|\le\beta_t^2$ and Lemma B.3. Thanks to Lemma B.9, we know that $w_q\in(0,1]$ and hence we have

$$\begin{aligned}
\mathbb{E}^{\mu^\pi}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_{t+1}-\boldsymbol{\theta}^\star\right\|_2^2 &\le \left\{\prod_{i=0}^t\left(1-\frac{w_q\beta_i}{1+\beta_i}\right)\right\}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_0-\boldsymbol{\theta}^\star\right\|_2^2+5G^2\sum_{i=0}^t\left\{\prod_{k=i+1}^t\left(1-\frac{w_q\beta_k}{1+\beta_k}\right)\right\}\beta_i^2\\
&\quad+2\sum_{i=0}^t\left\{\prod_{k=i+1}^t\left(1-\frac{w_q\beta_k}{1+\beta_k}\right)\right\}\mathbb{E}^{\mu^\pi}\{\Lambda_i(\widehat{\boldsymbol{\theta}}^{\text{st}}_i)\}\beta_i\\
&\le \left\{\prod_{i=0}^t\left(1-\frac{w_q\beta_i}{1+\beta_i}\right)\right\}\left\|\widehat{\boldsymbol{\theta}}^{\text{st}}_0-\boldsymbol{\theta}^\star\right\|_2^2+5G^2\sum_{i=0}^t\left\{\prod_{k=i+1}^t\left(1-\frac{w_q\beta_k}{1+\beta_k}\right)\right\}\beta_i^2\\
&\quad+12G^2\tau_{\beta_t}\beta_0\sum_{i=0}^{\tau_{\beta_t}}\left\{\prod_{k=i+1}^t\left(1-\frac{w_q\beta_k}{1+\beta_k}\right)\right\}\beta_i\\
&\quad+20G^2\tau_{\beta_t}\sum_{i=\tau_{\beta_t}+1}^t\left\{\prod_{k=i+1}^t\left(1-\frac{w_q\beta_k}{1+\beta_k}\right)\right\}\beta_{i-\tau_{\beta_t}}\beta_i.
\end{aligned}$$

where the second inequality follows from Lemma B.8. Since

$$1 - \frac{w_q \beta_i}{1+\beta_i} \leq \exp\left(-\frac{w_q \beta_i}{1+\beta_i}\right),$$

we have

$$\mathbb{E}^{\mu^\pi} \left\| \widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star \right\|_2^2 \leq \exp\left(-w_q \sum_{i=0}^{t} \frac{\beta_i}{1+\beta_i}\right) \left\| \widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star \right\|_2^2 \tag{20}$$

$$+ 5G^2 \sum_{i=0}^{t} \exp\left(-w_q \sum_{k=i+1}^{t} \frac{\beta_k}{1+\beta_k}\right) \beta_i^2 \tag{21}$$

$$+ 12G^2 \tau_{\beta_t} \beta_0 \sum_{i=0}^{\tau_{\beta_t}} \exp\left(-w_q \sum_{k=i+1}^{t} \frac{\beta_k}{1+\beta_k}\right) \beta_i \tag{22}$$

$$+ 20G^2 \tau_{\beta_t} \sum_{i=\tau_{\beta_t}+1}^{t} \exp\left(-w_q \sum_{k=i+1}^{t} \frac{\beta_k}{1+\beta_k}\right) \beta_{i-\tau_{\beta_t}} \beta_i. \tag{23}$$

We now provide a bound for each of the term above. Since

$$\sum \frac{\beta_k}{1+\beta_k} = \sum \beta_k - \sum \frac{\beta_k^2}{1+\beta_k} = \sum \beta_k + \mathcal{O}(1)$$

for $\beta_k = \beta_0/(k+1)^s$, $s \in (0.5, 1)$. For the first term, we have

$$\begin{aligned}(20) &\lesssim \exp\left\{-w_q \beta_0 \sum_{i=0}^{t} \frac{1}{(i+1)^s}\right\} \left\| \widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star \right\|_2^2 \\ &\lesssim \exp\left[\frac{-w_q \beta_0}{(1-s)} \left\{(t+1)^{1-s} - 1\right\}\right] \left\| \widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star \right\|_2^2 \end{aligned} \tag{24}$$

For the second term, by Lemma D.1, we have

$$(21) \lesssim \exp\left[\frac{-w_q \beta_0}{2(1-s)} \{(t+1)^{1-s} - 1\}\right] + \beta_t \tag{25}$$

For the third term, by Lemma D.2, we get

$$(22) \lesssim \tau_{\beta_t} \exp\left[\frac{-w_q \beta_0}{(1-s)} \{(t+1)^{1-s} - (\tau_{\beta_t}+1)^{1-s}\}\right] \tag{26}$$

For the last term, by Lemma D.2, we yield

$$(23) \lesssim \tau_{\beta_t} \left[\exp\left\{\frac{-w_q \beta_0}{2(1-s)} \{(t+1)^{1-s} - 1\}\right\} + \beta_{t-\tau_{\beta_t}}\right] \tag{27}$$

Combining (24)-(27) we get,

$$\mathbb{E}^{\mu^\pi} \left\| \widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star \right\|_2^2 \lesssim \exp\left[\frac{-w_q \beta_0}{(1-s)} \{(t+1)^{1-s} - 1\}\right] \left\| \widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star \right\|_2^2 + \mathcal{O}\left(\tau_{\beta_t} \beta_{t-\tau_{\beta_t}}\right).$$

C Proofs of SARSA

C.1 Preliminary Lemmas

For the SARSA algorithm, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, let us define

$$\begin{aligned} g_t(\boldsymbol{\theta}) &:= \boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}}) \left\{R_t + \gamma \boldsymbol{\phi}(\boldsymbol{S}_{t+1}^{\pi_{t-1}}, A_{t+1}^{\pi_t})^\top \boldsymbol{\theta} - \boldsymbol{\phi}(\boldsymbol{S}_t^{\pi_{t-2}}, A_t^{\pi_{t-1}})^\top \boldsymbol{\theta}\right\} \\ \bar{g}(\boldsymbol{\theta}) &:= \mathbb{E}^{\mu^{\boldsymbol{\theta}}} \left[\phi(\boldsymbol{S}, A) \left\{r(\boldsymbol{S}, A) + \gamma \phi(\boldsymbol{S}', A')^\top \boldsymbol{\theta} - \phi(\boldsymbol{S}, A)^\top \boldsymbol{\theta}\right\}\right]. \end{aligned}$$

Recall that $\mathbb{E}^{\mu^{\boldsymbol{\theta}}}$ denotes the expectation where the current state $\boldsymbol{S}$ follows the stationary distribution $\mu^{\pi_{\boldsymbol{\theta}}}$, the current action A is chosen according to the behavioral policy $\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{S})$, $\boldsymbol{S}'$ denotes the subsequent state transitioned following $\mathsf{P}(\cdot|\boldsymbol{S}, A)$, and A' is chosen based on the behavioral policy $\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{S}')$. With a slight abuse of notation, we define $\Lambda_t(\boldsymbol{\theta}) := \langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star, g_t(\boldsymbol{\theta}) - \bar{g}(\boldsymbol{\theta}) \rangle$. Recall that $\bar{g}(\boldsymbol{\theta}^\star) = 0$. We now provide the necessary Lemmas to establish finite-time error bounds.

Lemma C.1 (Lemma 2 of (Zou et al., 2019a)). *For any $\boldsymbol{\theta} \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}\|_2 \leq r$,*

$$\|g_t(\boldsymbol{\theta})\|_2 \leq G := b + 2r \;\; \textit{for all} \;\; t \geq 0$$

Lemma C.2 (Lemma 4 of (Zou et al., 2019a)). *For any $\boldsymbol{\theta} \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}\|_2 \leq r$,*

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star, \bar{g}(\boldsymbol{\theta}) - \bar{g}(\boldsymbol{\theta}^\star) \rangle \leq -w_s \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2$$

where $w_s > 0$ is the smallest eigenvalue of

$$-\frac{1}{2}\left\{\left(\boldsymbol{A}^{\boldsymbol{\theta}^\star} + C\lambda \boldsymbol{I}\right) + \left(\boldsymbol{A}^{\boldsymbol{\theta}^\star} + C\lambda \boldsymbol{I}\right)^\top\right\}.$$

Here, $\lambda := G|\mathcal{A}|\left(2 + \lceil \log_\rho \frac{1}{m} \rceil + \frac{1}{1-\rho}\right)$ and $C > 0$ is the Lipschitz constant of the behavioral policy π_θ.

