Policy Evaluation for Variance in Average Reward Reinforcement Learning

Shubhada Agrawal [1] **Prashanth L. A.** [2] **Siva Theja Maguluri** [1]

Abstract

We consider an average reward reinforcement learning (RL) problem and work with asymptotic variance as a risk measure to model safety-critical applications. We design a temporal-difference (TD) type algorithm tailored for policy evaluation in this context. Our algorithm is based on linear stochastic approximation of an equivalent formulation of the asymptotic variance in terms of the solution of the Poisson equation. We consider both the tabular and linear function approximation settings, and establish $\tilde{O}(1/k)$ finite time convergence rate, where k is the number of steps of the algorithm. Our work paves the way for developing actor-critic style algorithms for variance-constrained RL. To the best of our knowledge, our result provides the first sequential estimator for asymptotic variance of a Markov chain with provable finite sample guarantees, which is of independent interest.

1. Introduction

We consider an average reward Markov decision process (MDP) (Bertsekas, 2012; Puterman, 1994). We operate in a reinforcement learning (RL) (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 2018) framework, where the transition dynamics of the underlying MDP are unknown. The RL algorithm can obtain a sample of the MDP under any given policy, which specifies how actions are chosen in a given state. The traditional goal in an average reward RL problem is to find a policy that maximizes the long run average reward. While the need to optimize over average reward is well motivated, applications in safety-critical domains, for example, healthcare or finance, it is also crucial to control adverse outcomes. As an example in the financial domain, one can consider the portfolio optimization problem, where the objective is to find an investment strategy that maximizes the return (expected value), while keeping the variability under control through a constraint. This motivates the study of risk-sensitive RL, where a risk measure is incorporated either in the objective or as a constraint, see L.A. & Fu (2022) for a recent survey.

Within the domain of risk-sensitive MDPs, several risk measures have been considered, namely variance (Sobel, 1982; Filar et al., 1989), conditional value-at-risk (Chow & Ghavamzadeh, 2014), exponential utility (Whittle, 1990), the general class of coherent risk measures (Tamar et al., 2015), and cumulative prospect theory (Prashanth et al., 2016), which is not coherent. The choice of the risk measure largely depends on the application at hand. Nevertheless, variance is a popular risk measure that has been studied extensively in the literature.

Several previous works incorporate variance as a risk measure in a constrained setting, where the goal is to maximize the average reward (which is an expectation), while ensuring a certain bound on the variance. This is the so-called "mean-variance tradeoff", considered in the seminal work of Markowitz (Markowitz, 1952), and later in MDP contexts, cf. (Mandl, 1971; Sobel, 1982; Filar et al., 1989). An alternative to such a formulation is to consider the exponential utility formulation, see (Arrow, 1971; Howard & Matheson, 1972), where one optimizes the exponential. The constrained formulation is preferred over the exponential utility for two reasons. First, the mean-variance tradeoff can be controlled directly through a parameter that is a bound on the variance, while this trade-off is implicit in an exponential utility formulation. Second, the algorithms for the latter formulation do not extend easily when one considers feature-based representations and function approximation, see L.A. & Fu (2022, Section 7.2) for a detailed discussion.

We consider variance as the risk measure in a average reward RL framework. Broadly, for a given policy, two different notions of variance are suggested for an average reward MDP in (Filar et al., 1989). The first notion is the asymptotic variance, while the second one is the per-period stationary variance. The latter has been studied in an RL setting in (Prashanth & Ghavamzadeh, 2016), while the former has hardly been investigated in the literature, to the best of our

[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, USA. [2]Department of Computer Science and Engineering, Indian Institute of Technology Madras, India. Correspondence to: Shubhada Agrawal <sagrawal362@gatech.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

knowledge. We use the asymptotic variance as a measure of the risk associated with a given policy, which is the variance of the random variable whose mean is typically optimized. This asymptotic variance can be decomposed into two additive terms, where the first term coincides with per-step variance, while the other term involves correlations between states across time periods. In a setting where the state sequence is independent, the second term is zero, while it is not in any non-trivial MDP. A mean-variance optimization formulation would consider maximizing the expectation of a random variable which represents the long-run average reward, and it is natural to consider the variance of this random variable. This motivates our study of the asymptotic variance in a RL framework. Also see Section 3.2.

In this work, we consider the problem of policy evaluation for asymptotic variance, where the goal is to estimate the asymptotic variance associated with the Markov chain induced by a given stationary policy. Note that this goal is very different from studying the variance of iterates of any particular learning algorithm like temporal-difference (TD), or designing algorithms with minimum variance of the iterates of the algorithm, studied extensively in (Devraj & Meyn, 2017; Yin & Wang, 2020; Chen et al., 2020; Hao et al., 2021). To further elaborate on this distinction, for concreteness, consider the setting of MDPs, where the transition probabilities are known. Here, for evaluating a given policy, one can use value iteration (VI). Since the MDP is known, there is no variance associated with the iterates of VI. However, the asymptotic variance that we consider in this work (formally defined in Section 3) is still well-defined, and exists due to inherent randomness associated with the probabilistic transitions of the MDP. This work focuses on estimating this asymptotic variance.

The problem of estimating the asymptotic variance is a vital sub-problem in mean-variance policy optimization, for instance, as a critic in an actor-critic framework, cf. (Prashanth & Ghavamzadeh, 2016). For the discounted RL setting, a TD type algorithm for estimating variance has been proposed/analyzed in (Tamar et al., 2013), while a TD algorithm with provable finite sample guarantees, which caters to the variance risk measure in an average reward RL setting, is not present in the literature to the best of our knowledge.

1.1. Contributions

We now briefly describe the main contributions of this work.

1. We design a novel TD like linear stochastic approximation (SA) algorithm in both tabular and linear function approximation settings for estimating the asymptotic variance associated with the Markov Chain induced by a given stationary policy. Our algorithms are based on an equivalent formulation of the asymptotic variance of a given stationary policy in terms of the corresponding solution of the Poisson equation.
2. We develop the first finite sample error bounds for the policy evaluation problem for asymptotic variance in a tabular setting, proving $\tilde{O}(1/k)$ rate of convergence for the mean-squared error, where k is the time step. Here, $\tilde{O}(\cdot)$ notation hides $\log k$ and lower order dependencies.
3. We provide the first finite sample error bounds for the policy evaluation problem for asymptotic variance with linear function approximation. Again, we demonstrate $\tilde{O}(1/k)$ rate of convergence for the proposed algorithm. However, in this setting, our estimate for asymptotic variance suffers from an approximation error, which we characterize.
4. Estimating asymptotic variance of functions defined on state space of a Markov chain is a classical problem in statistics as it is useful in statistical inference of the mean (Wu, 2009). To the best of our knowledge, we develop the first fully-sequential estimator with finite sample guarantees on estimation error. Our contributions may be of independent interest to statistics community.

1.2. Related Literature

MDPs: Average Reward and Risk-sensitivity. MDPs have a long history. We refer the reader to classical books (Puterman, 1994; Bertsekas, 2012) for a textbook introduction to MDPs in general, and to (Howard, 1960; Blackwell, 1962; Brown, 1965; Veinott, 1966; Arapostathis et al., 1993) for an introduction to average-reward MDPs, in particular. Risk-sensitive objectives have also been well-studied in the MDP setting. For instance, see (Sobel, 1982; Filar et al., 1995; Mannor & Tsitsiklis, 2013) for variance in discounted and average reward MDPs, (Borkar & Meyn, 2002; Borkar & Jain, 2010; Whittle, 1990) for the exponential utility formulation, (Ruszczyński, 2010) for Markov risk measures, (Chow & Ghavamzadeh, 2014) for conditional value-at-risk (CVaR). However, in a MDP setting, algorithms require complete knowledge of the underlying model, which may not be feasible in many practical applications.

Risk-neutral RL. In the risk-neutral RL setting, expected value is the sole objective. TD type algorithms have been proposed for policy evaluation in discounted as well as average reward settings, and their asymptotic convergence is shown in (Tsitsiklis & Van Roy, 1996; 1999), respectively. TD algorithms have also been used in actor-critic style algorithms for solving the problem of control, cf. (Konda & Tsitsiklis, 2003; Bhatnagar et al., 2009). Asymptotic convergence of the classical Q-learning algorithm was established in (Borkar & Meyn, 2000; Tsitsiklis, 1994). In the non-asymptotic regime, finite-sample mean-square convergence bounds for classical discounted setting algorithms such as TD, TD(λ), n-step TD, and Q-learning, have been developed in (Chen et al., 2021). On the other hand, in the average

reward setting, finite-sample bounds for TD are derived in (Zhang et al., 2021).

Risk-sensitivity in RL. In a risk-sensitive optimization setting, the goal is either to optimize the usual expected value objective, while factoring a risk measure in a constraint, or optimize a risk measure in the objective. Tail-based risk measures such as variance, CVaR are meaningful to consider as a constraint, while risk measures such as exponential utility and prospect theory can be considered as the optimization objective, since they consider the entire distribution. In the context of RL, variance as a risk measure has been studied earlier in a discounted reward MDP setting in (Mihatsch & Neuneier, 2002; La & Ghavamzadeh, 2013), in a stochastic shortest path setting (Tamar et al., 2013; 2012; 2016), and in an average reward MDP setting in (Prashanth & Ghavamzadeh, 2016). Exponential utility has been explored in an average reward RL setting earlier, see (Borkar, 2010) for a survey and (Moharrami et al., 2022) for a recent contribution. Other risk measures such as CVaR, coherent risk measures, cumulative prospect theory have been explored in an RL setting in the literature, and some representative works include (Prashanth, 2014; Köse & Ruszczyński, 2021; Prashanth et al., 2016; Markowitz et al., 2023).

Variance Estimation of Markov Chain in Statistics. The estimators for the asymptotic variance are well studied in the statistics literature. (Wu, 2009; Flegal & Jones, 2010; Atchadé, 2011; Chien et al., 1997; Robert, 1995) develop efficient Monte-Carlo (MC) based batched or sequential estimators, but do not provide any finite sample guarantees. Benton (2022) proposes a TD like algorithm for estimating the asymptotic variance, with asymptotic convergence guarantees.

2. Average Reward RL

In this section, we formally introduce the problem. We begin by describing the underlying dynamics.

2.1. Markov Decision Process

Consider an infinite-horizon, average-reward MDP specified by $(\mathcal{S}, \mathcal{A}, r, p)$, where $\mathcal{S} = \{1, \ldots, |S|\}$ denotes the finite state-space, and $\mathcal{A} = \{1, \ldots, |A|\}$ denotes the action space. At each time t, the agent is in state $S_t \in \mathcal{S}$, takes an action $A_t \in \mathcal{A}$, receives a reward $r(S_t, A_t)$, and transitions to state S_{t+1}. Here, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the next state S_{t+1} is sampled according to $p(S_t, \cdot, A_t)$, where, $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, is the map that for states s, s' and action a associates probability $p(s, s', a)$ with the transition from state s to s', when action a is taken.

A stationary policy $\mu : \mathcal{S} \to \Sigma_{|\mathcal{A}|}$ is a map from state space to a probability-simplex in $\mathbb{R}^{|\mathcal{A}|}$, i.e., it maps each state to a distribution over the actions. Given a stationary policy μ and a state x, the associated probability of transitioning from state x to x' is given by $P_\mu(x, x') = \sum_{a \in \mathcal{A}} \mu(a|x) p(x, x', a)$.

Assumption 2.1. The sets $\mathcal{S}$ and $\mathcal{A}$ are finite. Under the stationary policy μ, the induced Markov chain $\mathcal{M}_1$ with state space in $\mathcal{S}$ and transition probabilities given by P_μ is irreducible and aperiodic.

This is a standard assumption in literature (see Tsitsiklis & Van Roy (1999); Bertsekas (2012)), and guarantees that each state is visited infinitely often. As a consequence of Assumption 2.1, we have a unique stationary distribution associated with $\mathcal{M}_1$, and the Markov chain starting from any initial distribution, converges to the stationary distribution geometrically-fast (Levin & Peres, 2017, Section 4.3). In particular, let π_μ denote the unique stationary distribution on $\mathcal{S}$ that satisfies $\pi_\mu^T P_\mu = \pi_\mu^T$.

Next, let $X_t := (S_t, A_t)$. Observe that the process $\mathcal{M}_2 := \{X_t\}$ is also a Markov chain with finite states. Let its state space be denoted by $\mathcal{X} \subset \mathcal{S} \times \mathcal{A}$. For any two states $x_1 := (s_1, a_1)$ and $x_2 := (s_2, a_2)$ in $\mathcal{X}$, the probability of transitioning from x_1 to x_2 under $\mathcal{M}_2$ is given by

$$P_2(x_1, x_2) := p(s_1, s_2, a_1)\mu(a_2|s_2). \tag{1}$$

For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, define

$$d_\mu(s, a) := \pi_\mu(s)\mu(a|s).$$

Remark 2.2. Under Assumption 2.1, $\mathcal{M}_2$ has a unique stationary distribution d_μ defined above, and it mixes geometrically-fast.

Remark 2.2 follows from Bhatnagar & Lakshmanan (2016, Proposition 1).

Notation. Let $D_\mu \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the diagonal array of stationary distribution d_μ, i.e., $D_\mu((s, a), (s, a)) = d_\mu(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. For vectors x and y in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let $< x, y >_{D_\mu} := x^T D_\mu y$, and $\|x\|_{D_\mu} := x^T D_\mu x$ denote the D_μ weighted inner product and the induced norm, respectively. For a vector $\mathbf{v}$, let $\mathbf{v^T}$ denote its transpose, and let $\|\mathbf{v}\|_\mathbf{2}$ and $\|\mathbf{v}\|_\infty$ denote the ℓ_2-norm and ℓ_∞-norm, respectively. These norms are by $\|\mathbf{v}\|_2 = \sqrt{\sum_i \mathbf{v}^2(i)}$, and $\|\mathbf{v}\|_\infty = \max_i |\mathbf{v}(i)|$. Additionally, let $\mathbf{e}$ denote the vector of all 1s in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

2.2. Average Reward Model

Average Reward. The long-term per-step expected reward accumulated by a stationary policy μ starting from state s is given by

$$\lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} r(S_t, A_t) | S_0 = s\right]. \tag{2}$$

Here, the expectation is with respect to both the transition dynamics of the environment, as well as the randomness in the policy μ. It is well known that under Assumption 2.1, this per-step expected reward in (2) is a constant that is independent of the starting state (Bertsekas, 2012). We denote this quantity by J_μ. It can be shown that

$$J_\mu = \mathbb{E}_{d_\mu}[r(S,A)] = \sum_{x\in\mathcal{S}, a\in\mathcal{A}} \pi_\mu(x)\mu(a|x)r(x,a). \quad (3)$$

Here, $\mathbb{E}_{d_\mu}[\cdot]$ denotes the expectation when (S,A) is sampled according to the distribution d_μ, which is defined after (1).

Q Function. Since $J_\mu = \mathbb{E}_{d_\mu}[r(S,A)]$, there exists a solution of the following Poisson equation or Bellman equation (see, Douc et al. (2018, Section 21.2)) $Q_\mu : \mathcal{X} \to \mathbb{R}$ such that for each $(s,a) \in \mathcal{X}$, and for P_2 given in (1),

$$r(s,a) - J_\mu = Q_\mu(s,a) - \sum_{(s',a')\in\mathcal{X}} P_2((s,a)),(s',a'))Q_\mu(s',a'). \quad (4)$$

Let Q_μ be a solution to (4). Then, $Q_\mu + c\mathbf{e}$ for any constant $c \in \mathbb{R}$ is also a solution. Let Q^*_μ denote the solution normalized so that it is orthogonal to $\mathbf{e}$, i.e., $Q^{*T}_\mu \mathbf{e} = 0$. It is well known that the set of solutions of (4) takes the following form (Puterman, 1994):

$$S_\mu := \{Q^*_\mu + c\mathbf{e} | c \in \mathbb{R}\}.$$

3. Asymptotic Variance

Similar to the (per-step) average reward of the stationary policy defined in (2), a natural notion of the (per-step) asymptotic variance of the rewards accumulated by the policy μ starting from state s is

$$\lim_{T\to\infty} \frac{1}{T} \mathrm{Var}\left[\sum_{t=0}^{T-1} r(S_t,A_t) \middle| S_0 = s\right]. \quad (5)$$

As earlier, this asymptotic variance is also computed with respect to randomness of both the dynamics of the MDP, and the policy μ.

3.1. Equivalent Expressions

In this work, we consider the problem of estimating the asymptotic variance of a given policy μ. To this end, we first present more tractable formulations for it, before presenting the proposed algorithm and its analysis.

In the proposition below, we show that like the per-step reward, the asymptotic variance is also a constant independent of the starting state. This follows from geometric mixing of $\mathcal{M}_2$ (Remark 2.2). Further, we also present equivalent formulations for the asymptotic variance defined in (5).

Proposition 3.1. *Given a stationary policy satisfying Assumption 2.1, asymptotic variance defined in (5) is a constant independent of the starting state, and is given by*

$$\kappa_\mu = \lim_{T\to\infty} \frac{1}{T} \mathrm{Var}\left[\sum_{t=0}^{T-1} r(S_t,A_t) \middle| (S_0,A_0) \sim d_\mu\right].$$

Furthermore, given any solution $Q_\mu \in S_\mu$ to (4),

$$\begin{aligned}
\kappa_\mu &= \mathbb{E}_{d_\mu}\big[(r(S,A) - J_\mu)^2\big] \quad (6)\\
&\quad + 2\lim_{T\to\infty}\sum_{j=1}^{T-1} \mathbb{E}_{d_\mu}[(r(S_0,A_0) - J_\mu)(r(S_j,A_j) - J_\mu)]\\
&= 2\mathbb{E}_{d_\mu}[(r(S,A) - J_\mu)Q_\mu(S,A)]\\
&\quad - \mathbb{E}_{d_\mu}\big[(r(S,A) - J_\mu)^2\big] \quad (7)\\
&= \mathbb{E}_{d_\mu}\big[Q^2_\mu(S,A)\big] - \mathbb{E}_{d_\mu}\big[(P_2Q_\mu)^2(S,A)\big]. \quad (8)
\end{aligned}$$

It is worth noting that the first term in equation (6) accounts for the per-step variance, while the second term encompasses the temporal correlation introduced by the Markov chain structure. These different representations for asymptotic variance are well known in literature. For example, see Douc et al. (2018, Theorem 21.2.6) for a similar formulation. For completeness, we give a proof of the proposition in Section A.1.

Continuing, the second representation in equation (7) formulates the variance of $r(\cdot,\cdot)$ in terms of the corresponding solution of the Poisson equation (4). This follows from Lemma A.1, which formulates the asymptotic variance of functions defined on the states of a Markov chain in terms of the solutions to the corresponding Poisson equation, and by noticing that $r(\cdot,\cdot)$ is the function of states $\mathcal{X}$ of the Markov chain $\mathcal{M}_2$ under consideration, and Q_μ is the corresponding solution to the Poisson equation (also see Douc et al. (2018, Theorem 21.2.5)). Note that this representation isn't affected by choice of Q_μ (any constant shift in Q_μ doesn't affect κ_μ). We will use the form in (7) to design a TD-type algorithm to estimate κ_μ for a given stationary policy μ.

Finally, the third equality follows by using the Poisson equation to replace $(r(S,A) - J_\mu)$ terms in (7) by $Q_\mu(S,A) - P_2Q_\mu(S,A)$.

3.2. Motivation for Asymptotic Variance

As discussed in Section 1, different notions of variance have been considered in the literature. We now motivate the choice of asymptotic variance in a risk-sensitive setting.

First, in the average reward setting, a classical goal is to optimize the long-term expected cumulative reward, which corresponds to expectation of the random variable $\sum_t r(S_t,A_t)$. The asymptotic variance considered in this work, corresponds to the variance of the same random variable. Since both the mean and variance of $\sum_t r(S_t,A_t)$

summed up to T terms are $O(T)$, we equivalently consider $\frac{1}{T}$ scaling of both the mean and variance to arrive at J_μ and κ_μ, respectively.

Second, since κ_μ corresponds to the variance in the central limit theorem for $\sum_t r(S_t, A_t)$ (see, Douc et al. (2018, Theorem 21.2.5)), we use the nomenclature "asymptotic variance".

Finally, consider an investor receiving return $r(S_t, A_t)$ on an investment at each time t. The total return of the investor in T steps is $\sum_{t=1}^{T} r(S_t, A_t)$. A risk-averse investor would aim to maximize the average cumulative return subject to its variance being small. Now, if the sequence $\{r(S_t, A_t)\}_{t\geq 1}$ were i.i.d. according to d_μ, the variance constraint would reduce to a bound on $T\mathbb{E}_{d_\mu}\big[(r(S,A) - J_\mu)^2\big]$. This corresponds to the first term in (6). However, in in a Markov sequence, this term alone does not capture the covariance across time. This is captured by the second term in (6).

It is important to emphasize that prior research on variance-constrained average reward RL, cf. (Prashanth & Ghavamzadeh, 2016), treats the first term in (6) as a surrogate for variance. However, in a Markovian context, the correlation between time steps is non-negligible, and should not be overlooked.

In what follows, we use linear SA for estimating κ_μ. We develop a TD-like linear SA algorithm for estimating the asymptotic variance associated with the Markov chain induced by the given stationary policy μ, i.e., κ_μ. In contrast to prior literature on TD such as (Devraj & Meyn, 2017; Yin & Wang, 2020), which focus on studying the variance of TD learning and its variants, we do *not* study the variance of a linear SA algorithm itself. We use linear SA simply as a tool for solving the variance estimation problem.

4. Variance Estimation: Tabular Setting

With the expressions for asymptotic variance derived in Proposition 3.1, in this section, we design a TD type algorithm for estimating κ_μ for a given stationary policy μ. While κ_μ is non-linear function of J_μ and Q_μ which need to be estimated from the samples, the proposed Algorithm 1 is a linear SA update. We simulate two independent trajectories under μ, which enables us to eliminate correlations at a minor expense of doubling the total number of samples required to ensure a given estimation error.

The following lemma expresses κ_μ in terms of averages computed from two independent trajectories evolving according to the specified policy μ.

Lemma 4.1. *Given a stationary policy μ satisfying Assumption 2.1, let (S, A), (S', A') be iid samples from d_μ. Then,*

$$\kappa_\mu = 2\mathbb{E}_{d_\mu}[(r(S,A) - r(S',A'))Q_\mu(S,A)] - \frac{1}{2}\mathbb{E}_{d_\mu}\big[(r(S,A) - r(S',A'))^2\big]\,.$$

The above lemma immediately follows from (7) and the observation that for a random variable X, $\mathrm{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$ can equivalently be expressed as

$$2\,\mathrm{Var}[X] = \mathbb{E}\big[(X - X')^2\big]\,,$$

where X' is an independent copy of X, and a similar expression for covariance. We refer the reader to Appendix B.1 for a proof.

4.1. Algorithm Design

For $T \geq 1$, and time steps $k = 1, 2, \ldots, T$, let (S_k, A_k) and (S'_k, A'_k) represent the states and actions chosen according to $\mu(\cdot|S_k)$ and $\mu(\cdot|S'_k)$ along the two independently simulated trajectories. To estimate κ_μ, we use SA for the expression in Lemma 4.1. This corresponds to (14) in Algorithm 1. To this end, we need to estimate Q_μ, which necessitates estimating J_μ along the same sample path. Estimating J_μ corresponds to the mean estimation update in (9). For estimating Q_μ, we again use SA to find a fixed point of (4). δ_k in (10) represents the SA adjustment term for Q_μ estimation. However, since the fixed points for (4) may not be unique, we particularly consider the one orthogonal to $\mathbf{e}$ in ℓ_2-norm. Thus, at each step, we project the δ_k update to the orthogonal subspace. This corresponds to the adjustments made in (11) and (12).

4.2. Convergence Rates

In this section, we bound the estimation error of the proposed algorithm. Let $Y_k := (S_k, A_k, S_{k+1}, A_{k+1}, S'_k, A'_k)$. Observe that $\mathcal{M}_3 := \{Y_k\}_{k\geq 1}$ is a Markov chain. Let its state space be denoted by $\mathcal{Y}$, which is a finite set. As earlier, under Assumption 2.1, $\mathcal{M}_3$ has a unique stationary distribution. Let us denote it by $\tilde{\pi}_\mu$. Further, $\mathcal{M}_3$ mixes geometrically fast (Bhatnagar & Lakshmanan, 2016, Proposition 1). This guarantees that there exist constants $C > 1$ and $\rho \in (0, 1)$ such that,

$$\sup_{y\in\mathcal{Y}}\; d_{TV}(\mathbb{P}(Y_k|Y_0 = y), \tilde{\pi}_\mu) \leq C\rho^k, \text{ for all } k \geq 1\,,$$

where $\mathbb{P}(Y_k|Y_0 = y)$ denotes the probability of $\mathcal{M}_3$ being in state Y_k at time k, starting from state y, and for probability measures P and Q, $d_{TV}(P, Q)$ denotes the total variation distance between P and Q.

Definition 4.2. Given $\delta > 0$, the mixing time $\tau(\delta)$ of the Markov chain $\mathcal{M}_3$ with precision δ is defined to be $\tau(\delta) := \min\{t \geq 0 \mid \sup_{y\in\mathcal{Y}}\; d_{TV}(\mathbb{P}(Y_t|Y_0 = y), \tilde{\pi}_\mu) \leq \delta\}$.

Algorithm 1: Policy Evaluation: Tabular Setting

Input: Time horizon $T > 0$, constants $c_1 > 0$, $c_2 > 0$, and step-size sequence $\{\alpha_k\}$.
Initialization: $J_0 = 0$, $Q_0 = \vec{\mathbf{0}}$, and $\kappa_0 = 0$.
for $k \leftarrow 1$ **to** T **do**

On first trajectory, take actions $A_k \sim \mu(\cdot|S_k)$ and $A_{k+1} \sim \mu(\cdot|S_{k+1})$ to observe $(S_k, A_k, r(S_k, A_k), S_{k+1}, A_{k+1})$.

On second trajectory, take action $A'_k \sim \mu(\cdot|S'_k)$ and observe $(S'_k, A'_k, r(S'_k, A'_k))$.

// **Average reward estimation**

$$J_{k+1} = J_k + c_1\alpha_k(r(S_k, A_k) - J_k). \quad (9)$$

// Q**-value estimation**

Define δ_k as

$$\delta_k := r(S_k, A_k) - J_k + (Q_k(S_{k+1}, A_{k+1}) - Q_k(S_k, A_k)). \quad (10)$$

For $(s, a) \neq (S_k, A_k)$,

$$Q_{k+1}(s, a) = Q_k(s, a) - \alpha_k\delta_k/|\mathcal{S}|\,|\mathcal{A}|. \quad (11)$$

For $(s, a) = (S_k, A_k)$,

$$Q_{k+1}(s, a) = Q_k(s, a) + \alpha_k\delta_k\,(1 - 1/|\mathcal{S}|\,|\mathcal{A}|)\,. \quad (12)$$

// **Variance estimation**

$$T_1 := 2(r(S_k, A_k) - r(S'_k, A'_k))Q_k(S_k, A_k), \quad (13)$$

$$T_2 := \frac{1}{2}\,(r(S_k, A_k) - r(S'_k, A'_k))^2\,,$$

$$\kappa_{k+1} = \kappa_k + c_2\alpha_k\,(T_1 - T_2 - \kappa_k)\,. \quad (14)$$

end
Output : Variance estimate κ_T

From geometric mixing, we have for any $\delta > 0$ that

$$\tau(\delta) \leq \frac{\log C}{\log(1/\rho)} \log\frac{1}{\delta} + 1 \leq L \log\frac{1}{\delta}, \quad (15)$$

where $L := 1 + \frac{\log C}{\log(1/\rho)}$. Let $\mathcal{R} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the vector of rewards and $r_{\max} := \|\mathcal{R}\|_\infty$. Recall that Q^*_μ was defined after (4). Next, define Δ_1 to be

$$\min\Big\{\mathbf{v}^T D_\mu(I - P_2)\mathbf{v} \mid \mathbf{v} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \|\mathbf{v}\|_2 = 1, \mathbf{v}^T\mathbf{e} = 0\Big\}\,.$$

Clearly, feasible vectors $\mathbf{v}$ in Δ_1 are non-constant vectors. Then, from Tsitsiklis & Van Roy (1997, Lemma 7), we have that $\mathbf{v}^T D_\mu(I - P_2)\mathbf{v} > 0$. Since the feasible region in Δ_1 is non-empty and compact, we get $\Delta_1 > 0$.

Next, let $\tilde{\Delta}_1 > 0$ be defined as $\tilde{\Delta}_1 := \Delta_1/(1 + 32r^2_{\max})$.

Finally, for the step-size constants c_1 and c_2 (inputs to Algorithm 1), define

$$\eta := \max\{(c_1^2 + 5 + 16c_2^2 r^2_{\max} + c_2^2)^{\frac{1}{2}}, r_{\max}(c_1^2 + 1 + 2c_2^2)^{\frac{1}{2}}\}. \quad (16)$$

The following theorem bounds the estimation error of Algorithm 1, which immediately implies $\tilde{O}(1/k)$ convergence rate for the mean-squared estimation error of κ_μ to 0.

Theorem 4.3. *Consider Algorithm 1 with c_1 and c_2 satisfying: $c_1 \geq \frac{1}{2}\left(\frac{1}{\Delta_1} + \Delta_1\right)$, and*

$$c_2 \in \left[2\tilde{\Delta}_1\left(1 - \frac{1}{\sqrt{2}}\right), 2\tilde{\Delta}_1\left(1 + \frac{1}{\sqrt{2}}\right)\right].$$

Let

$$\xi_1 := \left(1 + (J_\mu^2 + \kappa_\mu^2 + \|Q^*_\mu\|_2^2)^{\frac{1}{2}}\right)^2$$

and

$$\xi_2 := 520\eta^2\left(1 + (J_\mu^2 + \kappa_\mu^2 + \|Q^*_\mu\|_2^2)^{\frac{1}{2}}\right)^2,$$

where η is defined in (16).

(a) Let $\alpha_i = \alpha$ for all i, such that $\tilde{\Delta}\alpha < 2$, and

$$\alpha\tau(\alpha) \leq \min\left\{\frac{1}{4\eta}, \frac{\tilde{\Delta}_1}{260\eta^2}\right\}.$$

Then, for all $k \geq \tau(\alpha)$,

$$\mathbb{E}\big[(\kappa_k - \kappa_\mu)^2\big] \leq \xi_1\left(1 - \frac{\alpha\tilde{\Delta}_1}{2}\right)^{k-\tau(\alpha)} + \frac{\xi_2\alpha\tau(\alpha)}{\tilde{\Delta}_1}.$$

(b) Let $\alpha_i = \frac{\alpha}{i+h}$ for all i, with α and h chosen so that $2 < \alpha\tilde{\Delta}_1 < 2h$. Let k^* be the smallest positive integer such that $\sum_{i=0}^{k^*-1}\alpha_i \leq \frac{1}{2\eta}$, and for all $k \geq k^*$,

$$\sum_{i=k-\tau(\alpha_k)}^{k-1}\alpha_i \leq \min\left\{\frac{1}{4\eta}, \frac{\tilde{\Delta}_1}{260\eta^2}\right\}.$$

Then, for all $k \geq k^*$, and L defined after (15),

$$\mathbb{E}\big[(\kappa_k - \kappa_\mu)^2\big] \leq \xi_1\left(\frac{k^* + h}{k + h}\right)^{\alpha\tilde{\Delta}_1/2} + \left(\frac{4\alpha^2\xi_2 eL}{\tilde{\Delta}_1\alpha - 2}\right)\frac{\ln(k + h) - \ln\alpha}{k + h + 1}.$$

The above theorem follows from a more general result presented in the next section. We refer the reader to Appendix B.2 for a proof.

Theorem 4.3(a) bounds the mean-squared estimation error of the proposed algorithm in case of constant step size. Though the first term in this bound decays exponentially fast as the number of iterations k increase, the second term is constant and becomes a bottleneck after sufficiently large k. This is a well-known behavior of SA with constant step size. In fact, it suffers similar drawback of being non-adaptive as the vanilla Monte Carlo (Flegal & Jones, 2010; Chien et al., 1997). Specifically, in order to ensure the mean-squared error smaller than ϵ^2 for $\epsilon > 0$ (or mean absolute error smaller than ϵ), one needs to pick the step size α as a function of ϵ (see Corollary 4.5). This drawback of being non-adaptive is overcome by choosing a diminishing step size (as in Theorem 4.3(b)).

Second, it has been shown recently that linear SA with constant step size, in presence of Markov noise, suffers from an asymptotic bias (as $k \to \infty$), see Huo et al. (2023); Nagaraj et al. (2020). This is in contrast to the iid noise setting, where the asymptotic bias is shown to be zero (Mou et al., 2020; Lakshminarayanan & Szepesvari, 2018). This motivates us to study SA with a diminishing step size.