Lemma C.3 (Lemma 5 of (Zou et al., 2019a)). *For any $\boldsymbol{\theta} \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}\|_2 \leq r$, $|\Lambda_t(\boldsymbol{\theta})| \leq 2G^2$ for all $t \geq 0$.*

Lemma C.4 (Lemma 6 of (Zou et al., 2019a)). *For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^M$ such that $\|\boldsymbol{\theta}_1\|_2 \leq r, \|\boldsymbol{\theta}_2\|_2 \leq r$,*

$$|\Lambda_t(\boldsymbol{\theta}_1) - \Lambda_t(\boldsymbol{\theta}_2)| \leq (6 + \lambda C) G \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \;\; \textit{for all} \;\; t \geq 0.$$

Here, $\lambda := G|\mathcal{A}|\left(2 + \lceil \log_\rho \frac{1}{m} \rceil + \frac{1}{1-\rho}\right)$ and $C > 0$ is the Lipschitz constant of the behavioral policy π_θ.

Lemma C.5. *Given a step-size sequence $(\beta_k)_{k \geq 0}$, for any $i \leq \tau$,*

$$\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\right\} \leq 2G^2 + (6 + \lambda C) G^2 \sum_{k=0}^{\tau-1} \beta_k.$$

For any $i > \tau > 0$,

$$\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\right\} \leq 2C|\mathcal{A}|G^3\tau \sum_{k=i-\tau}^{i-1} \beta_k + 4G^2 m \rho^{\tau-1} + (6 + \lambda C) G^2 \sum_{k=i-\tau}^{i-1} \beta_k.$$

Proof. We provide a proof for the case $i \leq \tau$. For $i > \tau > 0$, the proof follows directly from Lemma 7 of (Zou et al., 2019a). For any $i \geq 0$, we have

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{k+1} - \widehat{\boldsymbol{\theta}}_k\right\|_2 &= \left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_k + \beta_k g_k(\widehat{\boldsymbol{\theta}}_k)\} - \Pi_r(\widehat{\boldsymbol{\theta}}_k)\right\|_2 \\
&\leq \left\|\beta_k g_k(\widehat{\boldsymbol{\theta}}_k)\right\|_2 && \text{(non-expansiveness of } \Pi_r\text{)} \\
&\leq \beta_k G. && \text{(by Lemma C.1)}
\end{aligned}$$

Therefore, by the triangle inequality,

$$\left\|\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_0\right\|_2 \leq \sum_{k=0}^{i-1} \left\|\widehat{\boldsymbol{\theta}}_{k+1} - \widehat{\boldsymbol{\theta}}_k\right\|_2 \leq G \sum_{k=0}^{i-1} \beta_k \leq G \sum_{k=0}^{\tau-1} \beta_k \;\; \text{for any } i \leq \tau.$$

Combined with the Lipschitzness of $\Lambda_i(\boldsymbol{\theta})$ in Lemma C.4, we have

$$|\Lambda_i(\widehat{\boldsymbol{\theta}}_i) - \Lambda_i(\widehat{\boldsymbol{\theta}}_0)| \leq (6 + \lambda C) G^2 \sum_{k=0}^{\tau-1} \beta_k \Longrightarrow \Lambda_i(\widehat{\boldsymbol{\theta}}_i) \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_0) + (6 + \lambda C) G^2 \sum_{k=0}^{\tau-1} \beta_k.$$

Applying Lemma C.3 and taking expectation on both sides, we get

$$\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\right\} \leq 2G^2 + (6 + \lambda C) G^2 \sum_{k=0}^{\tau-1} \beta_k.$$

□

Lemma C.6. *Given a step-size sequence $(\beta_k)_{k\geq 0}$, for any $i \leq \tau$,*

$$\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{st})\right\} \leq 2G^2 + (6+\lambda C)G^2 \sum_{k=0}^{\tau-1} \beta_k.$$

For any $i > \tau > 0$,

$$\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{st})\right\} \leq 2C|\mathcal{A}|G^3\tau \sum_{k=i-\tau}^{i-1} \beta_k + 4G^2 m\rho^{\tau-1} + (6+\lambda C)G^2 \sum_{k=i-\tau}^{i-1} \beta_k.$$

Proof. We proceed in three steps. The proof largely follows the approach of Lemma 7 in Zou et al. (2019a), with additional care in handling the self-tuned iterates.

Step 1. For any $k \geq 0$, we have

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_{k+1}^{\text{st}} - \widehat{\boldsymbol{\theta}}_k^{\text{st}}\right\|_2 &= \left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_k^{\text{st}} + \tilde{\beta}_k g_k(\widehat{\boldsymbol{\theta}}_k^{\text{st}})\} - \Pi_r(\widehat{\boldsymbol{\theta}}_k^{\text{st}})\right\|_2 \\
&\leq \left\|\tilde{\beta}_k g_k(\widehat{\boldsymbol{\theta}}_k^{\text{st}})\right\|_2 && \text{(non-expansiveness of } \Pi_r) \\
&\leq \beta_k G. && \text{(by Lemma C.1 and } \tilde{\beta}_k \leq \beta_k)
\end{aligned}$$

The proof for the case $i \leq \tau$ follows from analogous arguments to those in Lemma C.5. Therefore, we focus on the case $i > \tau > 0$. By the triangle inequality,

$$\left\|\widehat{\boldsymbol{\theta}}_i^{\text{st}} - \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}\right\|_2 \leq \sum_{k=i-\tau}^{i-1} \left\|\widehat{\boldsymbol{\theta}}_{k+1}^{\text{st}} - \widehat{\boldsymbol{\theta}}_k^{\text{st}}\right\|_2 \leq G \sum_{k=i-\tau}^{i-1} \beta_k \text{ for any } i > \tau > 0.$$

Combined with the Lipschitzness of $\Lambda_i(\boldsymbol{\theta})$ in Lemma C.4, we have

$$|\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}}) - \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}})| \leq (6+\lambda C)G^2 \sum_{k=i-\tau}^{i-1} \beta_k \Longrightarrow \Lambda_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}}) \leq \Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}) + (6+\lambda C)G^2 \sum_{k=i-\tau}^{i-1} \beta_k. \tag{28}$$

In the subsequent steps, we aim to upper bound the quantity $\Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}})$ in expectation.

Step 2. In this step, we obtain an upper bound on the closely related object induced by an auxiliary Markov chain. To clarify the approach, we use the abbreviated policy notation $\pi_i := \pi_{\widehat{\boldsymbol{\theta}}_i^{\text{st}}}$ and data tuple $O_i := (\boldsymbol{S}_i^{\pi_{i-2}}, A_i^{\pi_{i-1}}, \boldsymbol{S}_{i+1}^{\pi_{i-1}}, A_{i+1}^{\pi_i})$. To highlight the dependence on data, we rewrite $\Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}) = \Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, O_i)$. Additionally, conditioning on $\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}$ and $\boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}$, we construct the following auxiliary Markov chain

$$\boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}} \to \boldsymbol{S}_{i-\tau+2}^{\pi_{i-\tau}} \to \boldsymbol{S}_{i-\tau+3}^{\pi_{i-\tau}} \to \cdots \to \boldsymbol{S}_i^{\pi_{i-\tau}} \to \boldsymbol{S}_{i+1}^{\pi_{i-\tau}}, \tag{29}$$

whose transition probability is defined as

$$\begin{aligned}
\mathbb{P}(\boldsymbol{S}_{i-\tau+2}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) &= \sum_{a\in A} \pi_{i-\tau}(A_{i-\tau+1}^{\pi_{i-\tau}} = a \mid \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathsf{P}(\boldsymbol{S}_{i-\tau+2}^{\pi_{i-\tau}} \in \cdot \mid A_{i-\tau+1}^{\pi_{i-\tau}} = a, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \\
\mathbb{P}(\boldsymbol{S}_k^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{k-1}^{\pi_{i-\tau}}) &= \sum_{a\in A} \pi_{i-\tau}(A_{k-1}^{\pi_{i-\tau}} = a \mid \boldsymbol{S}_{k-1}^{\pi_{i-\tau}}) \mathsf{P}(\boldsymbol{S}_k^{\pi_{i-\tau}} \in \cdot \mid A_{k-1}^{\pi_{i-\tau}} = a, \boldsymbol{S}_{k-1}^{\pi_{i-\tau}}),
\end{aligned}$$

for $k = i-\tau+3, \ldots, i+1$. In short, conditioning on $\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}$ and $\boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}$, the constructed Markov chain is induced by repeatedly following the same behavior policy $\pi_{i-\tau}$. From the Assumption 4.6, it follows that conditioning on $(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}})$, for any $k \geq i-\tau+2$,

$$\left\|\mathbb{P}(\boldsymbol{S}_k^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mu^{\pi_{i-\tau}}\right\|_{TV} \leq m\rho^{k-(i-\tau+1)}.$$