Remark 4.4. For diminishing step sizes, Algorithm 1 achieves $\tilde{O}(1/k)$ rate of convergence (Theorem 4.3(b)). As earlier, the first term in the bound decays faster, and the second term is the rate-determining $\tilde{O}(1/k)$ term. We believe that the rate of convergence of $\Omega(1/k)$ for MSE is tight. In the special setting of iid noise with parametric distributions, this follows from the Cramér-Rao lower bound. We discuss a simple estimator that achieves this rate in iid setting in Appendix E.

Using bounds in Theorem 4.3, the corollary below presents the number of iterations of Algorithm 1 needed to have the mean estimation error bounded by ϵ. It follows from setting the mean-squared estimation error bounds in Theorem 4.3 to at most ϵ^2. We refer the reader to Section B.3 for a proof of Corollary 4.5.

Corollary 4.5. *To estimate κ_μ using iterates κ_k generated by Algorithm 1 up to mean estimation error $\mathbb{E}[|\kappa_k - \kappa_\mu|] \leq \epsilon$, we require*

$$k = O\left(\frac{\log^2 \frac{1}{\epsilon}}{\epsilon^2}\right) \tilde{O}\left(\frac{L\|Q^*_\mu\|_2^2}{\Delta_1^4}\right).$$

The sample complexity in Corollary 4.5 depends on $|S||A|$ via Δ_1, L, and $\|Q^*_\mu\|_2^2$. Here, $Q^*_\mu \in \mathbb{R}^{|S||A|}$, and hence, $\|Q^*_\mu\|_2^2 = O(|S|^2|A|^2)$. The dependence of L and Δ_1 on $|S||A|$ is more implicit, and relates to the mixing properties of the Markov chain induced by the policy. Below we make this dependence explicit in two specific examples.

If the underlying Markov chain is a random walk on a complete graph (with $|S||A|$ vertices), then $\Delta_1 = O(1)$, $L = O(1)$, and hence we get $\tilde{O}(|S|^2|A|^2)$ dependence. On the other hand, if the underlying Markov chain is a random walk on a cycle graph (Levin & Peres, 2017, Section 12.3.1), then $\Delta_1 = O(1/|S|^2|A|^2)$, $L = O(|S|^2|A|^2)$, and hence, we get a dependence of $\tilde{O}(|S|^{12}|A|^{12})$ in the sample complexity.

Remark 4.6. Using Jensen's inequality, we have

$$\mathbb{E}[|\kappa_k - \kappa_\mu|] \leq \mathbb{E}^{\frac{1}{2}}\left[(\kappa_k - \kappa_\mu)^2\right],$$

which then gives a bound on the mean estimation error. In particular, for the diminishing step-sizes of the form $\alpha_i = \frac{\alpha}{i+h}$ for all $i \geq 1$, Theorem 4.3(b) gives $\mathbb{E}[|\kappa_k - \kappa_\mu|] \leq \tilde{O}(1/\sqrt{k})$.

Remark 4.7. Consider the problem of estimating the standard deviation, $\sqrt{\kappa_\mu}$. Mean-squared error for the estimator $\sqrt{\kappa_k}$ satisfies $\mathbb{E}\left[(\sqrt{\kappa_k} - \sqrt{\kappa_\mu})^2\right] \leq \frac{1}{\kappa_\mu}\mathbb{E}\left[(\kappa_k - \kappa_\mu)^2\right]$, which is at most $\tilde{O}(1/k)$ for $\alpha_k = O(1/k)$.

5. Variance Estimation: Linear Function Approximation

When the underlying state and action spaces are large, estimating the Q_μ function for each state-action pair requires a lot of memory, and may be intractable. To address this, we consider evaluating an approximation of Q_μ that is its projection onto a linear subspace spanned by a given fixed set of d vectors $\{\tilde{\phi}_1, \ldots, \tilde{\phi}_d\}$, where $\tilde{\phi}_i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ for $i \in [d]$. In particular, for $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\theta \in \mathbb{R}^d$, we consider a linear function approximation $Q_\theta(s,a) = \phi(s,a)^T\theta$ of $Q_\mu(s,a)$, where $\phi^T(s,a) := [\tilde{\phi}_1(s,a) \ldots \tilde{\phi}_d(s,a)]$ is the feature vector for state-action pair (s,a) and $\phi(s,a) \in \mathbb{R}^d$. With this notation, let Φ be a $|\mathcal{S}|\,|\mathcal{A}| \times d$ matrix with $\tilde{\phi}_i$ being the i^{th} column, and let $W_\Phi = \{\Phi\theta : \theta \in \mathbb{R}^d\}$ denote the column space of Φ. Then, $Q_\theta = \Phi\theta$, where $Q_\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times 1}$ is an approximation for Q_μ using θ.

Assumption 5.1. *The matrix Φ is full rank, i.e., the set of feature vectors $\{\tilde{\phi}_1, \ldots, \tilde{\phi}_d\}$ are linearly independent. Additionally, $\|\phi(s,a)\|_2 \leq 1$, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

This is a standard assumption in literature, and can be achieved by feature normalization, see (Tsitsiklis & Van Roy, 1999; Bertsekas & Tsitsiklis, 1996).

We now introduce some notation that will be used in this section. Define subspace $S_{\Phi,e}$ of $\mathbb{R}^d$ as $S_{\Phi,e} := \text{span}\,(\{\theta|\Phi\theta = \mathbf{e}\})$. It equals $\{c\theta_e|c \in \mathbb{R}\}$ if $\mathbf{e} \in W_\Phi$, and θ_e is such that $\Phi\theta_e = \mathbf{e}$. Otherwise, $S_{\Phi,e} = \{0\}$. Let E be the subspace of $\mathbb{R}^d$ that is orthogonal complement (in ℓ_2-norm) of $S_{\Phi,e}$, i.e., $E = \{\theta \in \mathbb{R}^d : \theta^T\theta_e = 0\}$ if $\mathbf{e} \in W_\Phi$. It equals $\mathbb{R}^d$, otherwise. Additionally, let $\Pi_{2,E}$ denote the orthogonal projection of vectors in $\mathbb{R}^d$ (in 2-norm) on the subspace E.

Observe that for $\theta \in E$, $\Phi\theta$ is a non-constant vector. This follows since there does not exist $\theta \in E$ such that $\Phi\theta = \mathbf{e}$.

In particular, if $\mathbf{e} \in W_\Phi$, then the unique vector θ_e such that $\Phi\theta_e = \mathbf{e}$ does not belong to E. Similarly, if $\mathbf{e} \notin W_\Phi$, then there does not exist a vector $\theta \in \mathbb{R}^d$, hence in E, such that $\Phi\theta = \mathbf{e}$.

5.1. Algorithm

We now propose an algorithm for estimating the asymptotic variance of a given stationary policy μ with linear function approximation. Here, we estimate a good θ at each step. Call this estimate θ_k. The corresponding estimate for Q_μ is then $\Phi\theta_k$. The algorithm in this setting is a modification of Algorithm 1, with (10), (11) and (12), and (13) replaced by (17), (18), and (19), respectively.

$$\delta_k := r(S_k, A_k) - J_k + (\phi(S_{k+1}, A_{k+1}) - \phi(S_k, A_k))^T\theta_k, \quad (17)$$

$$\theta_{k+1} = \theta_k + \alpha_k \Pi_{2,E}\phi(S_k, A_k)\delta_k, \quad (18)$$

$$T_1 := 2(r(S_k, A_k) - r(S'_k, A'_k))\phi^T(S_k, A_k)\theta_k. \quad (19)$$

Here, updates in (17) and (18) correspond to TD for estimating Q_μ with linear function approximation, adjusted with projection on subspace E. Note that Algorithm 1 for the tabular setting is a special case with $d = |\mathcal{S}|\,|\mathcal{A}|$ and $\Phi = I$, the identity matrix in $|\mathcal{S}|\,|\mathcal{A}|$ dimensions.

5.2. Convergence Rates

In this section, we present a finite-sample bound on the estimation error of the proposed algorithm with linear function approximation. Recall the Markov chain $\mathcal{M}_3$ introduced in Section 4.2, along with its mixing time $\tau(\cdot)$. For E defined below Assumption 5.1, define

$$\Delta_2 := \min_{\|\theta\|_2=1, \theta\in E} \theta^T\Phi^T D_\mu(I - P_2)\Phi\theta.$$

As before, this can be shown to be strictly positive since for $\theta \in E$, $\Phi\theta$ is a non-constant vector. Let $\tilde{\Delta}_2 > 0$ be defined as $\tilde{\Delta}_2 := \Delta_2/(1 + 32r_{\max}^2)$. Finally, let $\Theta^{*T} := [J_\mu\ \theta^{*T}\ \kappa^*]$, where

$$\kappa^* = 2\mathbb{E}_{d_\mu}[(r(S, A) - r(S', A'))\,[\Phi\theta^*]\,(S, A)] - \frac{1}{2}\mathbb{E}_{d_\mu}\left[(r(S, A) - r(S', A'))^2\right], \quad (20)$$

and θ^* is the unique vector in E that is also a solution for

$$\Phi\theta = \Pi_{D_\mu, W_\Phi} T_\mu \Phi\theta,$$

where Π_{D_μ,W_Φ} is the projection matrix onto $W_\Phi := \left\{\Phi\theta | \theta \in \mathbb{R}^d\right\}$ with respect to D_μ norm. Specifically, $\Pi_{D_\mu,W_\Phi} = \Phi(\Phi^T D_\mu\Phi)^{-1}\Phi^T D_\mu$. Further, T_μ is an operator that for a vector $V \in \mathbb{R}^{|S||A|}$, satisfies $T_\mu V = \mathcal{R} - J_\mu\mathbf{e} + P_2 V$. Observe that κ^* differs from κ_μ in that Q_μ in the formulation in Lemma 4.1 is replaced by its estimate in the subspace spanned by Φ.

The following theorem bounds the mean-squared distance between the estimate at time k and the limit point κ^* under constant as well as diminishing step sizes.

Theorem 5.2. *Consider estimates κ_k generated by the algorithm with c_1 and c_2 satisfying: $c_1 \geq \frac{1}{2}\left(\frac{1}{\tilde{\Delta}_2} + \Delta_2\right)$, and*

$$c_2 \in \left[2\tilde{\Delta}_2\left(1 - \frac{1}{\sqrt{2}}\right), 2\tilde{\Delta}_2\left(1 + \frac{1}{\sqrt{2}}\right)\right].$$

Let $\xi_1 = (1 + \|\Theta^*\|_2)^2$ and $\xi_2 = 520\eta^2\,(\|\Theta^*\|_2 + 1)^2$, where η is defined in (16), and Θ^* before (20).

(a) Let $\alpha_i = \alpha$ for all i, such that $\tilde{\Delta}_2\alpha < 2$, and

$$\alpha\tau(\alpha) \leq \min\left\{\frac{1}{4\eta}, \frac{\tilde{\Delta}_2}{260\eta^2}\right\}.$$

Then, for all $k \geq \tau(\alpha)$,

$$\mathbb{E}\left[(\kappa_k - \kappa^*)^2\right] \leq \xi_1\left(1 - \frac{\alpha\tilde{\Delta}_2}{2}\right)^{k-\tau(\alpha)} + \frac{\xi_2\alpha\tau(\alpha)}{\tilde{\Delta}_2}.$$

(b) Let $\alpha_i = \frac{\alpha}{i+h}$ for all i, with α and h chosen so that $2 < \alpha\tilde{\Delta}_2 < 2h$. Let k^* be the smallest positive integer such that $\sum_{i=0}^{k^*-1}\alpha_i \leq \frac{1}{2\eta}$, and for all $k \geq k^*$,

$$\sum_{i=k-\tau(\alpha_k)}^{k-1}\alpha_i \leq \min\left\{\frac{1}{4\eta}, \frac{\tilde{\Delta}_2}{260\eta^2}\right\}.$$

Then, for all $k \geq k^*$, and L defined after (15),

$$\mathbb{E}\left[(\kappa_k - \kappa^*)^2\right] \leq \xi_1\left(\frac{k^* + h}{k + h}\right)^{\alpha\tilde{\Delta}_2/2} + \left(\frac{4\alpha^2\xi_2 eL}{\tilde{\Delta}_2\alpha - 2}\right)\frac{\ln(k + h) - \ln\alpha}{k + h + 1}.$$

To prove Theorem 5.2, we view the proposed algorithm as a linear SA update, and prove appropriate contraction properties for the associated matrices. Unlike in the discounted setting of RL, the operators in the average reward RL setting are not contractive under any norm. In fact, as in (Zhang et al., 2021), we establish semi-norm contraction, and establish convergence in an appropriate subspace (orthogonal to all 1s vector). This turns out to be sufficient since the formulation of asymptotic variance in (7) is unaffected by constant shifts in estimation of Q_μ.

Finally, we prove a much stronger statement where the above bounds hold for mean squared difference between the iterates of the algorithm $[J_k\ \theta_k^T\ \kappa_k]$ and Θ^* (see Theorem C.1). We refer the reader to Section 5.4 for a sketch of the proof, and to Section C.4 for a complete proof the the theorem.

5.3. Approximation Error

Since we approximate elements of the set S_μ of Q_μ functions (defined after (4)) by a linear combination of basis vectors, we incur an approximation error in estimating κ_μ using (7). In particular, κ^* may differ from κ_μ, depending on the approximation architecture. In this section, we bound the squared error: $(\kappa_\mu - \kappa^*)^2$.

To this end, we first define the approximation error associated with Q_μ approximation. Since each element of S_μ is a valid Q_μ function, for any $\theta \in \mathbb{R}^d$, we define the error of approximation by $\Phi\theta$ as the minimum weighted distance (in D_μ-norm) of $\Phi\theta$ from points in S_μ (Tsitsiklis & Van Roy (1999)), i.e., $\inf_{Q\in S_\mu} \|\Phi\theta - Q\|_{D_\mu}$. The minimum possible approximation error due to the chosen architecture is given by

$$\mathcal{E} := \inf_{\theta\in\mathbb{R}^d} \inf_{Q\in S_\mu} \|\Phi\theta - Q\|_{D_\mu}.$$

This essentially captures the distance between the two sets, S_μ and the column space of Φ.

Proposition 5.3. *Given d basis vectors represented as columns of Φ, there exists a constant $\lambda \in (0,1)$ such that*

$$(\kappa^* - \kappa_\mu)^2 \leq \frac{16 r_{\max}^2}{1-\lambda^2}\, \mathcal{E}^2. \tag{21}$$

Notice that $\mathcal{E}$, and hence the RHS above, equals 0 if the chosen basis functions are such that the span of these intersects with S_μ. This is particularly true for the tabular setting, where $d = |\mathcal{S}|\,|\mathcal{A}|$ and the basis vectors are the standard basis in this dimension.

To prove Proposition 5.3, we first bound the approximation error in estimation of Q_μ, similar to that for value function in (Tsitsiklis & Van Roy, 1999). Using this bound in (7), we arrive at the bound in (21). We refer the reader to Section C.5 for a proof of the proposition.

5.4. Proof Sketch for Theorem 5.2

To prove the bounds in Theorem 5.2, we view the proposed algorithm as a linear SA update, and use an appropriate Lyapunov drift argument. To this end, we first present the corresponding linear SA algorithm.

For $(s,a) \in \mathcal{S}\times\mathcal{A}$, $k \in \mathbb{N}_+$, define

$$r_k = r(S_k, A_k), r'_k = r(S'_k, A'_k), \phi_k = \phi(S_k, A_k). \tag{22}$$

Let $\vec{\mathbf{0}}$ be the 0 vector in $\mathbb{R}^d$. Define $A(Y_k) \in \mathbb{R}^{d+2} \times \mathbb{R}^{d+2}$,

$$A(Y_k) := \begin{bmatrix} -c_1 & \vec{\mathbf{0}}^T & 0 \\ -\Pi_{2,E}\phi_k & \Pi_{2,E}\phi_k(\phi_{k+1}^T - \phi_k^T) & \vec{\mathbf{0}} \\ 0 & 2c_2(r_k - r'_k)\phi_k^T & -c_2 \end{bmatrix},$$

and $b(Y_k) \in \mathbb{R}^{d+2}$ as

$$b(Y_k)^T := \begin{bmatrix} c_1 r_k & r_k\,[\Pi_{2,E}\phi_k]^T & -\frac{c_2}{2}(r_k - r'_k)^2 \end{bmatrix}.$$

Let $A = \mathbb{E}_{d_\mu}[A(Y_k)]$ and $b = \mathbb{E}_{d_\mu}[b(Y_k)]$ denote the stationary averages of $A(Y_k)$ and $b(Y_k)$, respectively (see Appendix C.1 for exact form of A and b), and let $\Theta^T := [J\ \theta^T\ \kappa]$ with $\theta \in E$. Observe that $\Theta = \Theta^*$ (defined before (20)) is the unique solution to $A\Theta + b = 0$ with $\theta \in E$ (see Appendix C.2 for a detailed justification). Since the algorithm doesn't have access to matrices A and b, we use SA to solve for Θ^*, which corresponds to the following update rule at step $k+1$ with step size α_k:

$$\Theta_{k+1} = \Theta_k + \alpha_k\left(A(Y_k)\Theta_k + b(Y_k)\right). \tag{23}$$

In fact, the above equation coincides with the update for the proposed algorithm, with $\Theta_k^T := [J_k\ \theta_k^T\ \kappa_k]$.

The lemma below establishes that A is contractive when restricted to an appropriate subspace, for appropriate choices for the step-size constants c_1 and c_2. This result is crucial in establishing the convergence of the iterates in (23).

Lemma 5.4. *Under Assumption 2.1, and conditions on c_1 and c_2 from Theorem 5.2, the matrix A satisfies*

$$\min_{\Theta\in\mathbb{R}\times E\times\mathbb{R}, \|\Theta\|_2^2=1} -\Theta^T A\Theta > \tilde{\Delta}_2/2.$$

Finally, as in Zhang et al. (2021), we use a Lyapunov-drift argument along with Lemma 5.4 to arrive at the finite-time bounds in Theorem 5.2. We refer the reader to Section C.3 for a proof of Lemma 5.4, and to Section C.4 for a complete proof of Theorem 5.2.

6. Conclusions and Future Work

We proposed a TD-like algorithm to estimate the asymptotic variance of a given stationary policy, and developed the first finite sample bounds on the estimation error in the mean-squared sense. We established $\tilde{O}(1/k)$ rate of convergence of the proposed algorithm in both the tabular, as well as linear function approximation settings. We also characterized the approximation error in the latter setting. Notably, using sampling along two independent trajectories, the proposed algorithm can be viewed as a version of a linear SA algorithm. Using Lyapunov drift arguments, we arrived at finite sample guarantees.

As future research, it would be interesting to design an algorithm that uses a single trajectory instead of two, to estimate the asymptotic variance with good finite sample guarantees. Additionally, our policy-evaluation algorithm can serve as a building block for designing sample-efficient actor-critic algorithms for identifying a policy that maximizes the long-run average reward subject to an asymptotic variance constraint.

Impact Statement

This work focuses on theoretical results for reinforcement learning (RL). While RL algorithms have important societal implications (e.g. in autonomous driving, healthcare, RL algorithms for network control, etc.), and thus designing provably-efficient algorithms and understanding their performance is important, we believe that the direct ethical consequences of our work is somewhat limited.

Acknowledgements

This work was partially supported by NSF grants EPCN-2144316, CPS-2240982, DST grant ITS/2024/002721, and the International Centre for Theoretical Sciences (ICTS) for participating in the meeting — Data Science: Probabilistic and Optimization Methods (code:ICTS/dspom2023/7).

References

Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, E., Ghosh, M. K., and Marcus, S. I. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 31:282–344, 1993.

Arrow, K. J. *Essays in the Theory of Risk Bearing*. Markham, Chicago, IL, 1971.

Atchadé, Y. F. Kernel estimators of asymptotic variance for adaptive markov chain monte carlo. *The Annals of Statistics*, 39(2):990–1011, 2011. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/29783664.

Benton, A. R. *Asymptotic variance sensitive Markov decision processes*. PhD thesis, 2022.

Bertsekas, D. P. *Dynamic Programming and Optimal Control, Vol. II, 4th edition*. Athena Scientific, 2012.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Bhatnagar, S. and Lakshmanan, K. Multiscale q-learning with linear function approximation. *Discrete Event Dynamic Systems*, 26:477–509, 2016.

Bhatnagar, S., Sutton, R., Ghavamzadeh, M., and Lee, M. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Blackwell, D. Discrete dynamic programming. *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

Borkar, V. S. Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5, 2010.

Borkar, V. S. and Jain, R. Risk-constrained Markov decision processes. In *IEEE Conference on Decision and Control*, pp. 2664–2669, 2010.

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

Borkar, V. S. and Meyn, S. P. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

Brown, B. W. On the iterative method of dynamic programming on a finite space discrete time markov process. *The annals of mathematical statistics*, pp. 1279–1285, 1965.

Chen, S., Devraj, A., Busic, A., and Meyn, S. Explicit mean-square error bounds for monte-carlo and linear stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4173–4183. PMLR, 2020.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

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

Chien, C., Goldsman, D., and Melamed, B. Large-sample results for batch means. *Management Science*, 43(9): 1288–1295, 1997.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, pp. 3509–3517, 2014.

Devraj, A. M. and Meyn, S. P. Fastest convergence for q-learning. *arXiv preprint arXiv:1707.03770*, 2017.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. *Markov chains*. Springer, 2018.

Filar, J., Kallenberg, L., and Lee, H. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

Filar, J., Krass, D., and Ross, K. Percentile performance criteria for limiting average Markov decision processes. *IEEE Transaction of Automatic Control*, 40(1):2–10, 1995.

Flegal, J. M. and Jones, G. L. Batch means and spectral variance estimators in markov chain monte carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/25662268.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084. PMLR, 2021.

Howard, R. A. *Dynamic programming and markov processes.* John Wiley, 1960.

Howard, R. A. and Matheson, J. E. Risk-sensitive Markov decision processes. *Management Science*, 18:356–369, 1972.

Huo, D., Chen, Y., and Xie, Q. Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 81–82, 2023.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

Köse, U. and Ruszczyński, A. Risk-averse learning by temporal difference methods with markov risk measures. *Journal of machine learning research*, 22(38):1–34, 2021.

L.A., P. and Fu, M. C. Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends® in Machine Learning*, 15(5):537–693, 2022. ISSN 1935-8237.

La, P. and Ghavamzadeh, M. Actor-critic algorithms for risk-sensitive mdps. *Advances in neural information processing systems*, 26, 2013.

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355. PMLR, 2018.

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

Mandl, P. On the variance in controlled markov chains. *Kybernetika*, 7(1):1–12, 1971.

Mannor, S. and Tsitsiklis, J. N. Algorithmic aspects of mean–variance optimization in Markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Markowitz, J., Gardner, R. W., Llorens, A., Arora, R., and Wang, I. A risk-sensitive approach to policy optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15019–15027, Jun. 2023.

Mihatsch, O. and Neuneier, R. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2):267–290, 2002.

Moharrami, M., Murthy, Y., Roy, A., and Srikant, R. A policy gradient algorithm for the risk-sensitive exponential cost mdp, 2022.

Mou, S. and Maguluri, S. T. Heavy traffic queue length behaviour in a switch under markovian arrivals. *arXiv preprint arXiv:2006.06150*, 2020.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pp. 2947–2997. PMLR, 2020.

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

Nielsen, F. Cramér-rao lower bound and information geometry. *Connected at Infinity II: A Selection of Mathematics by Indians*, pp. 18–37, 2013.

Prashanth, L. and Ghavamzadeh, M. Variance-constrained actor-critic algorithms for discounted and average reward mdps. *Machine Learning*, 105:367–417, 2016.

Prashanth, L. A. Policy gradients for CVaR-constrained MDPs. In *Algorithmic Learning Theory (ALT)*, pp. 155–169, 2014.

Prashanth, L. A., Jie, C., Fu, M. C., Marcus, S. I., and Szepesvári, C. Cumulative prospect theory meets reinforcement learning: prediction and control. In *International Conference on Machine Learning*, pp. 1406–1415, 2016.

Puterman, M. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Robert, C. P. Convergence control methods for markov chain monte carlo algorithms. *Statistical Science*, 10(3): 231–253, 1995.

Ruszczyński, A. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125:235–261, 2010.

Sobel, M. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pp. 794–802, 1982.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd ed. edition, 2018.

Tamar, A., Castro, D. D., and Mannor, S. Policy gradients with variance related risk criteria. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, pp. 387–396, 2012.

Tamar, A., Di Castro, D., and Mannor, S. Temporal difference methods for the variance of the reward to go. In *International Conference on Machine Learning*, pp. 495–503, 2013.

Tamar, A., Chow, Y., Ghavamzadeh, M., and Mannor, S. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1468–1476, 2015.

Tamar, A., Di Castro, D., and Mannor, S. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.

Tsitsiklis, J. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Tsitsiklis, J. N. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16:185–202, 1994.

Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 1, pp. 498–502. IEEE, 1997.

Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

Veinott, A. F. On finding optimal policies in discrete dynamic programming with no discounting. *The Annals of Mathematical Statistics*, 37(5):1284–1294, 1966.

Whittle, P. *Risk-sensitive Optimal Control*. Wiley-Interscience series in systems and optimization. Wiley, 1990. ISBN 9780471926221.

Wu, W. B. Recursive estimation of time-average variance constants. *The Annals of Applied Probability*, 19 (4):1529–1552, 2009. ISSN 10505164. URL http://www.jstor.org/stable/30243631.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR, 2020.

Zhang, S., Zhang, Z., and Maguluri, S. T. Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34: 1230–1242, 2021.

A. Proofs for Results in Section 3

A.1. Proof of Proposition 3.1

Consider the induced Markov chain $\mathcal{M}_2$ with the corresponding probabilities of transition given by (1) and having a unique stationary distribution d_μ (Remark 2.2). Observe that $r(\cdot,\cdot)$ is a function defined on $\mathcal{X}$, states of the discrete time Markov chain $\mathcal{M}_2$, with stationary expectation J_μ. Further, we have from (4) that Q_μ is a solution for the corresponding Poisson Equation. Then, from Lemma A.4 we get that the asymptotic variance is constant independent of the starting state.

Further, from Lemma A.1, we get the two formulations for κ_μ, the asymptotic variance for function $f = r(\cdot,\cdot)$.

Finally, as discussed in the main text, the third formulation follows by replacing $(r(S,A) - J_\mu)$ terms in (7) by $Q_\mu(S,A) - P_2 Q_\mu(S,A)$, which follows from the Poisson Equation (4). □

A.2. Variance of Functions of a Discrete Time Markov Chain (DTMC)

Lemma A.1. *Let $X = (X_k : k \geq 0)$ be an irreducible and aperiodic DTMC on a finite-state space $\mathcal{S}$ with transition probability matrix P and a unique stationary distribution π. Let $f : \mathcal{S} \to \mathbb{R}$ be any function, and let $\bar{f} := \sum_{x\in\mathcal{S}} \pi(x) f(x)$ denote its stationary expectation. Let V denote the solution to the Poisson equation for $f - \bar{f}$, i.e., for each $x \in \mathcal{S}$, V satisfies*

$$V(x) - \sum\nolimits_{x'\in\mathcal{S}} P(x,x')V(x') = f(x) - \bar{f}.$$

Then, for $j \in \mathbb{N}$ and $\gamma(k) := \mathbb{E}_\pi\big[\big(f(X_k) - \bar{f}\big)\big(f(X_0) - \bar{f}\big)\big]$, the asymptotic variance of f is given by

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k) | X_0 \sim \pi_0\right] &= \gamma(0) + 2\lim_{n\to\infty}\sum_{k=1}^{n-1}\gamma(k) \\ &= 2\sum_{x\in\mathcal{S}} \pi(x)(f(x)-\bar{f})V(x) - \sum_{x\in\mathcal{S}}\pi(x)(f(x)-\bar{f})^2.\end{aligned}$$

Proof. From Lemma A.4, we have

$$\lim_{n\to\infty}\frac{1}{n}\operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k)|X_0\sim\pi_0\right] = \lim_{n\to\infty}\frac{1}{n}\operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k)|X_0\sim\pi\right].$$

We show that the r.h.s. in the above equation is same as the expression in the lemma. Recall that for any $k \geq 1$, $\mathbb{E}[f(X_k)|X_0 \sim \pi] = \bar{f}$. For simplicity of notation, we denote $\mathbb{E}_{X_0\sim\pi_0}[\cdot]$ by $\mathbb{E}_{\pi_0}[\cdot]$. Then,

$$\begin{aligned}\operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k)|X_0\sim\pi\right] &= \mathbb{E}_\pi\left[\left(\sum_{k=0}^{n-1}\big(f(X_k)-\bar{f}\big)\right)^2\right] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{n-1}(f(X_k)-\bar{f})^2\right] + 2\sum_{k=0}^{n-1}\sum_{i<k}\mathbb{E}_\pi\big[\big(f(X_i)-\bar{f}\big)\big(f(X_k)-\bar{f}\big)\big] \\ &\overset{(a)}{=} \mathbb{E}_\pi\left[\sum_{k=0}^{n-1}(f(X_k)-\bar{f})^2\right] + 2\sum_{k=1}^{n-1}(n-k)\mathbb{E}_\pi\big[\big(f(X_k)-\bar{f}\big)\big(f(X_0)-\bar{f}\big)\big] \\ &= n\gamma(0) + 2\sum_{k=1}^{n-1}(n-k)\gamma(k),\end{aligned}$$

where (a) follows from Markov property. Dividing by n and taking limits, we get

$$\lim_{n\to\infty}\frac{1}{n}\operatorname{Var}\left[\sum_{k=0}^{n-1}f(X_k)|X_0\sim\pi\right] = \gamma(0)+\lim_{n\to\infty}2\sum_{k=1}^{n-1}\frac{n-k}{n}\gamma(k)$$

$$\stackrel{(a)}{=} \gamma(0)+\lim_{n\to\infty}2\sum_{k=1}^{n-1}\gamma(k) \tag{24}$$

$$= \lim_{n\to\infty}2\sum_{k=0}^{n-1}\frac{n-k}{n}\gamma(k)-\gamma(0)$$

$$\stackrel{(b)}{=} \lim_{n\to\infty}2\sum_{k=0}^{n-1}\gamma(k)-\gamma(0), \tag{25}$$

where (a) and (b) follow from Lemma A.2. Expression in (24) corresponds to the first expression in the lemma.

Next, we show that the expression in (25) is same as the second expression in the lemma.

Since V is the solution of the Poisson Equation, we have the following form for V Douc et al. (2018, Proposition 21.2.3, Lemma 21.2.2)

$$V(x)=\sum_{k=0}^{\infty}\mathbb{E}\big[f(X_k)-\bar{f}|X_0=x\big]+c,$$

for any constant c. Substituting this in the expression in the lemma statement gives:

$$2\sum_{x\in\mathcal{S}}\pi(x)(f(x)-\bar{f})V(x)-\gamma(0) = 2\sum_{x\in\mathcal{S}}\pi(x)(f(x)-\bar{f})\left(\sum_{k=0}^{\infty}\mathbb{E}\big[f(X_k)-\bar{f}|X_0=x\big]\right)-\gamma(0)$$

$$\stackrel{(a)}{=} 2\sum_{k=0}^{\infty}\mathbb{E}_\pi\big[(f(X_0)-\bar{f})\left(f(X_k)-\bar{f}\right)\big]-\gamma(0)$$

$$= 2\lim_{n\to\infty}\sum_{k=1}^{n-1}\gamma(k)-\gamma(0),$$

which is same as (25). Here, to change the limits and expectation in (a), we used the bounded convergence theorem since f is bounded (underlying Markov chain is on a finite state space). □

Lemma A.2. *Let $X=(X_k:k\geq 0)$ be an irreducible and aperiodic DTMC on a finite state space $\mathcal{S}$ with transition probability matrix P and a unique stationary distribution π. Let $f:\mathcal{S}\to\mathbb{R}$ be any function, and let $\bar{f}:=\sum_{x\in\mathcal{S}}\pi(x)f(x)$ denote its stationary expectation. Then,*

$$\lim_{n\to\infty}\sum_{k=0}^{n-1}\frac{n-k}{n}\gamma(k)=\lim_{n\to\infty}\sum_{k=0}^{n-1}\gamma(k),$$

where

$$\gamma(k):=\mathbb{E}_\pi\big[(f(X_k)-\bar{f})(f(X_0)-\bar{f})\big].$$

Proof. The proof of this lemma follows along the lines of the proof of Mou & Maguluri (2020, Lemma 3). However, we give the proof for completeness. Let $f_{\max}:=\max_{s\in\mathcal{S}}f(x)$. Since X is an irreducible and aperiodic Markov chain on a finite state space, it mixes geometrically fast (Levin & Peres, 2017), i.e., there exist constants $C>0$ and $\alpha\in(0,1)$ such that for all $k\in\mathbb{N}_+$,

$$\sup_{x\in\mathcal{X}}d_{TV}(P(X_k|X_0=x),\pi)\leq C\alpha^k,$$

where $d_{TV}(\cdot,\cdot)$ represents the total variation distance between the two input distributions. Below, we first show that $\gamma(k)$ is bounded.

$$\begin{aligned}
|\gamma(k)| &= \left|\mathbb{E}_\pi\left[\left(f(X_k)-\bar{f}\right)\left(f(X_0)-\bar{f}\right)\right]\right| \\
&= \left|\mathbb{E}_\pi\left[\mathbb{E}\left[\left(f(X_k)-\bar{f}\right)\left(f(X_0)-\bar{f}\right)\right]|X_0=x\right]\right| \\
&= \left|\mathbb{E}_{x\sim\pi}\left[\mathbb{E}\left[\left(f(X_k)-\bar{f}\right)|X_0=x\right]\left(f(x)-\bar{f}\right)\right]\right| \\
&\leq \mathbb{E}_{x\sim\pi}\left[\left|\mathbb{E}\left[\left(f(X_k)-\bar{f}\right)|X_0=x\right]\right|\left|f(x)-\bar{f}\right|\right] \\
&\leq 4f_{\max}^2 C\alpha^k,
\end{aligned}$$

where we used Lemma A.3 to get the last inequality above.