Now, let us define auxiliary data tuple $\tilde{O}_i := (\boldsymbol{S}_i^{\pi_{i-\tau}}, A_i^{\pi_{i-\tau}}, \boldsymbol{S}_{i+1}^{\pi_{i-\tau}}, A_{i+1}^{\pi_{i-\tau}})$. Applying Lemma C.3, Lemma 9 in (Bhandari et al., 2018) and the triangle inequality with

$$\left\|\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mu^{\pi_{i-\tau}}\right\|_{TV} \leq m\rho^{\tau-1}$$
$$\left\|\mathbb{P}(\boldsymbol{S}_{i+1}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mu^{\pi_{i-\tau}}\right\|_{TV} \leq m\rho^{\tau},$$

we obtain

$$\mathbb{E}\left\{\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \tilde{O}_i) \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}\right\} - \mathbb{E}\left\{\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, O_i') \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}\right\} \leq 4G^2 m\rho^{\tau-1}$$

where O_i' are independently generated by the stationary distribution $\mu^{\pi_{i-\tau}}$ and the policy $\pi_{i-\tau}$. Since $\mathbb{E}^{\mu^{\boldsymbol{\theta}}}\{\Lambda_i(\boldsymbol{\theta})\} = 0$ for any fixed $\boldsymbol{\theta}$, we know

$$\mathbb{E}\left\{\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, O_i') \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}\right\} = 0,$$

which yields

$$\mathbb{E}\left\{\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \tilde{O}_i)\right\} \leq 4G^2 m\rho^{\tau-1}. \tag{30}$$

Step 3. In this final step, we analyze the difference between the target quantity $\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}})\right\}$ and the auxiliary quantity $\mathbb{E}\left\{\Lambda(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \tilde{O}_i)\right\}$. First, note that

$$\begin{aligned}
&\mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, O_i) \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}\right\} - \mathbb{E}\left\{\Lambda_i(\widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \tilde{O}_i) \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}\right\} \\
&\leq 2G^2 \left\|\mathbb{P}(O_i \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\tilde{O}_i \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}})\right\|_{TV}.
\end{aligned}$$

From the result in section B of (Zou et al., 2019a), we have

$$\begin{aligned}
&\left\|\mathbb{P}(O_i \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\tilde{O}_i \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}})\right\|_{TV} \\
&\leq \left\|\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}})\right\|_{TV} + C|\mathcal{A}|G \sum_{j=i-2}^{i-1} \sum_{k=i-\tau}^{j} \beta_k.
\end{aligned}$$

We obtain the upper bound of the first term recursively as follows:

$$\begin{aligned}
&\left\|\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}})\right\|_{TV} \\
&= \frac{1}{2} \int_{s' \in \mathcal{S}} \left|\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}})\right| \\
&= \frac{1}{2} \int_{s' \in \mathcal{S}} \Big| \int_{s \in \mathcal{S}} \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-3}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
&\quad - \int_{s \in \mathcal{S}} \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s)\Big|
\end{aligned} \tag{31}$$

Pushing absolute value operation inside the integral, we upper bound (31) as

$$\begin{aligned}
(31) \leq \frac{1}{2} \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} \Big|&\mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-3}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
&- \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s)\Big|.
\end{aligned}$$

Using the triangle inequality, we know that

$$
\begin{aligned}
(31) \leq \frac{1}{2} \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} \Bigg\{ & \Big| \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-3}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-2}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
& - \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-2}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \Big| \\
& + \Big| \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-2}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
& - \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = ds \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-\tau}} = ds' \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s) \Big| \Bigg\},
\end{aligned}
$$

which leads to

$$
\begin{aligned}
(31) \leq & \left\| \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-3}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \right\|_{TV} \\
& + \sup_{s \in \mathcal{S}} \left\| \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-2}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) - \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s) \right\|_{TV}.
\end{aligned}
$$

We re-express the term inside the supremum operator as follows:

$$
\begin{aligned}
& \left\| \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-2}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) - \mathbb{P}(\boldsymbol{S}_{i}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s) \right\|_{TV} \\
& = \frac{1}{2} \int_{s' \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \mathsf{P}(ds' \mid s, a) \mathbb{E} \left\{ \pi_{i-2}(a \mid s) \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s \right\} - \mathsf{P}(ds' \mid s, a) \pi_{i-\tau}(a \mid s) \right|
\end{aligned} \tag{32}
$$

where the $\mathbb{E}$ is with respect to $\widehat{\boldsymbol{\theta}}_{i-2}^{\text{st}}$ used in the policy π_{i-2}. We provide the justification of (32) in the following subsection. Then we have

$$
\begin{aligned}
(32) & \leq \frac{1}{2} \int_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathsf{P}(ds' \mid s, a) \left| \mathbb{E} \left\{ \pi_{i-2}(a \mid s) - \pi_{i-\tau}(a \mid s) \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s \right\} \right| \\
& \leq C \int_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathsf{P}(ds' \mid s, a) \left\| \widehat{\boldsymbol{\theta}}_{i-2}^{\text{st}} - \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}} \right\|_2 \leq C|\mathcal{A}|G \sum_{k=i-\tau}^{i-3} \beta_k,
\end{aligned}
$$

where the first inequality is due to the triangle inequality and the second inequality is the consequence of the Lipschitzness of the policy. In the third inequality, we used the fact that

$$
\left\| \widehat{\boldsymbol{\theta}}_{i-2}^{\text{st}} - \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}} \right\|_2 \leq \sum_{k=i-\tau}^{i-3} \left\| \widehat{\boldsymbol{\theta}}_{k+1}^{\text{st}} - \widehat{\boldsymbol{\theta}}_{k}^{\text{st}} \right\|_2 \leq G \sum_{k=i-\tau}^{i-3} \tilde{\beta}_k \leq G \sum_{k=i-\tau}^{i-3} \beta_k,
$$

Thus, we obtain a recursive bound of (31) as follows:

$$
(31) \leq \left\| \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-3}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\boldsymbol{S}_{i-1}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \right\|_{TV} + C|\mathcal{A}|G \sum_{k=i-\tau}^{i-3} \beta_k.
$$

Applying this recursion repeatedly, we obtain

$$
\begin{aligned}
& \left\| \mathbb{P}(O_i \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\tilde{O}_i \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \right\|_{TV} \\
& \leq \left\| \mathbb{P}(\boldsymbol{S}_{i-\tau+2}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) - \mathbb{P}(\boldsymbol{S}_{i-\tau+2}^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}_{i-\tau}^{\text{st}}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}) \right\|_{TV} + C|\mathcal{A}|G \sum_{j=i-\tau}^{i-1} \sum_{k=i-\tau}^{j} \beta_k \\
& \leq C|\mathcal{A}|G\tau \sum_{k=i-\tau}^{i-1} \beta_k
\end{aligned}
$$

Combining all three steps, we obtain the desired result. □

C.1.1 JUSTIFICATION OF (32)

Notice that

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
&= \sum_{a\in\mathcal{A}} \mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot, A_{i-1}^{\pi_{i-2}} = a \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
&= \sum_{a\in\mathcal{A}} \mathsf{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot \mid \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s, A_{i-1}^{\pi_{i-2}} = a)\mathbb{P}(A_{i-1}^{\pi_{i-2}} = a \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s). \qquad (33)
\end{aligned}$$

Since

$$\begin{aligned}
&\mathbb{P}(A_{i-1}^{\pi_{i-2}} = a \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
&= \mathbb{E}\left\{\mathbb{P}(A_{i-1}^{\pi_{i-2}} = a \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-2}, \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+2}^{\pi_{i-\tau}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s\right\} \\
&= \mathbb{E}\left\{\pi_{i-2}(a \mid s) \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s\right\}
\end{aligned}$$

where the expectation is with respect to the randomness of $\widehat{\boldsymbol{\theta}}^{\text{st}}_{i-2}$, we have

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s) \\
&= \sum_{a\in\mathcal{A}} \mathsf{P}(\boldsymbol{S}_i^{\pi_{i-2}} \in \cdot \mid \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s, A_{i-1}^{\pi_{i-2}} = a)\mathbb{E}\left\{\pi_{i-2}(a \mid s) \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-3}} = s\right\}.
\end{aligned}$$

Finally, one can also observe that

$$\mathbb{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} \in \cdot \mid \widehat{\boldsymbol{\theta}}^{\text{st}}_{i-\tau}, \boldsymbol{S}_{i-\tau+1}^{\pi_{i-\tau-1}}, \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s) = \sum_{a\in\mathcal{A}} \mathsf{P}(\boldsymbol{S}_i^{\pi_{i-\tau}} \in \cdot \mid \boldsymbol{S}_{i-1}^{\pi_{i-\tau}} = s, \boldsymbol{A}_{i-1}^{\pi_{i-\tau}} = a)\pi_{i-\tau}(a|s). \qquad (34)$$

Lastly, we upper bound w_s, the smallest eigenvalue of $-\frac{1}{2}\{(\boldsymbol{A}^{\boldsymbol{\theta}^\star} + C\lambda\boldsymbol{I}) + (\boldsymbol{A}^{\boldsymbol{\theta}^\star} + C\lambda\boldsymbol{I})^\top\}$, which governs the decay of the initialization error in our finite-time bounds.

Lemma C.7. *Under Assumptions 4.6–4.7 and the normalization $\|\boldsymbol{\phi}(\boldsymbol{s}, a)\|_2 \le 1$, $w_s \in (0, \frac{1+\gamma}{M}]$. Hence, $w_s < 1/2$ whenever $M \ge 4$, i.e., whenever at least four basis functions are used.*

Proof. Positivity of w_s is the negative definiteness required in Assumption 4.7. Write $\boldsymbol{A} := \boldsymbol{A}^{\boldsymbol{\theta}^\star}$, $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{S}, A)$ and $\boldsymbol{\phi}' = \boldsymbol{\phi}(\boldsymbol{S}', A')$. Since

$$-\tfrac{1}{2}\{(\boldsymbol{A} + C\lambda\boldsymbol{I}) + (\boldsymbol{A} + C\lambda\boldsymbol{I})^\top\} = -\tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^\top) - C\lambda\boldsymbol{I},$$

we know $w_s = \lambda_{\min}\left\{-\frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^\top)\right\} - C\lambda$. Because $-\frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^\top)$ is symmetric with trace $-\operatorname{tr}(\boldsymbol{A})$,

$$w_s \le -\frac{\operatorname{tr}(\boldsymbol{A})}{M} - C\lambda \le -\frac{\operatorname{tr}(\boldsymbol{A})}{M}.$$

By definition of $\boldsymbol{A}$ and linearity of the trace, $\operatorname{tr}(\boldsymbol{A}) = \mathbb{E}^{\mu^{\theta^\star}}\left\{\gamma\boldsymbol{\phi}^\top\boldsymbol{\phi}' - \|\boldsymbol{\phi}\|_2^2\right\}$, and Cauchy–Schwarz with the normalization gives $\boldsymbol{\phi}^\top\boldsymbol{\phi}' \ge -\|\boldsymbol{\phi}\|_2\|\boldsymbol{\phi}'\|_2 \ge -1$, so $-\operatorname{tr}(\boldsymbol{A}) \le 1 + \gamma$. Thus $w_s \le (1+\gamma)/M < 2/M$. For $M \ge 4$ this gives $w_s < 1/2$. □

C.2 PROOF OF THEOREM 4.8

Standard SARSA: we first obtain the finite-time error bound of the projected SARSA iterates with a step-size sequence of the form $\beta_i = \frac{\beta_0}{(i+1)^s},\ s \in (0.5, 1),\ i \in \mathbb{N}$. Consider

$$\begin{aligned}
\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 &= \mathbb{E}\left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_t + \beta_t g_t(\widehat{\boldsymbol{\theta}}_t)\} - \Pi_r\boldsymbol{\theta}^\star\right\|_2^2 \\
&\leq \mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_t + \beta_t g_t(\widehat{\boldsymbol{\theta}}_t) - \boldsymbol{\theta}^\star\right\|_2^2 \\
&= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2\left\|g_t(\widehat{\boldsymbol{\theta}}_t)\right\|_2^2 + 2\beta_t\langle\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star, g_t(\widehat{\boldsymbol{\theta}}_t)\rangle\right\} \\
&= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2\left\|g_t(\widehat{\boldsymbol{\theta}}_t)\right\|_2^2 + 2\beta_t\langle\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star, \bar{g}(\widehat{\boldsymbol{\theta}}_t) - \bar{g}(\boldsymbol{\theta}^\star)\rangle + 2\beta_t\Lambda_t(\widehat{\boldsymbol{\theta}}_t)\right\}
\end{aligned}$$

where the inequality is thanks to the non-expansiveness of Π_r, and the last equality is due to $\bar{g}(\boldsymbol{\theta}^\star) = 0$. From Lemma C.1 and Lemma C.2, we obtain

$$\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \leq (1 - 2\beta_t w_s)\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2 G^2 + 2\beta_t\mathbb{E}\Lambda_t(\widehat{\boldsymbol{\theta}}_t).$$

Recursively applying the inequality with the assumption $\beta_0 \in (0, \frac{1}{2w_s})$, we have

$$\begin{aligned}
\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 &\leq \left\{\prod_{i=0}^{t}(1 - 2\beta_i w_s)\right\}\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + G^2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\beta_i^2 \\
&\quad + 2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\mathbb{E}\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i)\}\beta_i \\
&\lesssim \left\{\prod_{i=0}^{t}(1 - 2\beta_i w_s)\right\}\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + \sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\beta_i^2 \\
&\quad + \sum_{i=0}^{\tau_{\beta_t}}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\left(\sum_{k=0}^{\tau_{\beta_t}-1}\beta_k\right)\beta_i \\
&\quad + \sum_{i=\tau_{\beta_t}+1}^{t}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\left(\tau_{\beta_t}\sum_{k=i-\tau_{\beta_t}}^{i-1}\beta_k\right)\beta_i \\
&\lesssim \left\{\prod_{i=0}^{t}(1 - 2\beta_i w_s)\right\}\left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + \sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\beta_i^2 \\
&\quad + \tau_{\beta_t}\sum_{i=0}^{\tau_{\beta_t}}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\beta_i \\
&\quad + \left(\tau_{\beta_t}^2\beta_{i-\tau_{\beta_t}} + \tau_{\beta_t}\beta_{i-\tau_{\beta_t}}\right)\sum_{i=\tau_{\beta_t}+1}^{t}\left\{\prod_{k=i+1}^{t}(1 - 2\beta_k w_s)\right\}\beta_i
\end{aligned}$$

where the second inequality follows from Lemma C.5. Thanks to $1 - 2\beta_i w_s \leq \exp(-2\beta_i w_s)$, we have

$$\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \lesssim \exp\left(-2w_s \sum_{i=0}^{t} \beta_i\right) \left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 \tag{35}$$

$$+ \sum_{i=0}^{t} \exp\left(-2w_s \sum_{k=i+1}^{t} \beta_k\right) \beta_i^2 \tag{36}$$

$$+ \tau_{\beta_t} \sum_{i=0}^{\tau_{\beta_t}} \exp\left(-2w_s \sum_{k=i+1}^{t} \beta_k\right) \beta_i \tag{37}$$

$$+ \left(\tau_{\beta_t}^2 \beta_{i-\tau_{\beta_t}} + \tau_{\beta_t} \beta_{i-\tau_{\beta_t}}\right) \sum_{i=\tau_{\beta_t}+1}^{t} \exp\left(-2w_s \sum_{k=i+1}^{t} \beta_k\right) \beta_i. \tag{38}$$