Next, define

$$V_1 := \lim_{n\to\infty}\sum_{k=0}^{n-1}\gamma(k), \quad \text{and} \quad V_2 := \lim_{n\to\infty}\sum_{k=0}^{n-1}\frac{n-k}{n}\gamma(k).$$

Consider the following:

$$\begin{aligned}
|V_1 - V_2| &= \lim_{n\to\infty}\left|\sum_{k=0}^{n-1}\frac{k}{n}\gamma(k)\right| \\
&\leq \lim_{n\to\infty}\sum_{k=0}^{n-1}\frac{k}{n}|\gamma(k)| \\
&\overset{(a)}{\leq} 4f_{\max}^2 C\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n-1}k\alpha^k \\
&\leq 0,
\end{aligned}$$

since $\alpha < 1$. This implies $V_1 = V_2$. □

Lemma A.3 (Mou & Maguluri (2020, Lemma 2))**.** *Let $X = (X_k : k \geq 0)$ be DTMC on a finite-state space $\mathcal{S}$ with a unique stationary distribution π. Let $f : \mathcal{S} \to \mathbb{R}$ be any function, and let $\bar{f} := \sum_{x\in\mathcal{S}}\pi(x)f(x)$ denote its stationary expectation. Then, for any $k \in \mathbb{N}_+$, there exist constants $\alpha \in (0,1)$ and $C > 0$, such that for any initial distribution $X_0 \sim \pi_0$, we have*

$$\left|\mathbb{E}\left[f(X_k) - \bar{f}|X_0 \sim \pi_0\right]\right| \leq 2f_{\max}C\alpha^k,$$

where $f_{\max} := \max_{x\in\mathcal{S}} f(x)$.

Proof. Consider the following inequalities:

$$\begin{aligned}
\left|\mathbb{E}\left[f(X_k) - \bar{f}|X_0 \sim \pi_0\right]\right| &= \left|\sum_{x'\in\mathcal{S}}\pi_0(x')\sum_{x\in\mathcal{S}}f(x)\left(\mathbb{P}\left(X_k = x|X_0 = x'\right) - \pi(x)\right)\right| \\
&\leq \sum_{x'\in\mathcal{S}}\pi_0(x')\sum_{x\in\mathcal{S}}|f(x)|\left|\mathbb{P}\left(X_k = x|X_0 = x'\right) - \pi(x)\right| \\
&\leq f_{\max}\sum_{x'\in\mathcal{S}}\pi_0(x')\sum_{x\in\mathcal{S}}\left|\mathbb{P}\left(X_k = x|X_0 = x'\right) - \pi(x)\right| \\
&\leq f_{\max}\sup_{x'\in\mathcal{S}}\sum_{x\in\mathcal{S}}\left|\mathbb{P}\left(X_k = x|X_0 = x'\right) - \pi(x)\right| \\
&\leq 2f_{\max}C\alpha^k,
\end{aligned}$$

where the last inequality follows from the definition of total variation distance and geometric mixing of the underlying Markov chain. □

Lemma A.4. *Let $X = (X_k : k \geq 0)$ be an irreducible and aperiodic DTMC on a finite-state space $\mathcal{S}$ with transition probability matrix P and a unique stationary distribution π. Let $f : \mathcal{S} \to \mathbb{R}$ be any function, and let $\bar{f} := \sum_{x\in\mathcal{S}} \pi(x)f(x)$ denote its stationary expectation. Then,*

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k)|X_0 \sim \pi_0\right] = \lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k)|X_0 \sim \pi\right].$$

Proof. For $k \in \mathbb{N}_+$, let $\bar{f}_k := \mathbb{E}_{\pi_0}[f(X_k)]$. Then,

$$\begin{aligned}\frac{1}{n} \operatorname{Var}\left[\sum_{k=0}^{n-1} f(X_k)|X_0 \sim \pi_0\right] &= \frac{1}{n} \mathbb{E}_{\pi_0}\left[\left(\sum_{k=0}^{n-1} \big(f(X_k) - \bar{f}_k\big)\right)^2\right] \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_{\pi_0}\big[(f(X_k) - \bar{f}_k)^2\big] + \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} \mathbb{E}_{\pi_0}\big[\big(f(X_i) - \bar{f}_i\big)\big(f(X_k) - \bar{f}_k\big)\big]. \quad (26)\end{aligned}$$

Clearly,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_{\pi_0}\big[(f(X_k) - \bar{f}_k)^2\big] = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{E}_{\pi}\big[(f(X_t) - \bar{f})^2\big]. \quad (27)$$

This follows from the following:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_{\pi_0}\big[(f(X_k) - \bar{f}_k)^2\big] &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_{\pi_0}\big[(f(X_k) - \bar{f})^2\big] - \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} (\bar{f} - \bar{f}_k)^2 \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_{\pi}\big[(f(X_k) - \bar{f})^2\big],\end{aligned}$$

where we used that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} (\bar{f} - \bar{f}_k)^2 = 0,$$

which follows from geometric mixing of the underlying Markov chain.

Let us now show that the second term in (26) converges to the right limit. Towards this, we first re-write it as

$$\mathcal{B} = \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} \left(\mathbb{E}_{\pi_0}\big[(f(X_i) - \bar{f})(f(X_k) - \bar{f})\big] + (\bar{f}_i - \bar{f})(\bar{f} - \bar{f}_k)\right),$$

and let the corresponding term under π be

$$\mathcal{B}' = \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} \mathbb{E}_{\pi}\big[(f(X_i) - \bar{f})(f(X_k) - \bar{f})\big].$$

Let

$$T_1 := \mathbb{E}_{\pi_0}\big[(f(X_i) - \bar{f})(f(X_k) - \bar{f})\big] \quad \text{and} \quad T_2 := \mathbb{E}_{\pi}\big[(f(X_i) - \bar{f})(f(X_k) - \bar{f})\big].$$

Then,

$$\begin{aligned}|\mathcal{B} - \mathcal{B}'| &= \left|\frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} (T_1 - T_2) + \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} (\bar{f}_i - \bar{f})(\bar{f} - \bar{f}_k)\right| \\ &\leq \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} |T_1 - T_2| + \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} \big|\bar{f}_i - \bar{f}\big|\,\big|\bar{f} - \bar{f}_k\big| \\ &\overset{(a)}{\leq} \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} |T_1 - T_2| + \frac{2}{n} \sum_{k=0}^{n-1} \sum_{i<k} 4f_{\max}^2 C\alpha^k, \quad (28)\end{aligned}$$

where we used Lemma A.3 to get the bound in (a) above, and recall that $f_{\max} := \max_x f(x)$. Clearly, the second term converges to 0 as $n \to \infty$. Let us now bound the first term.

$$\begin{aligned}
|T_1 - T_2| &= \left| \sum_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} \sum_{z\in\mathcal{S}} (\pi_0(x) - \pi(x))\, P^i(x,y) P^{k-i}(y,z)(f(y)-\bar{f})(f(z)-\bar{f}) \right| \\
&\le 4 f_{\max}^2 \sum_{z\in\mathcal{S}} \left| \sum_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} (\pi_0(x) - \pi(x))\, P^i(x,y) P^{k-i}(y,z) \right| \\
&= 4 f_{\max}^2 \sum_{z\in\mathcal{S}} \left| \sum_{x\in\mathcal{S}} \big(\pi_0(x) P^k(x,z) - \pi(x) P^k(x,z)\big) \right| \\
&= \sum_{z\in\mathcal{S}} \left| \sum_{x\in\mathcal{S}} \pi_0(x) P^k(x,z) - \pi(z) \right| f_{\max}^2 \\
&\le \sum_{z\in\mathcal{S}} \sum_{x\in\mathcal{S}} \pi_0(x) \left| P^k(x,z) - \pi(z) \right| f_{\max}^2 \\
&= 2L^2 \sum_{x\in\mathcal{S}} \pi_0(x) \| P^k(x,\cdot) - \pi \|_{\text{T.V.}} \\
&\le 2 f_{\max}^2 C \alpha^k.
\end{aligned}$$

Using the above bound in (28), we get that $\lim_{n\to\infty} |\mathcal{B} - \mathcal{B}'| \le 0$, proving the desired result. □

B. Proof of Results in Section 4

B.1. Proof of Lemma 4.1

Clearly,

$$2\mathbb{E}_{d_\mu}[(r(S,A) - r(S',A'))Q_\mu(S,A)] = 2\mathbb{E}_{d_\mu}[(r(S,A) - J_\mu)Q_\mu(S,A)]\,.$$

This follows from linearity of expectation and the fact that (S,A) and (S',A') are independent samples.

Next, consider the second term in the expression in lemma.

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_{d_\mu}\big[(r(S,A) - r(S',A'))^2\big] &= \frac{1}{2}\mathbb{E}_{d_\mu}\big[r^2(S,A) + r^2(S',A') - 2r(S,A)r(S',A')\big] \\
&\overset{(a)}{=} \mathbb{E}_{d_\mu}\big[r^2(S,A)\big] - J_\mu^2 \\
&= \mathbb{E}_{d_\mu}\big[(r(S,A) - J_\mu)^2\big]\,,
\end{aligned}$$

where again (a) follows from linearity of expectation and independence.

Combining the two terms, we get that the rhs of expression in lemma is exactly κ_μ from (7). □

B.2. Proof of Theorem 4.3

For $(s,a) \in \mathcal{S}\times\mathcal{A}$, let $\mathbf{1}(s,a) \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|}$ be the indicator for (s,a), i.e., it equals 1 at the $(s,a)^{\text{th}}$ coordinate, and 0 otherwise. Furthermore, for $k \in \{1,2,\ldots\}$, define $\mathbf{1}_k := \mathbf{1}(S_k, A_k)$, indicator for (S_k, A_k). Then, setting $d = |\mathcal{S}|\,|\mathcal{A}|$, basis vectors in the setup in Section 5 as standard basis, and for $k \in \{1,2,\ldots\}$, $\phi_k = \mathbf{1}_k$, we recover Algorithm 1 from updates in (17), (18), and (19).

Further, since in this setting, the set of fixed points Q_μ of (4) lies in the span of the basis vectors, κ^* from (20) is same as κ_μ. Hence, guarantees in Theorem 5.2 with these adjustments, reduce to those in Theorem 4.3, proving the result.

B.3. Proof of Corollary 4.5

We prove the sample complexity bound separately for the algorithm with a constant step size, and that with the diminishing step size. But first, observe that $c_1 = O(1/\Delta_1)$ implying that $\eta = O(1/\Delta_1)$. Furthermore, $\xi_1 = O(\|Q^*_\mu\|_2^2)$, and $\xi_2 = O(\|Q^*_\mu\|_2^2/\Delta_1^2)$.

(a) For Algorithm 1 with a constant step size α, to estimate κ_μ up to a mean estimation error of at most ϵ, we require

$$k \geq L \log \frac{1}{\alpha} + O\left(\frac{\log(\xi_1/\epsilon^2)}{\alpha\Delta_1}\right) = O\left(\frac{\log^2 \frac{1}{\epsilon}}{\epsilon^2}\right)\tilde{O}\left(\frac{L\|Q^*_\mu\|_2^2}{\Delta_1^4}\right).$$

This follows from Theorem 4.3(a), where setting the second term to at most ϵ^2 gives

$$\alpha = O\left(\frac{\epsilon^2}{\log \frac{1}{\epsilon}}\right)\tilde{O}\left(\frac{\tilde{\Delta}_1}{L\xi_2}\right).$$

Choosing α to satisfy this condition, and setting the first term in the error bound to ϵ^2, we get the required sample complexity bound.

(b) Next, for Algorithm 1 with diminishing step size $\alpha_i = \frac{\alpha}{i+h}$ for all $i \geq 1$, for estimating κ_μ up to a mean estimation error at most ϵ, we require

$$k = O\left(\frac{\log^2 \frac{1}{\epsilon}}{\epsilon^2}\right)\tilde{O}\left(\frac{L\|Q^*_\mu\|_2^2}{\Delta_1^4}\right).$$

This follows from the mean-square estimation error bound in Theorem 4.3(b) by setting the rate-determining second term to at most ϵ^2, after optimizing over α. Note that α such that $\alpha\tilde{\Delta}_1 = 2$ (or $\alpha = O(1/\tilde{\Delta}_1)$) is the optimal choice.

C. Proofs for Results in Section 5.2

C.1. Stationary-Average Matrices

The average matrices A and b are given by

$$A = \begin{bmatrix} -c_1 & \vec{\mathbf{0}}^T & 0 \\ -\Pi_{2,E}\Phi^T D_\mu \vec{\mathbf{1}} & \Pi_{2,E}\Phi^T D_\mu (P_2 - I)\Phi & \vec{\mathbf{0}} \\ 0 & 2c_2(\mathcal{R}^T - J_\mu \vec{\mathbf{1}}^T) D_\mu \Phi & -c_2 \end{bmatrix},$$

and

$$b = \begin{bmatrix} c_1 \mathbb{E}_{d_\mu}[r(S,A)] \\ \Pi_{2,E}\Phi^T D_\mu \mathcal{R} \\ -\frac{c_2}{2}\mathbb{E}_{d_\mu}\Big[\big(r(S,A) - r(S',A')\big)^2\Big] \end{bmatrix},$$

where P_2 is the transition matrix defined in (1), and I is the identity matrix of the same dimensions. This follows from the observation that the stationary expectation of each entry of matrix $A(\cdot)$ and $b(\cdot)$ is given by the corresponding entry of A and b, respectively.

C.2. Algorithm's Limit: A Discussion

Let $\Theta^* = [J_\mu\ \theta^*\ \kappa^*]^T$, where θ^* is the unique vector in E that satisfies (uniqueness follows from Assumption 5.1)

$$\Phi\theta^* = \Pi_{D_\mu, W_\Phi}(T_\mu \Phi\theta^*),$$

where T_μ is an operator that for a vector $V \in \mathbb{R}^{|S||A|}$, satisfies $T_\mu V = \mathcal{R} - J_\mu \mathbf{e} + P_2 V$, and κ^* is given by

$$\kappa^* = 2\mathbb{E}_{d_\mu}[(r(S,A) - J_\mu)[\Phi\theta^*](S,A)] - \frac{1}{2}\mathbb{E}_{d_\mu}\big[(r(S,A) - J_\mu)^2\big],$$

where $[\Phi\theta^*](S,A)$ denotes the $(S,A)^{th}$ entry of the vector $\Phi\theta^*$, which corresponds to $Q_\mu(S,A)$ in the tabular setting.

Here, Π_{D_μ,W_Φ} is the projection in D_μ norm on the column space of Φ and is given by

$$\Pi_{D_\mu,W_\Phi} = \Phi(\Phi^T D_\mu \Phi)^{-1}\Phi^T D_\mu.$$

Let $\Theta_\infty := [J_\infty\ \theta_\infty\ \kappa_\infty]$, where $\theta_\infty \in E$. We now show that $\Theta_\infty = \Theta^*$ is the unique solution to $A\Theta_\infty + b = 0$. This equation corresponds to

$$-c_1 J_\infty + c_1 \mathbb{E}_{d_\mu}[r(S,A)] = 0, \tag{29}$$

$$-\Pi_{2,E}\Phi^T D_\mu e J_\infty + \Pi_{2,E}\Phi^T D_\mu (P_2 - I)\Phi\theta_\infty + \Pi_{2,E}\Phi^T D_\mu \mathcal{R} = 0, \tag{30}$$

$$2c_2(\mathcal{R}^T - J_\mu e^T)D_\mu\Phi\theta_\infty - c_2\kappa_\infty - \frac{c_2}{2}\mathbb{E}_{d_\mu}\big[(r(S,A) - r(S',A'))^2\big] = 0. \tag{31}$$

Clearly, (29) implies $J_\infty = J_\mu$. Moreover, if $\theta_\infty = \theta^* + a\theta_e$, for $a \in \mathbb{R}$, where θ_e is the unique vector in $\mathbb{R}^d$ such that $\Phi\theta_e = \mathbf{e}$, then (31) implies $\kappa_\infty = \kappa^*$.