We now provide a bound for each of the term above. For the first term, we have

$$\begin{aligned}(35) &= \exp\left\{-2w_s\beta_0 \sum_{i=0}^{t} \frac{1}{(i+1)^s}\right\} \left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 \\ &\lesssim \exp\left[\frac{-2w_s\beta_0}{(1-s)}\left\{(t+1)^{1-s} - 1\right\}\right] \left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2. \end{aligned} \tag{39}$$

For the second term, we have

$$\begin{aligned}(36) &\leq 2G^2(K_b e^{-w_s \sum_{k=0}^{t} \beta_k} + \beta_t) \frac{\exp(w_s\beta_0)}{2w_s} \qquad \text{(by Lemma D.1)} \\ &\lesssim \exp\left[\frac{-w_s\beta_0}{(1-s)}\{(t+1)^{1-s} - 1\}\right] + \beta_t. \end{aligned} \tag{40}$$

For the third term, by Lemma D.2, we get

$$(37) \lesssim \tau_{\beta_t} \exp\left[\frac{-2w_s\beta_0}{(1-s)}\{(t+1)^{1-s} - (\tau_{\beta_t}+1)^{1-s}\}\right]. \tag{41}$$

For the last term, by Lemma D.2, we yield

$$(38) \lesssim \left(\tau_{\beta_t}^2 + \tau_{\beta_t}\right) \beta_{t-\tau_{\beta_t}}. \tag{42}$$

Combining (39)-(42) we get,

$$\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^\star\right\|_2^2 \lesssim \exp\left[\frac{-2w_s\beta_0}{(1-s)}\{(t+1)^{1-s} - 1\}\right] \left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^\star\right\|_2^2 + \mathcal{O}\left\{\beta_{t-\tau_{\beta_t}}\left(\tau_{\beta_t}^2 + \tau_{\beta_t}\right)\right\}$$

completing the proof for the standard SARSA finite-time bound.

Self-tuned SARSA: we next obtain the finite-time error bound of the projected self-tuned SARSA iterates with a step-size sequence of the form $\beta_i = \frac{\beta_0}{(i+1)^s},\ s \in (0.5, 1),\ i \in \mathbb{N}$. With a slight abuse of notation, let $\tilde{\beta}_i := \frac{\beta_i}{1+\beta_i\|\boldsymbol{\phi}(\boldsymbol{S}_i^{\pi_{i-2}}, A_i^{\pi_{i-1}})\|_2^2}$. Consider

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 &= \mathbb{E}\left\|\Pi_r\{\widehat{\boldsymbol{\theta}}_t^{\text{st}} + \tilde{\beta}_t g_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}})\} - \Pi_r\boldsymbol{\theta}^\star\right\|_2^2 \\ &\leq \mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_t^{\text{st}} + \tilde{\beta}_t g_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}}) - \boldsymbol{\theta}^\star\right\|_2^2 \\ &\leq \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_t^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2\left\|g_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}})\right\|_2^2 + 2\tilde{\beta}_t\langle\widehat{\boldsymbol{\theta}}_t^{\text{st}} - \boldsymbol{\theta}^\star, g_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}})\rangle\right\} \\ &= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_t^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + \beta_t^2\left\|g_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}})\right\|_2^2 + 2\tilde{\beta}_t\langle\widehat{\boldsymbol{\theta}}_t^{\text{st}} - \boldsymbol{\theta}^\star, \bar{g}(\widehat{\boldsymbol{\theta}}_t^{\text{st}}) - \bar{g}(\boldsymbol{\theta}^\star)\rangle + 2\tilde{\beta}_t\Lambda_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}})\right\}\end{aligned}$$

where the first inequality is thanks to the non-expansiveness of Π_r, and the second inequality is due to $\tilde{\beta}_t \leq \beta_t$. In the last equality, we used $\bar{g}(\boldsymbol{\theta}^\star) = 0$. From Lemma C.1 and Lemma C.2, we obtain

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 &\leq \mathbb{E}\left\{(1 - 2\tilde{\beta}_t w_s)\left\|\widehat{\boldsymbol{\theta}}_t^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2\right\} + \beta_t^2 G^2 + 2\beta_t \mathbb{E}\Lambda_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}}) + 2\mathbb{E}\left\{(\tilde{\beta}_t - \beta_t)\Lambda_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}})\right\} \\ &\leq \left(1 - \frac{2\beta_t w_s}{1+\beta_t}\right)\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_t^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + 5\beta_t^2 G^2 + 2\beta_t \mathbb{E}\Lambda_t(\widehat{\boldsymbol{\theta}}_t^{\text{st}}).\end{aligned}$$

where in the second inequality, we used the bound $\tilde{\beta}_t \geq \frac{\beta_t}{1+\beta_t}$, which holds almost surely as well as the fact $|\tilde{\beta}_t - \beta_t| \leq \beta_t^2$ with Lemma C.3. Thanks to Lemma C.7, as long as we use at least four basis functions, we know that $w_s \in (0, 1/2]$ and hence we have

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 \leq \left\{\prod_{i=0}^{t}\left(1 - \frac{2w_s\beta_i}{1+\beta_i}\right)\right\}\left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + 5G^2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\beta_i^2 \\ + 2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\mathbb{E}\{\Lambda_i(\widehat{\boldsymbol{\theta}}_i^{\text{st}})\}\beta_i\end{aligned}$$

Applying Lemma C.6, we have

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 &\lesssim \left\{\prod_{i=0}^{t}\left(1 - \frac{2w_s\beta_i}{1+\beta_i}\right)\right\}\left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + \sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\beta_i^2 \\ &\quad + \sum_{i=0}^{\tau_{\beta_t}}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\left(\sum_{k=0}^{\tau_{\beta_t}-1}\beta_k\right)\beta_i \\ &\quad + \sum_{i=\tau_{\beta_t}+1}^{t}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\left(\tau_{\beta_t}\sum_{k=i-\tau_{\beta_t}}^{i-1}\beta_k\right)\beta_i \\ &\lesssim \left\{\prod_{i=0}^{t}\left(1 - \frac{2w_s\beta_i}{1+\beta_i}\right)\right\}\left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + \sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\beta_i^2 \\ &\quad + \tau_{\beta_t}\sum_{i=0}^{\tau_{\beta_t}}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\beta_i \\ &\quad + \left(\tau_{\beta_t}^2\beta_{i-\tau_{\beta_t}} + \tau_{\beta_t}\beta_{i-\tau_{\beta_t}}\right)\sum_{i=\tau_{\beta_t}+1}^{t}\left\{\prod_{k=i+1}^{t}\left(1 - \frac{2w_s\beta_k}{1+\beta_k}\right)\right\}\beta_i\end{aligned}$$

Since $1 - \frac{2w_s\beta_i}{1+\beta_i} \leq \exp\left(-\frac{2w_s\beta_i}{1+\beta_i}\right)$ and

$$\sum \frac{\beta_k}{1+\beta_k} = \sum \beta_k - \sum \frac{\beta_k^2}{1+\beta_k} = \sum \beta_k + \mathcal{O}(1)$$

for $\beta_k = \beta_0/(k+1)^s$, $s \in (0.5, 1)$, we have

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 &\lesssim \exp\left(-2w_s\sum_{i=0}^{t}\beta_i\right)\left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 && (43) \\ &\quad + \sum_{i=0}^{t}\exp\left(-2w_s\sum_{k=i+1}^{t}\beta_k\right)\beta_i^2 && (44) \\ &\quad + \tau_{\beta_t}\sum_{i=0}^{\tau_{\beta_t}}\exp\left(-2w_s\sum_{k=i+1}^{t}\beta_k\right)\beta_i && (45) \\ &\quad + \left(\tau_{\beta_t}^2\beta_{i-\tau_{\beta_t}} + \tau_{\beta_t}\beta_{i-\tau_{\beta_t}}\right)\sum_{i=\tau_{\beta_t}+1}^{t}\exp\left(-2w_s\sum_{k=i+1}^{t}\beta_k\right)\beta_i. && (46)\end{aligned}$$

We now provide a bound for each of the term above. For the first term, we have

$$(43) = \exp\left\{-2w_s\beta_0 \sum_{i=0}^{t} \frac{1}{(i+1)^s}\right\} \left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2$$
$$\lesssim \exp\left[\frac{-2w_s\beta_0}{(1-s)}\left\{(t+1)^{1-s} - 1\right\}\right] \left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2. \tag{47}$$