Next, we show that $\theta_\infty = \theta^*$. Let's use $J_\infty = J_\mu$, and re-write the LHS of (30) as below:

$$\Pi_{2,E}\Phi^T D_\mu \left((\mathcal{R} - J_\mu \mathbf{e}) + (P_2 - I)\Phi\theta_\infty\right).$$

Clearly, it equals 0 for $\theta_\infty = \theta^*$. This follows from the definition of θ^*. Hence, θ^* is a solution for (30). Furthermore, $\theta \neq \theta^* + c\theta_e$ for any $c \in \mathbb{R}$, since $\theta^* + c\theta_e \notin E$. We now show that there does not exist any $\theta \in E$ different from θ^* that satisfies (30). To this end, suppose such a $\theta' \in E$ exists. Then, (30) evaluated at θ' re-writes as

$$\Pi_{2,E}\Phi^T D_\mu \underbrace{(-J_\infty \mathbf{e} + (P_2 - I)\Phi\theta^* + \mathcal{R})}_{=0} + \Pi_{2,E}\Phi^T D_\mu (P_2 - I)\Phi(\theta' - \theta^*),$$

where the first term equals 0. Now, recall that $\Pi_{2,E} = I - \theta_e\theta_e^T$, with the convention that $\theta_e = 0$ if $\mathbf{e} \notin W_\Phi$, the column space of Φ. Further, $\theta_e^T\Phi^T = \mathbf{e}^T \in \mathbb{R}^{|S||A|}$, which implies

$$\theta_e^T \Phi^T D_\mu (P_2 - I) = 0.$$

Using these, the second term above equals

$$\Phi^T D_\mu (P_2 - I)\Phi(\theta' - \theta^*) - \theta_e \underbrace{\theta_e^T \Phi^T D_\mu (P_2 - I)\Phi(\theta' - \theta^*)}_{=0}.$$

Now, the first term above is non-zero since $\theta' - \theta^* \in E$, and hence, $\Phi(\theta' - \theta^*)$ is non-constant vector that does not belong to the null space of $P_2 - I$. Thus, $\nexists \theta' \in E$ different from θ^* that also satisfies (30).

C.3. Proof of Lemma 5.4

In this appendix, we will show that

$$\min_{\substack{\Theta \in \mathbb{R}\times E\times\mathbb{R} \\ \|\Theta\|_2^2 = 1}} -\Theta^T A\Theta > 0. \tag{32}$$

Recall that for $\theta \in E$, $\Phi\theta$ is a non-constant vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Thus, $\theta^T\Phi^T D_\mu(I - P_2)\Phi\theta > 0$, for $\theta \in E$ (Tsitsiklis & Van Roy, 1997, Lemma 7). Since the set $\left\{\theta \in E : \|\theta\|_2^2 = 1\right\}$ is non-empty and compact, by extreme value theorem, we have

$$\Delta_2 := \min_{\|\theta\|_2 = 1, \theta \in E} \theta^T\Phi^T D_\mu (I - P_2)\Phi\theta > 0.$$

Next, expressing $\Theta \in \mathbb{R} \times E \times \mathbb{R}$ as $\Theta^T = [J\ \theta^T\ \kappa]$, we re-write the minimization problem in (32) as

$$\begin{aligned}
&\min_{\substack{\sqrt{J^2+\|\theta\|_2^2+\kappa^2}=1 \\ J\in\mathbb{R},\kappa\in\mathbb{R},\theta\in E}} -\Theta^T A\Theta \\
=&\min_{\substack{\sqrt{J^2+\|\theta\|_2^2+\kappa^2}=1 \\ J\in\mathbb{R},\kappa\in\mathbb{R},\theta\in E}} c_1J^2+c_2\kappa^2+J\theta^T\Pi_{2,E}\Phi^T D_\mu\mathbf{e}+\theta^T\Pi_{2,E}\Phi^T D_\mu(I-P_2)\Phi\theta-2c_2\kappa(\mathcal{R}^T-J_\mu\mathbf{e}^T)D_\mu\Phi\theta \\
\overset{(a)}{=}&\min_{\substack{\sqrt{J^2+\|\theta\|_2^2+\kappa^2}=1 \\ J\in\mathbb{R},\kappa\in\mathbb{R},\theta\in E}} c_1J^2+c_2\kappa^2+J\theta^T\Phi^T D_\mu\mathbf{e}+\theta^T\Phi^T D_\mu(I-P_2)\Phi\theta-2c_2\kappa(\mathcal{R}^T-J_\mu\mathbf{e}^T)D_\mu\Phi\theta \\
\overset{(b)}{\geq}&\min_{\substack{\sqrt{J^2+\|\theta\|_2^2+\kappa^2}=1 \\ J\in\mathbb{R},\kappa\in\mathbb{R},\theta\in E}} c_1J^2+c_2\kappa^2+J\theta^T\Phi^T D_\mu\mathbf{e}+\Delta_2\|\theta\|_2^2+2c_2\theta^T\Phi^T D_\mu\kappa(J_\mu\mathbf{e}-\mathcal{R}), \qquad (33)
\end{aligned}$$

where (a) follows since $\theta \in E$, and (b) follows since $\theta^T\Phi^T D_\mu(I-P_2)\Phi\theta \geq \Delta_2\|\theta\|_2^2$.

Next, consider

$$\begin{aligned}
\left|\theta^T\Phi^T D_\mu(J\mathbf{e}+2c_2\kappa(J_\mu\mathbf{e}-\mathcal{R}))\right| &= \left|J\theta^T\Phi^T D_\mu\mathbf{e}+2c_2\kappa\theta^T\Phi^T D_\mu(J_\mu\mathbf{e}-\mathcal{R})\right| \\
&\leq |J|\left|\theta^T\Phi^T D_\mu\mathbf{e}\right|+2c_2\left|\kappa\right|\left|\theta^T\Phi^T D_\mu(J_\mu\mathbf{e}-\mathcal{R})\right| \\
&\leq |J|\left\|\Phi\theta\right\|_\infty\|D_\mu\mathbf{e}\|_1+2c_2\left|\kappa\right|\left\|\Phi\theta\right\|_\infty\|D_\mu(J_\mu\mathbf{e}-\mathcal{R})\|_1.
\end{aligned}$$

Next, let $r_{\max} := \|\mathcal{R}\|_\infty$. Since $\|D_\mu\mathbf{e}\|_1 = 1$, and $\|D_\mu(J_\mu\mathbf{e}-\mathcal{R})\|_1 \leq 2r_{\max}$, continuing the above inequalities, we further get

$$\left|\theta^T\Phi^T D_\mu(J\mathbf{e}+2c_2\kappa(J\mathbf{e}-\mathcal{R}))\right| \leq |J|\left\|\Phi\theta\right\|_\infty+4c_2\left|\kappa\right|\left\|\Phi\theta\right\|_\infty r_{\max}.$$

Next, observe that

$$\|\Phi\theta\|_\infty \leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\phi(s,a)\|_2\|\theta\|_2 \leq \|\theta\|_2,$$

where the first inequality above follows from the definition of $\|\cdot\|_\infty$ and Holder's inequality, and the second from Assumption 5.1, giving

$$\left|\theta^T\Phi^T D_\mu(J\mathbf{e}+2c_2\kappa(J\mathbf{e}-\mathcal{R}))\right| \leq |J|\left\|\theta\right\|_2+4c_2\left|\kappa\right|\left\|\theta\right\|_2 r_{\max}.$$

Using this in (33), we have

$$\min_{\substack{\sqrt{J^2+\|\theta\|_2^2+\kappa^2}=1 \\ J\in\mathbb{R},\kappa\in\mathbb{R},\theta\in E}} -\Theta^T A\Theta \ \geq \ \min_{\substack{\sqrt{J^2+\|\theta\|_2^2+\kappa^2}=1 \\ J\in\mathbb{R},\kappa\in\mathbb{R},\theta\in E}} c_1J^2+c_2\kappa^2+\Delta_2\|\theta\|_2^2-|J|\left\|\theta\right\|_2-4c_2\left|\kappa\right|\left\|\theta\right\|_2 r_{\max}.$$

The above minimization problem can be re-written as

$$\min_{\kappa\in[-1,1]}\ \min_{\substack{\theta\in E,J\in\mathbb{R} \\ \|\theta\|_2^2+J^2=1-\kappa^2}} c_2\kappa^2+c_1J^2-(|J|+4c_2\left|\kappa\right|r_{\max})\left\|\theta\right\|_2+\Delta_2\|\theta\|_2^2.$$

Now, consider the following bounds on the above minimization problem:

$$\begin{aligned}
&\min_{\kappa\in[-1,1]}\left\{c_2\kappa^2+\min_{\substack{\theta\in E,J\in\mathbb{R} \\ \|\theta\|_2^2+J^2=1-\kappa^2}}\left\{\,c_1J^2-(|J|+4c_2\left|\kappa\right|r_{\max})\left\|\theta\right\|_2+\Delta_2\|\theta\|_2^2\right\}\right\} \\
=&\min_{\kappa\in[-1,1]}\left\{c_2\kappa^2+\min_{J\in\left[-\sqrt{1-\kappa^2},\sqrt{1-\kappa^2}\right]}\left\{c_1J^2-(|J|+4c_2\left|\kappa\right|r_{\max})\sqrt{1-\kappa^2-J^2}+\Delta_2\big(1-\kappa^2-J^2\big)\right\}\right\} \\
=&\min_{\kappa\in[-1,1]}\left\{(c_2-\Delta_2)\kappa^2+\Delta_2+\min_{J\in\left[0,\sqrt{1-\kappa^2}\right]}\left\{\,(c_1-\Delta_2)J^2-(J+4c_2\left|\kappa\right|r_{\max})\sqrt{1-\kappa^2-J^2}\right\}\right\} \\
\geq&\Delta_2+\min_{\kappa\in[-1,1]}\left\{(c_2-\Delta_2)\kappa^2-4c_2\left|\kappa\right|r_{\max}\sqrt{1-\kappa^2}+\min_{J\in\left[0,\sqrt{1-\kappa^2}\right]}\left\{\,(c_1-\Delta_2)J^2-J\sqrt{1-\kappa^2-J^2}\right\}\right\}. \qquad (34)
\end{aligned}$$

Let's first solve the inner minimization problem. Let $y_1 \in \mathbb{R}$ be such that

$$c_1 - \Delta_2 \geq y_1. \tag{35}$$

Then, for $c_1 \geq 0$ and ≥ 0,

$$\begin{aligned}
&\min_{J \in \left[0, \sqrt{1-\kappa^2}\right]} \left\{ (c_1 - \Delta_2) J^2 - J\sqrt{1 - \kappa^2 - J^2} \right\} \\
= &\min_{x \in \left[0, 1-\kappa^2\right]} \left\{ (c_1 - \Delta_2) x - \sqrt{x\left(1-\kappa^2\right) - x^2} \right\} \\
\geq &\min_{x \in \left[0, 1-\kappa^2\right]} \left\{ y_1 x - \sqrt{x\left(1-\kappa^2\right) - x^2} \right\} \\
= &\frac{1-\kappa^2}{2}\left(y_1 - \sqrt{y_1^2 + 1}\right),
\end{aligned}$$

where we used Lemma D.1 to get the last equality. Using this in (34), we have the required minimum is at least

$$\begin{aligned}
&\Delta_2 + \min_{\kappa \in [-1,1]} \left\{ (c_2 - \Delta_2)\kappa^2 - 4c_2 |\kappa| r_{\max} \sqrt{1-\kappa^2} + \frac{1-\kappa^2}{2}\left(y_1 - \sqrt{y_1^2+1}\right) \right\} \\
= &\Delta_2 + \frac{y_1 - \sqrt{y_1^2+1}}{2} + \min_{\kappa \in [-1,1]} \left\{ \left(c_2 - \Delta_2 - \frac{y_1 - \sqrt{y_1^2+1}}{2}\right)\kappa^2 - 4c_2 |\kappa| r_{\max}\sqrt{1-\kappa^2} \right\} \\
= &\Delta_2 + \frac{y_1 - \sqrt{y_1^2+1}}{2} + \min_{x \in [0,1]} \left\{ \left(c_2 - \Delta_2 - \frac{y_1 - \sqrt{y_1^2+1}}{2}\right) x - 4c_2 r_{\max}\sqrt{x - x^2} \right\}.
\end{aligned}$$

Next, let $y_2 \in \mathbb{R}$ such that

$$c_2 - \Delta_2 - \frac{y_1 - \sqrt{y_1^2+1}}{2} \geq y_2. \tag{36}$$

Then the previous expression is lower bounded by

$$\Delta_2 + \frac{y_1 - \sqrt{y_1^2+1}}{2} + \min_{x \in [0,1]} \left\{ y_2 x - 4c_2 r_{\max}\sqrt{x - x^2} \right\},$$

which equals (using Lemma D.1)

$$\Delta_2 + \frac{y_1 - \sqrt{y_1^2+1}}{2} + \frac{y_2 - \sqrt{y_2^2 + 16c_2^2 r_{\max}^2}}{2}. \tag{37}$$

Choosing

$$y_1 = \frac{1}{2\Delta_2} - \frac{\Delta_2}{2}, \quad \text{and} \quad y_2 = \frac{c_2^2}{4\Delta_2}(1 + 32 r_{\max}^2) - \frac{16\Delta_2 r_{\max}^2}{1 + 32 r_{\max}^2}$$

ensures that (37) is at least

$$\frac{\Delta_2}{2(1 + 32 r_{\max}^2)},$$

which is strictly positive. Choosing

$$c_1 \geq \frac{1}{2}\left(\frac{1}{\Delta_2} + \Delta_2\right),$$

and

$$c_2 \in \left[\frac{2\Delta_2}{1 + 32 r_{\max}^2}\left(1 - \frac{1}{\sqrt{2}}\right), \frac{2\Delta_2}{1 + 32 r_{\max}^2}\left(1 + \frac{1}{\sqrt{2}}\right)\right]$$

satisfies (35) and (36).

C.4. Proof of Theorem 5.2

Results in Theorem 5.2 follow from much stronger results stated in Theorem C.1 below, which we will prove in this section.

Theorem C.1. *Consider the iterates $\Theta_k^T = [J_k \ \theta_k^T \ \kappa_k]$ of the proposed algorithm in Section 5.1 with c_1 and c_2 satisfying:*

$$c_1 \geq \frac{1}{2}\left(\frac{1}{\Delta_2} + \Delta_2\right), \quad \text{and} \quad c_2 \in \left[2\tilde{\Delta}_2\left(1 - \frac{1}{\sqrt{2}}\right), 2\tilde{\Delta}_2\left(1 + \frac{1}{\sqrt{2}}\right)\right].$$

Let $\xi_1 = (1 + \|\Theta^*\|_2)^2$ and $\xi_2 = 520\eta^2\left(\|\Theta^*\|_2 + 1\right)^2$, where η is defined in (16), and Θ^* before (20).

(a) Let $\alpha_i = \alpha$ for all i, such that $\tilde{\Delta}_2\alpha < 2$, and

$$\alpha\tau(\alpha) \leq \min\left\{\frac{1}{4\eta}, \frac{\tilde{\Delta}_2}{260\eta^2}\right\}.$$

Then, for all $k \geq \tau(\alpha)$,

$$\mathbb{E}\left[(J_k - J_\mu)^2 + \|\theta_k - \theta^*\|_2^2 + (\kappa_k - \kappa^*)^2\right] \leq \xi_1\left(1 - \frac{\alpha\tilde{\Delta}_2}{2}\right)^{k-\tau(\alpha)} + \frac{\xi_2\alpha\tau(\alpha)}{\tilde{\Delta}_2}.$$

(b) Let $\alpha_i = \frac{\alpha}{i+h}$ for all i, with α and h chosen so that $2 < \alpha\tilde{\Delta}_2 < 2h$. Let k^* be the smallest positive integer such that $\sum_{i=0}^{k^*-1}\alpha_i \leq \frac{1}{2\eta}$, and for all $k \geq k^*$,

$$\sum_{i=k-\tau(\alpha_k)}^{k-1}\alpha_i \leq \min\left\{\frac{1}{4\eta}, \frac{\tilde{\Delta}_2}{260\eta^2}\right\}.$$

Then, for all $k \geq k^*$, and L defined after (15),

$$\mathbb{E}\left[(J_k - J_\mu)^2 + \|\theta_k - \theta^*\|_2^2 + (\kappa_k - \kappa^*)^2\right] \leq \xi_1\left(\frac{k^*+h}{k+h}\right)^{\alpha\tilde{\Delta}_2/2} + \left(\frac{4\alpha^2\xi_2 eL}{\tilde{\Delta}_2\alpha - 2}\right)\frac{\ln(k+h) - \ln\alpha}{k+h+1}.$$

Notation. Recall that $\Pi_{2,E}$ denotes the orthogonal projection of vectors in $\mathbb{R}^d$ (in 2-norm) on the subspace $E \in \mathbb{R}^d$. Consider the sequence of iterates $\Theta_k := [J_k \ \theta_k^T \ \kappa_k]$ generated by the proposed algorithm in Section 5.1 with $\Theta_0 = \vec{\mathbf{0}}$ and $\theta_k \in E$. Recall from Section 5.4 that these iterates can be rewritten as a linear stochastic-approximation update given below:

$$\Theta_{k+1} = \Theta_k + \alpha_k\left(A(Y_k)\Theta_k + b(Y_k)\right),$$

where $Y_k := (S_k, A_k, S_{k+1}, A_{k+1}, S_k', A_k')$,

$$A(Y_k) := \begin{bmatrix} -c_1 & \vec{\mathbf{0}}^T & 0 \\ -\Pi_{2,E}\phi_k & \Pi_{2,E}\phi_k(\phi_{k+1}^T - \phi_k^T) & \vec{\mathbf{0}} \\ 0 & 2c_2(r_k - r_k')\phi_k^T & -c_2 \end{bmatrix}, \quad b(Y_k) := \begin{bmatrix} c_1 r_k \\ r_k\Pi_{2,E}\phi_k \\ -\frac{c_2}{2}(r_k - r_k')^2 \end{bmatrix},$$

and recall that $\phi_k^T := \phi^T(S_k, A_k) = [\tilde{\phi}_1(S_k, A_k) \ \ldots \ \tilde{\phi}_d(S_k, A_k)]$. Moreover, recall that

$$A = \begin{bmatrix} -c_1 & \vec{\mathbf{0}}^T & 0 \\ -\Pi_{2,E}\Phi^T D_\mu\vec{\mathbf{1}} & \Pi_{2,E}\Phi^T D_\mu(P_2 - I)\Phi & \vec{\mathbf{0}} \\ 0 & 2c_2(\mathcal{R}^T - J_\mu\vec{\mathbf{1}}^T)D_\mu\Phi & -c_2 \end{bmatrix}, \quad \text{and} \quad b = \begin{bmatrix} c_1\mathbb{E}_{d_\mu}[r(S,A)] \\ \Pi_{2,E}\Phi^T D_\mu\mathcal{R} \\ -\frac{c_2}{2}\mathbb{E}_{d_\mu}\left[(r(S,A) - r(S',A'))^2\right] \end{bmatrix},$$

where P_2 is a transition probability matrix defined in (1), and I is the identity matrix of the same dimensions.

Next, define $\tilde{A}(Y_k)$ and $\tilde{b}(Y_k)$ so that $A(Y_k) = \Pi\tilde{A}(Y_k)$ and $b(Y_k) = \Pi\tilde{b}(Y_k)$, where

$$\Pi := \begin{bmatrix} 1 & \vec{\mathbf{0}}^T & 0 \\ \vec{\mathbf{0}} & \Pi_{2,E} & \vec{\mathbf{0}} \\ 0 & \vec{\mathbf{0}}^T & 1 \end{bmatrix},$$

that is,

$$\tilde{A}(Y_k) = \begin{bmatrix} -c_1 & \vec{\mathbf{0}}^T & 0 \\ -\phi_k & \phi_k(\phi_{k+1}^T - \phi_k^T) & \vec{\mathbf{0}} \\ 0 & 2c_2(r_k - r_k')\phi_k^T & -c_2 \end{bmatrix}, \quad \tilde{b}(Y_k) = \begin{bmatrix} c_1 r_k \\ r_k\phi_k \\ -\frac{c_2}{2}(r_k - r_k')^2 \end{bmatrix}.$$

Similarly, define $\tilde{A}$ and $\tilde{b}$ so that $A = \Pi\tilde{A}$ and $b = \Pi\tilde{b}$, i.e.,

$$\tilde{A} := \begin{bmatrix} -c_1 & \vec{\mathbf{0}}^T & 0 \\ -\Phi^T D_\mu \vec{\mathbf{1}} & \Phi^T D_\mu (P_2 - I)\Phi & \vec{\mathbf{0}} \\ 0 & 2c_2(\mathcal{R}^T - J_\mu \vec{\mathbf{1}}^T) D_\mu \Phi & -c_2 \end{bmatrix}, \quad \text{and} \quad \tilde{b} = \begin{bmatrix} c_1 \mathbb{E}_{d_\mu}[r(S,A)] \\ \Phi^T D_\mu \mathcal{R} \\ -\frac{c_2}{2}\mathbb{E}_{d_\mu}\Big[\big(r(S,A) - r(S',A')\big)^2\Big] \end{bmatrix},$$

We now bound $\|\Theta_{k+1} - \Theta^*\|_2^2$. Towards this, we first establish the properties of the associated matrices.

Bounds on $A(Y_t)$ and $b(Y_t)$. For a matrix M, let $\|M\|_2$ and $\|M\|_F$ denote its induced 2-norm and Frobenius norm, respectively. Then, recall that $\|M\|_2 \leq \|M\|_F$.

$$\begin{aligned} \|A(Y_t)\|_2 &= \|\Pi\tilde{A}(Y_t)\|_2 \\ &\leq \|\Pi\|_2 \|A(\tilde{Y_t})\|_2, \end{aligned}$$

where the above inequality follows from sub-multiplicativity of the induced 2-norm. Now, since Π is an orthogonal projection matrix, $\|\Pi\|_2 = 1$. Thus, we have

$$\begin{aligned} \|A(Y_t)\|_2 &\leq \|\tilde{A}(Y_t)\|_2 \\ &\overset{(a)}{\leq} \|\tilde{A}(Y_t)\|_F \\ &= \sqrt{c_1^2 + 1 + \|\phi_k(\phi_{k+1}^T - \phi_k^T)\|_F^2 + 16c_2^2 r_{\max}^2 + c_2^2} \\ &\overset{(b)}{\leq} \sqrt{c_1^2 + 1 + (\|\phi_k\phi_{k+1}^T\|_F + \|\phi_k\phi_k^T\|_F)^2 + 16c_2^2 r_{\max}^2 + c_2^2} \\ &\overset{(c)}{=} \sqrt{c_1^2 + 1 + (\|\phi_k\|_2\|\phi_{k+1}\|_2 + \|\phi_k\|_2\|\phi_k\|_2)^2 + 16c_2^2 r_{\max}^2 + c_2^2} \\ &\overset{(d)}{\leq} \sqrt{c_1^2 + 5 + 16c_2^2 r_{\max}^2 + c_2^2}, \end{aligned}$$

where (a) follows since for a matrix M, $\|M\|_2 \leq \|M\|_F$, (b) uses triangle inequality for $\|\cdot\|_F$, and (d) uses Assumption 5.1. To conclude (c), let $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}^d$. Observe the following equalities for $\|x_1 x_2^T\|_F$.

$$\|x_1 x_2^T\|_F = \sqrt{\text{Trace}(x_1 x_2^T x_2 x_1^T)} = \sqrt{\text{Trace}(x_1 x_1^T \|x_2\|_2^2)} = \|x_2\|_2 (x_1^T x_1) = \|x_1\|_2 \|x_2\|_2.$$

Next, to bound $b(Y_t)$,

$$\|b(Y_t)\|_2 = \|\Pi\,\tilde{b}(Y_t)\|_2 \leq \|\tilde{b}(Y_t)\|_2 = r_{\max}\sqrt{c_1^2 + 1 + 2c_2^2}.$$

General finite-time bound. Let $\Theta_k^T := \begin{bmatrix} J_k & \theta_k^T & \kappa_k \end{bmatrix}$, denote the k^{th} iterate of the proposed algorithm in Section 5.1, and let $\Theta^{*T} := [J_\mu \;\; \theta^{*T} \;\; \kappa^*]$, where $\theta^* \in E \subset \mathbb{R}^d$ and $\kappa^* \in \mathbb{R}$ are such that $A\Theta^* + b = 0$. For simplicity of notation, let

$$\mathbb{E}_{k-\tau_k}[\cdot] := \mathbb{E}\big[\cdot|\Theta_{k-\tau(\alpha_k)}, Y_{k-\tau(\alpha_k)}\big],$$

where for $\delta > 0$, $\tau(\delta)$ is the mixing time of the Markov chain $\mathcal{M}_3$, and is introduced in Definition 4.2. Also, for $0 < k_1 < k_2$, let

$$\alpha_{k_1,k_2} := \sum_{i=k_1}^{k_2} \alpha_i.$$

Define $A_k := A(Y_k), \quad b_k := b(Y_k)$, and let

$$A_{\max} := \sqrt{c_1^2 + 5 + 16c_2^2 r_{\max}^2 + c_2^2}, \quad b_{\max} := r_{\max}\sqrt{c_1^2 + 1 + 2c_2^2}, \quad \eta := \max\left\{A_{\max}, b_{\max}\right\}.$$

For any $k \geq 0$, we have

$$\begin{aligned}
&\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k+1} - \Theta^*\|_2^2 - \|\Theta_k - \Theta^*\|_2^2\right] \\
=&\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k+1} - \Theta_k + \Theta_k - \Theta^*\|_2^2 - \|\Theta_k - \Theta^*\|_2^2\right] \\
=&\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k+1} - \Theta_k\|_2^2 + 2(\Theta_k - \Theta^*)^T(\Theta_{k+1} - \Theta_k)\right] \\
=&\alpha_k^2\mathbb{E}_{k-\tau_k}\left[\|A_k\Theta_k + b_k\|_2^2\right] + 2\alpha_k\mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta^*)^T(A_k\Theta_k + b_k)\right] \\
=&\alpha_k^2\mathbb{E}_{k-\tau_k}\left[\|A_k\Theta_k + b_k\|_2^2\right] + 2\alpha_k\mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta^*)^T(A_k\Theta_k + b_k - A\Theta_k - b)\right] \\
&\qquad + 2\alpha_k\mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta^*)^T(A\Theta_k + b)\right].
\end{aligned} \tag{38}$$

We will now bound each of the three terms in the above expressions.

Bounding $\|A_k\Theta_k + b_k\|_2^2$ in (38). Here we will use that the update matrices at each step are bounded, specifically that $\|A_k\|_2 \leq A_{\max}$ and $\|b_k\|_2 \leq b_{\max}$ for all $k \geq 1$.

$$\begin{aligned}
\|A_k\Theta_k + b_k\|_2^2 &\leq \left(\|A_k\|_2\|\Theta_k\|_2 + \|b_k\|_2\right)^2 \\
&\leq \left(A_{\max}\|\Theta_k\|_2 + b_{\max}\right)^2 \\
&\leq \eta^2\left(\|\Theta_k\|_2 + 1\right)^2 \\
&\leq \eta^2\left(\|\Theta_k - \Theta^*\|_2 + \|\Theta^*\|_2 + 1\right)^2 \\
&\leq 2\eta^2\left(\|\Theta_k - \Theta^*\|_2^2 + \left(\|\Theta^*\|_2 + 1\right)^2\right).
\end{aligned}$$

Bounding $(\Theta_k - \Theta^*)^T(A\Theta_k + b)$ in (38). To bound this term, we will use the fact that the limit point Θ^* satisfies $A\Theta^* + b = 0$. Using this, we have

$$(\Theta_k - \Theta^*)^T(A\Theta_k + b) = (\Theta_k - \Theta^*)^T(A\Theta_k - A\Theta^*) = (\Theta_k - \Theta^*)^T A(\Theta_k - \Theta^*) \overset{(a)}{\leq} -\frac{\tilde{\Delta}_2\|\Theta_k - \Theta^*\|_2^2}{2},$$

where (a) follows from Lemma 5.4 since $(\Theta_k - \Theta^*) \in \mathbb{R} \times E \times \mathbb{R}$.

Bounding $\mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta^*)^T(A_k\Theta_k + b_k - A\Theta_k - b)\right]$ in (38).

$$\begin{aligned}
&\mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta^*)^T(A_k\Theta_k + b_k - A\Theta_k - b)\right] \\
&= \mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta_{k-\tau(\alpha_k)} + \Theta_{k-\tau(\alpha_k)} - \Theta^*)^T(A_k\Theta_k + b_k - A\Theta_k - b)\right] \\
&= \mathbb{E}_{k-\tau_k}\left[(\Theta_k - \Theta_{k-\tau(\alpha_k)})^T(A_k\Theta_k + b_k - A\Theta_k - b)\right] \\
&\qquad + \mathbb{E}_{k-\tau_k}\left[(\Theta_{k-\tau(\alpha_k)} - \Theta^*)^T(A_k\Theta_k + b_k - A\Theta_k - b)\right].
\end{aligned} \tag{39}$$

Bounding the first term in (39),

$$\begin{aligned}
&\mathbb{E}_{k-\tau_k}\left[(\Theta_k-\Theta_{k-\tau(\alpha_k)})^T(A_k\Theta_k+b_k-A\Theta_k-b)\right]\\
&\leq \mathbb{E}_{k-\tau_k}\left[\left|(\Theta_k-\Theta_{k-\tau(\alpha_t)})^T(A_k\Theta_k+b_k-A\Theta_k-b)\right|\right]\\
&\leq \mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta_{k-\tau(\alpha_k)}\|_2\|A_k\Theta_k+b_k-A\Theta_k-b\|_2\right]\\
&\leq \mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta_{k-\tau(\alpha_k)}\|_2\left(\|A_k-A\|_2\|\Theta_k\|_2+\|b_k-b\|_2\right)\right]\\
&\leq 2\eta\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta_{k-\tau(\alpha_k)}\|_2\left(\|\Theta_k\|_2+1\right)\right]\\
&\overset{(a)}{\leq} 8\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[\left(\|\Theta_k\|_2+1\right)^2\right]\\
&\leq 8\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[\left(\|\Theta_k-\Theta^*\|_2+\|\Theta^*\|_2+1\right)^2\right]\\
&\leq 16\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2+\left(\|\Theta^*\|_2+1\right)^2\right],
\end{aligned}\tag{40}$$

where (a) follows since for any $k_1>0$ and $k_2>0$ such that

$$\alpha_{k_1,k_2}\leq\frac{1}{4\eta}\implies\|\Theta_{k_2}-\Theta_{k_1}\|\leq 4\eta\alpha_{k_1,k_2-1}(\|\Theta_{k_2}\|+1),\tag{41}$$

and the assumption on α_k that for $k\geq k^*$, $\alpha_{k-\tau(\alpha_k),k-1}\leq\frac{1}{4\eta}$. This follows from Chen et al. (2022, Lemma 2.3).