For the second term, by Lemma D.1, we have

$$(44) \lesssim \exp\left[\frac{-2w_s\beta_0}{2(1-s)}\{(t+1)^{1-s} - 1\}\right] + \beta_t. \tag{48}$$

For the third term, by Lemma D.2, we get

$$(45) \lesssim \tau_{\beta_t} \exp\left[\frac{-2w_s\beta_0}{(1-s)}\{(t+1)^{1-s} - (\tau_{\beta_t}+1)^{1-s}\}\right]. \tag{49}$$

For the last term, by Lemma D.2, we yield

$$(46) \lesssim \left(\tau_{\beta_t}^2 + \tau_{\beta_t}\right)\beta_{t-\tau_{\beta_t}}. \tag{50}$$

Combining (47)-(50) we get,

$$\mathbb{E}\left\|\widehat{\boldsymbol{\theta}}_{t+1}^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 \lesssim \exp\left[\frac{-2w_s\beta_0}{(1-s)}\{(t+1)^{1-s} - 1\}\right] \left\|\widehat{\boldsymbol{\theta}}_0^{\text{st}} - \boldsymbol{\theta}^\star\right\|_2^2 + \mathcal{O}\left\{\beta_{t-\tau_{\beta_t}}\left(\tau_{\beta_t}^2 + \tau_{\beta_t}\right)\right\}$$

completing the proof.

D Auxiliary Lemma

Lemma D.1. *For $t \in \mathbb{N}$, let $\beta_t = \frac{\beta_0}{(t+1)^s}$ and $s \in (0.5, 1)$. With $q > 0$,*

$$\sum_{i=0}^{t}\left(e^{-q\sum_{k=i+1}^{t}\beta_k}\right)\beta_i^2 \le 2\left(K_b e^{-\frac{q}{2}\sum_{k=0}^{t}\beta_k} + \beta_t\right)\frac{e^{\frac{q\beta_0}{2}}}{q},$$

where $K_b = \beta_0 e^{\frac{q}{2}\sum_{k=0}^{i_0}\beta_k}$ for some $i_0 \in \mathbb{N}$.

Proof. Let $T_t = \sum_{i=0}^{t-1}\beta_i$ and use the convention $\sum_{k=t+1}^{t}\beta_k = 0$ and $\sum_{k=t+1}^{t}\beta_k^2 = 0$. Notice that

$$\sum_{i=0}^{t}\left(e^{-\frac{q}{2}\sum_{k=i+1}^{t}\beta_k}\right)\beta_i \le \left(\sup_{i\ge 0} e^{\frac{q}{2}\beta_i}\right)\left\{\sum_{i=0}^{t}\left(e^{-\frac{q}{2}\sum_{k=i}^{t}\beta_k}\right)\beta_i\right\}$$
$$= \left(\sup_{i\ge 0} e^{\frac{q}{2}\beta_i}\right)\left\{\sum_{i=0}^{t}\left(e^{-\frac{q}{2}(T_{t+1}-T_i)}\right)\beta_i\right\}$$
$$\le \left(\sup_{i\ge 0} e^{\frac{q}{2}\beta_i}\right)\int_0^{T_{t+1}} e^{-\frac{q}{2}(T_{t+1}-s)}ds$$
$$\le \left(\sup_{i\ge 0} e^{\frac{q}{2}\beta_i}\right)\frac{2}{q} \le \frac{2e^{\frac{q\beta_0}{2}}}{q}, \tag{51}$$

where we have used the definition of the left-Riemann sum in the first inequality. The last inequality is due to the fact that $\{\beta_t\}$ is a non-increasing sequence. Now consider

$$\sum_{i=0}^{t}\left(e^{-q\sum_{k=i+1}^{t}\beta_k}\right)\beta_i^2 \le \sup_{0\le i\le t}\left(\beta_i e^{-\frac{q}{2}\sum_{k=i+1}^{t}\beta_k}\right)\left\{\sum_{i=0}^{t}\left(e^{-\frac{q}{2}\sum_{k=i+1}^{t}\beta_k}\right)\beta_i\right\}$$
$$\le \sup_{0\le i\le t}\left(\beta_i e^{-\frac{q}{2}\sum_{k=i+1}^{t}\beta_k}\right)\frac{2e^{\frac{q\beta_0}{2}}}{q} \tag{52}$$

where the last inequality follows from (51). Note that $\beta_i e^{-\frac{q}{2}\sum_{k=i+1}^{t}\beta_k}$ is eventually increasing, i.e., after some time $i_0 \in \mathbb{N}$, for all $t \geq i_0$, we have

$$\sup_{i_0 \leq i \leq t} \left\{ \beta_i \exp\left(-\frac{q}{2} \sum_{k=i+1}^{t} \beta_k \right) \right\} \leq \beta_t,$$

where we used the convention $\sum_{k=t+1}^{t} \beta_k = 0$. Therefore, we have

$$\begin{aligned}
(52) &\leq \left\{ \sup_{0 \leq i \leq i_0} \left(\beta_i e^{-\frac{q}{2}\sum_{k=i+1}^{t}\beta_k} \right) + \beta_t \right\} \frac{2e^{\frac{q\beta_0}{2}}}{q} \\
&\leq \left\{ e^{-\frac{q}{2}\sum_{k=0}^{t}\beta_k} \sup_{0 \leq i \leq i_0} \left(\beta_i e^{\frac{q}{2}\sum_{k=0}^{i}\beta_k} \right) + \beta_t \right\} \frac{2e^{\frac{q\beta_0}{2}}}{q} \\
&\leq \left(K_b e^{-\frac{q}{2}\sum_{k=0}^{t}\beta_k} + \beta_t \right) \frac{2e^{\frac{q\beta_0}{2}}}{q},
\end{aligned}$$

where $K_b = \beta_0 e^{\frac{q}{2}\sum_{k=0}^{i_0}\beta_k}$. □

Lemma D.2. *For $t \in \mathbb{N}$, let $\beta_t = \frac{\beta_0}{(t+1)^s}$ and $s \in (0.5, 1)$. With $q > 0$ and $\tau \in \{0, \cdots, t\}$,*

1. $\sum_{i=0}^{\tau} e^{-q\sum_{k=i+1}^{t}\beta_k}\beta_i \leq \frac{e^{q\beta_0}}{q} e^{-\frac{q\beta_0}{(1-s)}\left\{(t+1)^{1-s}-(1+\tau)^{1-s}\right\}}$.
2. $\sum_{i=\tau_{\beta_t}+1}^{t} \left(e^{-q\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\beta_i \right) \leq \left[e^{\frac{-q\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}} D_\beta \mathbb{I}_{\left(\tau_{\beta_t}+1 < i_\beta\right)} + \beta_{t-\tau_{\beta_t}} \right] \frac{2e^{q\beta_0/2}}{q}$

where $D_\beta = \exp\{(q/2)\sum_{k=0}^{i_\beta}\beta_k\}\beta_0$ for some $i_\beta \in \mathbb{N}$.

Proof. Let $T_t = \sum_{i=0}^{t-1}\beta_i$ and use the convention $\sum_{k=t+1}^{t}\beta_k = 0$. For the first statement,

$$\begin{aligned}
\sum_{i=0}^{\tau} e^{-q\sum_{k=i+1}^{t}\beta_k}\beta_i &\leq \max_{i\geq 0}\left(e^{q\beta_i}\right) \sum_{i=0}^{\tau} e^{-q\sum_{k=i}^{t}\beta_k}\beta_i = e^{q\beta_0}\sum_{i=0}^{\tau} e^{-q(T_{t+1}-T_i)}\beta_i \\
&\leq e^{q\beta_0}\int_0^{T_{\tau+1}} e^{-q(T_{t+1}-s)}ds \leq \frac{e^{q\beta_0}}{q} e^{-q(T_{t+1}-T_{\tau+1})} \\
&= \frac{e^{q\beta_0}}{q} e^{-q\beta_0\sum_{k=\tau+1}^{t}1/(1+k)^s} \leq \frac{e^{q\beta_0}}{q} e^{-\frac{q\beta_0}{(1-s)}\left\{(t+1)^{1-s}-(1+\tau)^{1-s}\right\}}.
\end{aligned}$$

For the second statement, first notice that

$$\begin{aligned}
\sum_{i=\tau_{\beta_t}+1}^{t} e^{-q\sum_{k=i+1}^{t}\beta_k}\beta_i &\leq \max_{i\geq 0}\left(e^{q\beta_i}\right) \sum_{i=\tau_{\beta_t}+1}^{t} e^{-q\sum_{k=i}^{t}\beta_k}\beta_i = e^{q\beta_0}\sum_{i=\tau_{\beta_t}+1}^{t} e^{-q(T_{t+1}-T_i)}\beta_i \\
&\leq e^{q\beta_0}\int_{T_{\tau_{\beta_t}+1}}^{T_{t+1}} e^{-q(T_{t+1}-s)}ds = \frac{e^{q\beta_0}}{q}\left\{1 - e^{-q\left(T_{t+1}-T_{\tau_{\beta_t}+1}\right)}\right\} \leq \frac{e^{q\beta_0}}{q}.
\end{aligned} \tag{53}$$

Then, we have

$$\begin{aligned}
\sum_{i=\tau_{\beta_t}+1}^{t} e^{-q\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\beta_i &\leq \max_{i\in\left[\tau_{\beta_t}+1,t\right]}\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\} \sum_{i=\tau_{\beta_t}+1}^{t} e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_i \\
&\leq \max_{i\in\left[\tau_{\beta_t}+1,t\right]}\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\}\frac{2e^{q\beta_0/2}}{q}
\end{aligned} \tag{54}$$

where the second inequality follows from (53). To bound the first term in (54), note that the sequence $\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\}_{i\in\mathbb{N}}$ is eventually increasing. In other words, there exists $i_\beta \in \mathbb{N}$ such that,

$$\max_{i\in\left[\tau_{\beta_t}+1,t\right]}\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\} = \beta_{t-\tau_{\beta_t}} \quad \text{if} \quad \tau_{\beta_t}+1 \geq i_\beta.$$

If $\tau_{\beta_t}+1 < i_\beta$, then

$$\begin{aligned}
&\max_{i\in\left[\tau_{\beta_t}+1,t\right]}\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\} \\
&\leq \max_{i\in[2\tau_\beta+1,i_\beta]}\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\} + \max_{i\in[i_\beta+1,t]}\left\{e^{(-q/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\right\} \\
&\leq e^{(-q/2)\sum_{k=0}^{t}\beta_k}\max_{i\in\left[\tau_{\beta_t}+1,i_\beta\right]}\left\{e^{(q/2)\sum_{k=0}^{i}\beta_k}\beta_{i-\tau_{\beta_t}}\right\} + \beta_{t-\tau_{\beta_t}} \\
&\leq e^{-(q/2)\sum_{k=0}^{t}\beta_k}e^{(q/2)\sum_{k=0}^{i_\beta}\beta_k}\beta_0 + \beta_{t-\tau_{\beta_t}} \\
&\leq e^{\frac{-q\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}}D_\beta + \beta_{t-\tau_{\beta_t}}
\end{aligned}$$

where $D_\beta = e^{(q/2)\sum_{k=0}^{i_\beta}\beta_k}\beta_0$. Combining everything, we get

$$\sum_{i=\tau_{\beta_t}+1}^{t} e^{-q\sum_{k=i+1}^{t}\beta_k}\beta_{i-\tau_{\beta_t}}\beta_i \leq \left[e^{\frac{-q\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}}D_\beta\mathbb{I}_{\left(\tau_{\beta_t}+1<i_\beta\right)} + \beta_{t-\tau_{\beta_t}}\right]\frac{2e^{q\beta_0/2}}{q}.$$

□

E Experimental Details

E.1 Discrete State Environments

Environment Descriptions.

- **Cliff Walking** is a 4×12 gridworld from `gymnasium` (Towers et al., 2024) with 48 discrete states representing grid positions. The agent starts at the bottom-left corner and must reach the bottom-right goal while avoiding a cliff along the bottom edge. The action space consists of 4 movement directions (up, down, left, right). Stepping off the cliff incurs a reward of -100 and resets the agent to the start, while each non-terminal step incurs a reward of -1.
- **Taxi** is a 5×5 gridworld from `gymnasium` with 500 discrete states encoding the taxi position, passenger location, and destination. The action space consists of 6 actions (move in 4 directions, pick up, drop off). The agent receives a reward of $+20$ for a successful delivery, -10 for illegal pickup or dropoff attempts, and -1 for each step taken.
- **Blackjack** is instantiated via `gymnasium`. The state space consists of 704 discrete states encoding the player's current sum, the dealer's showing card, and whether the player holds a usable ace. The action space consists of 2 actions (hit or stick). The agent is trained for 100,000 episodes with a maximum of 100,000 steps per episode.
- **Frozen Lake** uses the 4×4 non-slippery variant from `gymnasium` with 16 discrete states representing grid positions. The action space consists of 4 movement actions (left, down, right, up). The agent receives a reward of $+1$ upon reaching the goal and 0 otherwise. The agent is trained for 500 episodes with a maximum of 1,000 steps per episode.
- **GridWorld** is a custom 5×5 grid environment with 25 discrete states. The agent starts at position $(0, 0)$ and aims to reach the goal at $(4, 4)$, while avoiding three hole states at $(1, 1)$, $(2, 3)$, and $(3, 1)$. The action space consists of 4 movement directions. The agent receives a reward of $+1$ for reaching the goal, -1 for falling into a hole, and -0.1 for each step taken. The agent is trained for 100 episodes with a maximum of 200 steps per episode.

- **Dynamic Pricing** is a custom environment with 36 discrete states encoding the cross product of 6 demand buckets and 6 inventory buckets, where demand evolves stochastically at each step. The action space consists of 5 price tiers $\{5, 10, 15, 20, 25\}$. The reward at each step is defined as the revenue from units sold minus a holding cost proportional to current inventory. Episodes terminate after 20 steps or when inventory is depleted. The agent is trained for 300 episodes with a maximum of 300 steps per episode.

Experimental Settings for Cliff Walking and Taxi. Both environments are trained for 400 episodes with a maximum of 10,000 steps per episode, and results are averaged over 50 independent runs. A constant step-size β is swept over the grid $\{0.2, 0.4, \ldots, 2.0\}$. Furthermore, we considered a decreasing step-size of the form $\beta_t = \beta_0/(t+1)^{2/3}$, where β_0 is swept over the grid $\{0.1, 1000, 2000, 3000\}$. We use a discount factor $\gamma = 0.99$ and projection radius $r = 5000$. Following Mnih et al. (2015), the exploration parameter decays linearly from $\epsilon_0 = 0.1$ to $\epsilon_{400} = 0.01$ over the 400 episodes. Q-learning and ST-Q-learning employ an ϵ-greedy policy, while SARSA and ST-SARSA employ an ϵ-softmax policy with temperature $\iota = 0.05$.

Experimental Settings for Blackjack, Frozen Lake, GridWorld, and Dynamic Pricing. All four environments use a decreasing step-size schedule of the form $\beta_t = \beta_0/(t+1)^{2/3}$, where t denotes the global update counter and $\beta_0 \in (0, 300]$ is the initial step-size. We also consider constant step sizes from the grid $\{0.1, 1.0, \ldots, 10.0\}$. Results are averaged over 50 independent runs. We use discount factor $\gamma = 0.99$ (except Blackjack, where $\gamma = 0.95$) and projection radius $r = 5000$. The exploration parameter decays linearly from $\epsilon_0 = 1.0$ to $\epsilon_{\text{final}} = 0.05$ over the first 100 episodes. The same policy assignments as above apply. Environment-specific episode counts are provided in the respective environment descriptions above.

E.2 Continuous State Environments

Environment Descriptions.

- **Acrobot** (`Acrobot-v1`) is a classic underactuated control problem from `gymnasium` with a 6-dimensional continuous state space encoding the joint angles and angular velocities of a two-link pendulum. The agent applies one of 3 discrete torques (left, neutral, right) to the joint between the two links and aims to swing the free end above a target height as quickly as possible. Each step incurs a reward of -1 until the goal is reached or the episode terminates after 1000 steps, so shorter episode lengths directly reflect more efficient policies.
- **Mountain Car** (`MountainCar-v0`) is a continuous control benchmark with a 2-dimensional state space encoding the car's position and velocity. The agent selects one of 3 discrete actions (push left, neutral, push right) and must leverage momentum to drive an underpowered car up a steep hill. Each step incurs a reward of -1, and the episode terminates upon reaching the goal or after 1000 steps, making episode length the primary performance indicator.

Function Approximation. Since both environments possess continuous state spaces, tabular methods are not directly applicable. We employ a linear function approximator for the action-value function $Q(s, a)$, parameterized by a separate weight vector. State observations are first standardized using a `StandardScaler` fitted on 10,000 randomly sampled observations. The standardized states are then mapped to a 400-dimensional feature vector via four RBF samplers with bandwidths $\sigma \in \{0.5, 1.0, 2.0, 5.0\}$, each contributing 100 components. The Q-value for state-action pair (s, a) is then computed as $Q(s, a) = \phi(s, a)^\top \boldsymbol{\theta}$, where $\phi(s) \in \mathbb{R}^{400}$ denotes the feature vector.

Experimental Settings. Both environments are trained for 30 episodes and results are averaged over 50 independent runs. We employ a decreasing step-size schedule of the form $\beta_t = \beta_0/(t+1)^{2/3}$, where β_0 is the initial step-size, swept over $\{0.1, 2.0, 4.0, 6.0, 8.0, 10.0\}$. We use a discount factor $\gamma = 0.99$, projection radius $r = 1000$, and a fixed exploration parameter $\epsilon = 0.1$ (no decay) throughout training. Q-learning and ST-Q-learning use an ϵ-greedy policy, while SARSA and ST-SARSA use an ϵ-softmax policy with temperature $\iota = 0.05$. Performance is assessed via the average episode length as a direct measure of policy quality.

F Additional Tabular Environment Experimental Results

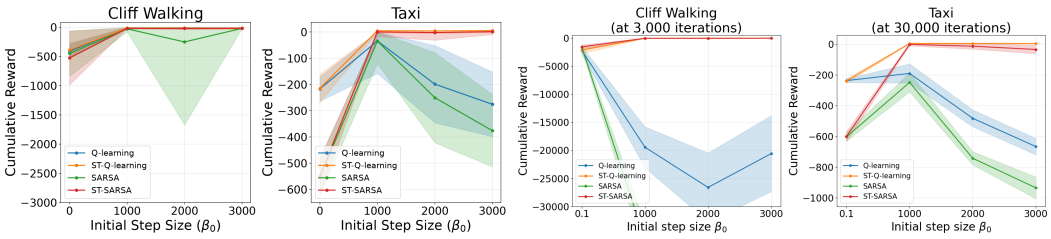

Figure 7: Performance on Cliff Walking and Taxi environments with decreasing step-size.

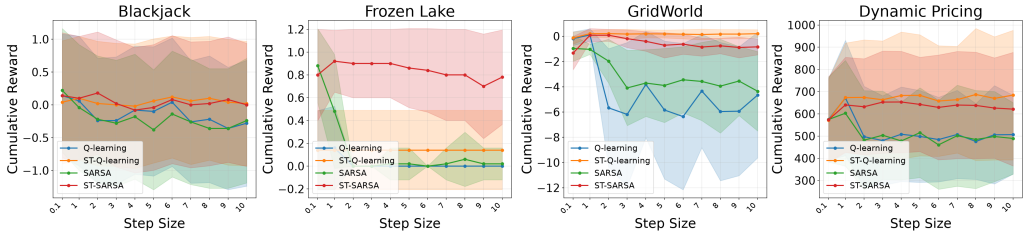

Figure 8: Performance on tabular environments with constant step-size.

Figures 7 and 8 further illustrate the robustness of the self-tuned methods to aggressive step-size choices. Under the decreasing schedule $\beta_t = \beta_0/(t+1)^{2/3}$, Figure 7 shows that increasing β_0 to very large values leads to substantial performance degradation and increased variability for the standard methods, whereas self-tuned Q-learning and self-tuned SARSA remain stable and maintain strong performance. A similar pattern is observed in Figure 8 under constant step-sizes: as the step-size increases, the performance of the standard algorithms can deteriorate considerably, while the self-tuned variants remain comparatively insensitive across the range considered. These observations are consistent with the qualitative robustness suggested by our theory under decreasing step-sizes. Although our finite-time analysis does not directly cover constant step-sizes, Figure 8 shows that the empirical reward performance of the self-tuned methods remains stable even for substantially larger step-sizes. This suggests that, in practice, the self-tuned procedures exhibit greater efficiency than what is captured by our theoretical guarantees.

G Extensions to Deep Q-learning

Environment Descriptions.

- **Freeway** uses the `ALE/Freeway-v5` environment with RAM observations. The agent controls a chicken attempting to cross multiple lanes of traffic, receiving reward according to increases in the game score when successful crossings are completed. We use the reduced Atari action space, a frame skip of 4, and normalize the RAM observations by dividing by 255. The agent is trained for 50 episodes with a maximum of 2,000 interaction steps per episode.
- **Breakout** uses the `ALE/Breakout-v5` environment with RAM observations. The agent controls a paddle to keep the ball in play and remove bricks, with rewards corresponding to increases in the game score from breaking bricks. We use the reduced Atari action space, a frame skip of 4, and normalize the RAM observations by dividing by 255. The agent is trained for 50 episodes with a maximum of 2,000 interaction steps per episode.

Experimental Settings. We further evaluated deep Q-learning (Mnih et al., 2015; Van Hasselt et al., 2016), self-tuned deep Q-learning, and Q-AMSGrad (Weng et al., 2020) on the Atari Freeway and Breakout environments using RAM observations. The Q-function is represented by a two-hidden-layer

neural network with 128 ReLU units per layer. We run 50 independent trials of 50 episodes and use the decreasing nominal step size $\beta_t = \frac{\beta_0}{(t+1)^{2/3}}, \beta_0 \in \{0.1, 2.0, 4.0\}$ with ε-greedy exploration decaying from 0.1 to 0.01. For the self-tuned method, the effective step size is $\widetilde{\beta}_t = \frac{\beta_t}{1+\beta_t \|\nabla_{\boldsymbol{\theta}} Q(s_t, a_t; \boldsymbol{\theta}_t)\|_2^2}$, where the feature norm appearing in the linear setting is replaced by the squared parameter-gradient norm of the nonlinear Q-function.

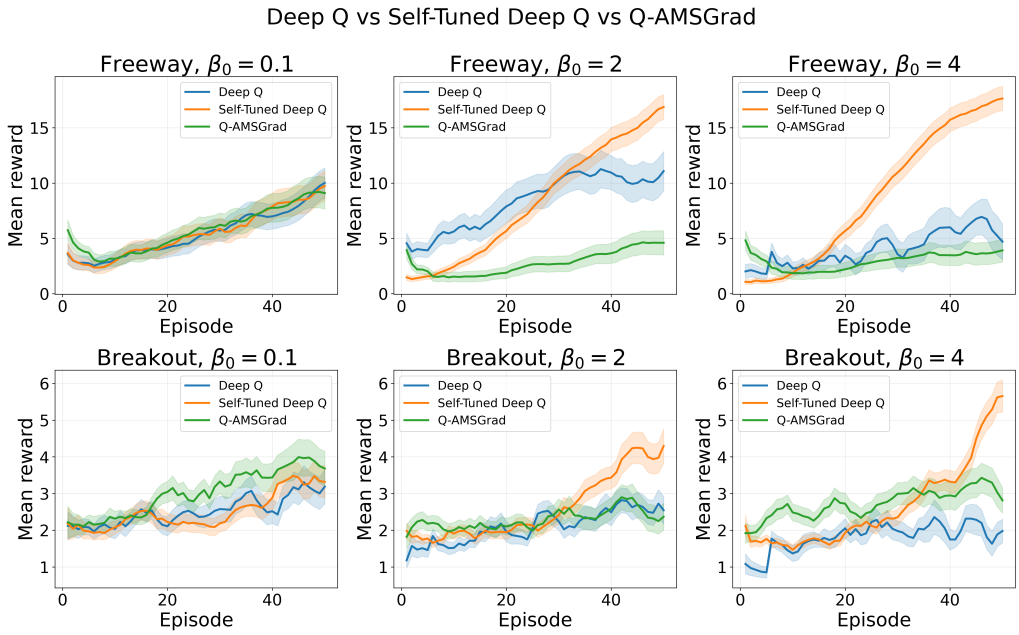

Figure 9: Performance comparison of deep Q-learning, self-tuned deep Q-learning, and Q-AMSGrad.