Next, consider the second term in (39).

$$\begin{aligned}
&\mathbb{E}_{k-\tau_k}\left[(\Theta_{k-\tau(\alpha_k)}-\Theta^*)^T(A_k\Theta_k+b_k-A\Theta_k-b)\right]\\
=&\mathbb{E}_{k-\tau_k}\left[(\Theta_{k-\tau(\alpha_k)}-\Theta^*)^T\left((A_k-A)\Theta_{k-\tau(\alpha_k)}+b_k-b+(A_k-A)(\Theta_k-\Theta_{k-\tau(\alpha_k)})\right)\right]\\
\leq&\underbrace{\left|(\Theta_{k-\tau(\alpha_k)}-\Theta^*)^T\mathbb{E}_{k-\tau_k}\left[(A_k-A)\Theta_{k-\tau(\alpha_k)}+b_k-b\right]\right|}_{A_1}\\
&+\underbrace{\left|(\Theta_{k-\tau(\alpha_k)}-\Theta^*)^T\mathbb{E}_{k-\tau_k}\left[(A_k-A)(\Theta_k-\Theta_{k-\tau(\alpha_k)})\right]\right|}_{A_2}.
\end{aligned}\tag{42}$$

We further bound the two terms in (42) separately.

$$\begin{aligned}
A_1&\leq\|\Theta_{k-\tau(\alpha_k)}-\Theta^*\|_2\left\|\mathbb{E}_{k-\tau_k}\left[(A_k-A)\Theta_{k-\tau(\alpha_k)}+b_k-b\right]\right\|_2\\
&=\|\Theta_{k-\tau(\alpha_k)}-\Theta^*\|_2\left\|\mathbb{E}_{k-\tau_k}\left[A_k-A\right]\Theta_{k-\tau(\alpha_k)}+\mathbb{E}_{k-\tau_k}\left[b_k-b\right]\right\|_2\\
&\overset{(a)}{\leq}\|\Theta_{k-\tau(\alpha_k)}-\Theta^*\|_2\left(\left\|\mathbb{E}_{k-\tau_k}\left[A_k\right]-A\right\|_2\|\Theta_{k-\tau(\alpha_k)}\|_2+\left\|\mathbb{E}_{k-\tau_k}\left[b_k\right]-b\right\|_2\right)\\
&\overset{(b)}{\leq}2\eta\alpha_k\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k-\tau(\alpha_k)}-\Theta^*\|_2\left(1+\|\Theta_{k-\tau(\alpha_k)}\|_2\right)\right]\\
&=2\eta\alpha_k\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k-\tau(\alpha_k)}-\Theta_k+\Theta_k-\Theta^*\|_2\left(1+\|\Theta_{k-\tau(\alpha_k)}-\Theta_k+\Theta_k-\Theta^*+\Theta^*\|_2\right)\right]\\
&\leq 2\eta\alpha_k\mathbb{E}_{k-\tau_k}\left[\left(\|\Theta_{k-\tau(\alpha_k)}-\Theta_k\|_2+\|\Theta_k-\Theta^*\|_2\right)\left(1+\|\Theta_{k-\tau(\alpha_k)}-\Theta_k\|_2+\|\Theta_k-\Theta^*\|_2+\|\Theta^*\|_2\right)\right].
\end{aligned}$$

The inequality (a) follows since for any matrix M and a vector x, we have $\|Mx\|_2\leq\|M\|_2\|x\|_2$, where $\|M\|_2$ is the induced 2-norm (or operator 2-norm) for M. Inequality (b) follows since, from the definition of the mixing time, we have

$$\left\|\mathbb{E}_{k-\tau_k}\left[A_k\right]-A\right\|_2\leq 2\eta\alpha_k\quad\text{and}\quad\left\|\mathbb{E}_{k-\tau_k}\left[b_k\right]-b\right\|\leq 2\eta\alpha_k,\tag{43}$$

and $\Theta_{k-\tau(\alpha_k)}$ is a constant with respect to the conditioning.

To see (43), consider the following

$$\begin{aligned}
\|\mathbb{E}_{k-\tau_k}[b_k] - b\|_2 &\leq \sup_{y' \in \mathcal{Y}} \left\| \sum_{y \in \mathcal{Y}} \left(\mathbb{P}(Y_k = y | Y_{k-\tau(\alpha_k)} = y') - \tilde{\pi}_\mu(Y_k = y) \right) b(y) \right\|_2 \\
&\leq \sup_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \left| \mathbb{P}(Y_k = y | Y_{k-\tau(\alpha_k)} = y') - \tilde{\pi}_\mu(Y_k = y) \right| \|b(y)\|_2 \\
&\leq 2\eta C \rho^{\tau(\alpha_k)} \\
&\leq 2\eta\rho\alpha_k \\
&\leq 2\eta\alpha_k,
\end{aligned}$$

where the last inequality follows since $\rho < 1$. Similarly,

$$\begin{aligned}
\|\mathbb{E}_{k-\tau_k}[A_k] - A\|_2 &\leq \sup_{y' \in \mathcal{Y}} \left\| \sum_{y \in \mathcal{Y}} \left(\mathbb{P}(Y_k = y | Y_{k-\tau(\alpha_k)} = y') - \tilde{\pi}_\mu(Y_k = y) \right) A(y) \right\|_2 \\
&\leq \sup_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \left| \mathbb{P}(Y_k = y | Y_{k-\tau(\alpha_k)} = y') - \tilde{\pi}_\mu(Y_k = y) \right| \|A(y)\|_2 \\
&\leq 2\eta\alpha_k.
\end{aligned}$$

Next, since for $0 < k_1 < k_2$,

$$\alpha_{k_1,k_2-1} \leq \frac{1}{4\eta} \implies \|\Theta_{k_2} - \Theta_{k_1}\|_2 \leq 1 + \|\Theta_{k_2}\|_2,$$

from the assumption that $\alpha_{k-\tau(\alpha_k),k-1} \leq \frac{1}{4\eta}$, we have that

$$\|\Theta_k - \Theta_{k-\tau(\alpha_k)}\|_2 \leq \|\Theta_k\|_2 + 1.$$

This follows from Chen et al. (2022, Lemma 2.3). Using this to further bound A_1, we have

$$A_1 \leq 2\eta\alpha_k \mathbb{E}_{k-\tau_k}\left[(1 + \|\Theta_k\|_2 + \|\Theta_k - \Theta^*\|_2)\,(2 + \|\Theta_k\|_2 + \|\Theta_k - \Theta^*\|_2 + \|\Theta^*\|_2) \right].$$

Continuing bounding the r.h.s. above,

$$\begin{aligned}
A_1 &\leq 2\eta\alpha_t \mathbb{E}_{k-\tau_k}\left[(1 + \|\Theta_k\|_2 + \|\Theta_k - \Theta^*\|_2)\,(2 + \|\Theta_k\|_2 + \|\Theta_k - \Theta^*\|_2 + \|\Theta^*\|_2) \right] \\
&\leq 4\eta\alpha_k \mathbb{E}_{k-\tau_k}\left[(1 + \|\Theta^*\|_2 + 2\|\Theta_k - \Theta^*\|_2)\,(1 + \|\Theta_k - \Theta^*\|_2 + \|\Theta^*\|_2) \right] \\
&\leq 8\eta\alpha_k \mathbb{E}_{k-\tau_k}\left[\left(1 + \|\Theta^*\|_2 + \|\Theta_k - \Theta^*\|_2\right)^2 \right] \\
&\leq 16\eta\alpha_k \mathbb{E}_{k-\tau_k}\left[\left(1 + \|\Theta^*\|_2\right)^2 + \|\Theta_k - \Theta^*\|_2^2 \right] \\
&\leq 16\eta^2\alpha_{k-\tau(\alpha_k),k-1} \mathbb{E}_{k-\tau_k}\left[\left(1 + \|\Theta^*\|_2\right)^2 + \|\Theta_k - \Theta^*\|_2^2 \right]. \qquad (44)
\end{aligned}$$

The last inequality follows since by choice, $\eta \geq 1$, and $\alpha_k \leq \alpha_{k-1} \leq \alpha_{k-\tau(\alpha_k),k-1}$. Hence, we have $\alpha_k \leq \eta\alpha_{k-\tau(\alpha_k),k-1}$. Let us now bound the other term in (42).

$$\begin{aligned}
A_2 &= \left| (\Theta_{k-\tau(\alpha_k)} - \Theta^*)^T \mathbb{E}_{k-\tau_k}\left[(A_k - A)(\Theta_k - \Theta_{k-\tau(\alpha_k)}) \right] \right| \\
&\leq \|\Theta_{k-\tau(\alpha_k)} - \Theta^*\|_2 \left\| \mathbb{E}_{k-\tau_k}\left[(A_k - A)\,(\Theta_k - \Theta_{k-\tau(\alpha_k)}) \right] \right\|_2 \\
&\leq 2\eta \|\Theta_{k-\tau(\alpha_k)} - \Theta^*\|_2 \mathbb{E}_{k-\tau_k}\left[\left\| \Theta_k - \Theta_{k-\tau(\alpha_k)} \right\|_2 \right],
\end{aligned}$$

where the last inequality follows from convexity of $\|\cdot\|_2$, sub-multiplicativity of the norm, and the definition of η. Again

using (41),

$$\begin{aligned}
A_2 &\leq 8\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k-\tau(\alpha_k)}-\Theta^*\|_2(\|\Theta_k\|_2+1)\right]\\
&\leq 8\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(\|\Theta_k-\Theta_{k-\tau(\alpha_k)}\|_2+\|\Theta_k-\Theta^*\|_2)(\|\Theta_k\|_2+1)\right]\\
&\overset{(a)}{\leq} 8\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta_k\|_2+\|\Theta_k-\Theta^*\|_2)(\|\Theta_k\|_2+1)\right]\\
&\leq 8\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2+2\|\Theta_k-\Theta^*\|_2)(\|\Theta_k-\Theta^*\|_2+\|\Theta^*\|_2+1)\right]\\
&\leq 16\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2+\|\Theta_k-\Theta^*\|_2)^2\right]\\
&\leq 32\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2)\right], \qquad (45)
\end{aligned}$$

where (a) uses (41).

Combining bounds in (44) and (45) and substituting in (42), we have:

$$\begin{aligned}
&\mathbb{E}_{k-\tau_k}\left[(\Theta_{k-\tau(\alpha_k)}-\Theta^*)^T(A_k\Theta_k+b_k-A\Theta_k-b)\right]\\
&\qquad\leq 16\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2\right]\\
&\qquad\qquad+32\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2)\right]\\
&\qquad\leq 48\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2\right]. \qquad (46)
\end{aligned}$$

Combining the above with (40), and substituting in (39), we have

$$\mathbb{E}_{k-\tau_k}\left[(\Theta_k-\Theta^*)^T(A_k\Theta_k+b_k-A\Theta_k-b)\right]\leq 64\eta^2\alpha_{k-\tau(\alpha_k),k-1}\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2\right].$$

Putting everything together and substituting the bounds in (38), we have

$$\begin{aligned}
&\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k+1}-\Theta^*\|^2-\|\Theta_k-\Theta^*\|_2^2\right]\\
&\qquad\leq 2\alpha_k^2\eta^2\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2+(1+\|\Theta^*\|_2)^2\right]-2\alpha_k\frac{\tilde{\Delta}\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2\right]}{2}\\
&\qquad\qquad+128\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2\right]\\
&\qquad\leq 2\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2+(1+\|\Theta^*\|_2)^2\right]-\alpha_k\tilde{\Delta}\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2\right]\\
&\qquad\qquad+128\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k\mathbb{E}_{k-\tau_k}\left[(1+\|\Theta^*\|_2)^2+\|\Theta_k-\Theta^*\|_2^2\right]\\
&\qquad=\left(130\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k-\alpha_K\tilde{\Delta}\right)\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2\right]\\
&\qquad\qquad+130\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k(1+\|\Theta^*\|_2)^2.
\end{aligned}$$

On rearranging, we have

$$\begin{aligned}
\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k+1}-\Theta^*\|_2^2\right]\leq&\left(1+130\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k-\alpha_k\tilde{\Delta}\right)\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2\right]\\
&+130\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k(1+\|\Theta^*\|_2)^2. \qquad (47)
\end{aligned}$$

From condition on k^*, for $k\geq k^*$,

$$\alpha_{k-\tau(\alpha_k),k-1}\leq\frac{\tilde{\Delta}_2}{260\eta^2},\ \text{ i.e., }\ 260\eta^2\alpha_{k-\tau(\alpha_k),k-1}\alpha_k\leq\alpha_k\tilde{\Delta}_2,$$

and

$$\frac{\alpha_{k-\tau(\alpha_k),k-1}}{\tau(\alpha_k)\alpha_k}\leq 2,\ \text{ i.e., }\ \alpha_{k-\tau(\alpha_k),k-1}\alpha_k\leq 2\alpha_k^2\tau(\alpha_k).$$

The above follows since each of the α_i for $i\in[k-\tau(\alpha_k),k-1]$ are less than $2\alpha_k$.

Using these in (47),

$$\begin{aligned}\mathbb{E}_{k-\tau_k}\left[\|\Theta_{k+1}-\Theta^*\|_2^2\right] &\leq \left(1-\frac{\alpha_k\tilde{\Delta}_2}{2}\right)\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2\right]+260\eta^2\alpha_k^2\tau(\alpha_k)(1+\|\Theta^*\|_2)^2\\ &\leq \left(1-\frac{\alpha_k\tilde{\Delta}_2}{2}\right)\mathbb{E}_{k-\tau_k}\left[\|\Theta_k-\Theta^*\|_2^2\right]+\frac{\xi_2}{2}\alpha_k^2\tau(\alpha_k),\end{aligned}$$

where we used that

$$\frac{\xi_2}{2}\geq 260\eta^2(1+\|\Theta^*\|_2)^2.$$

Now, recursively using this inequality, we have for $k\geq k^*$,

$$\mathbb{E}\big[\|\Theta_k-\Theta^*\|_2^2\big]\leq \mathbb{E}\big[\|\Theta_{k^*}-\Theta^*\|_2^2\big]\prod_{i=k^*}^{k-1}\left(1-\frac{\alpha_t\tilde{\Delta}_2}{2}\right)+\frac{\xi_2}{2}\sum_{i=k^*}^{k-1}\tau(\alpha_i)\alpha_i^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right).$$

Again, by assumption on k^*, $\alpha_{0,k^*-1}\leq\frac{1}{2\eta}$. Using this,

$$\begin{aligned}\mathbb{E}\big[\|\Theta_{k^*}-\Theta^*\|_2^2\big] &\leq \mathbb{E}\big[(\|\Theta_{k^*}-\Theta_0\|_2+\|\Theta_0-\Theta^*\|_2)^2\big]\\ &\overset{(a)}{\leq}(1+\|\Theta_0\|_2+\|\Theta_0-\Theta^*\|_2)^2\\ &\leq(1+2\|\Theta_0\|_2+\|\Theta^*\|_2)^2\\ &=(1+\|\Theta^*\|_2)^2\\ &=\xi_1,\end{aligned}$$

where we used that $\Theta_0=0$. To see (a), consider the following inequalities:

$$\begin{aligned}\|\Theta_{k^*}-\Theta_0\|_2 &= \|(\Theta_{k^*}-\Theta_{k^*-1})+\Theta_{k^*-1}-\Theta_0\|_2\\ &=\|\alpha_{k^*-1}(A_{k^*-1}\Theta_{k^*-1}+b_{k^*-1})+\Theta_{k^*-1}-\Theta_0\|_2\\ &\leq\alpha_{k^*-1}\|A_{k^*-1}\Theta_{k^*-1}+b_{k^*-1}\|+\|\Theta_{k^*-1}-\Theta_0\|_2\\ &\leq\alpha_{k^*-1}\eta(1+\|\Theta_{k^*-1}-\Theta_0\|_2+\|\Theta_0\|_2)+\|\Theta_{k^*-1}-\Theta_0\|_2\\ &=(1+\eta\alpha_{k^*-1})\|\Theta_{k^*-1}-\Theta_0\|_2+\eta\alpha_{k^*-1}(1+\|\Theta_0\|_2)\\ &\leq\prod_{j=1}^{k^*-1}(1+\eta\alpha_j)\|\Theta_1-\Theta_0\|_2+\eta(1+\|\Theta_0\|_2)\sum_{j=1}^{k^*-1}\alpha_j\prod_{i=j+1}^{k^*-1}(1+\eta\alpha_i)\\ &\leq e^{\eta\alpha_{1,k^*-1}}\|\Theta_1-\Theta_0\|_2+\eta(1+\|\Theta_0\|_2)\sum_{j=1}^{k^*-1}\alpha_j e^{\eta\alpha_{1,k^*-1}}\\ &\leq e^{\eta\alpha_{0,k^*-1}}\|\Theta_1-\Theta_0\|_2+\eta(1+\|\Theta_0\|_2)\alpha_{1,k^*-1}e^{\eta\alpha_{0,k^*-1}}\\ &\leq 2\alpha_0\eta(1+\|\Theta_0\|_2)+2\eta(1+\|\Theta_0\|_2)\alpha_{1,k^*-1}\\ &=2\eta(1+\|\Theta_0\|_2)\alpha_{0,k^*-1}\\ &\leq(1+\|\Theta_0\|_2).\end{aligned}$$

This gives the desired bound:

$$\mathbb{E}\big[\|\Theta_k-\Theta^*\|_2^2\big]\leq\xi_1\prod_{i=k^*}^{k-1}\left(1-\frac{\alpha_i\tilde{\Delta}_2}{2}\right)+\frac{\xi_2}{2}\sum_{i=k^*}^{k-1}\tau(\alpha_i)\alpha_i^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right).\tag{48}$$

Theorem C.1(a). Let's now bound the terms in (48), when $\alpha_i = \alpha$. In this case, $k^* = \tau(\alpha)$, and the bound in (48) holds for all $k \geq \tau(\alpha)$. First, observe that

$$\prod_{i=\tau(\alpha)}^{k-1}\left(1-\frac{\alpha\tilde{\Delta}_2}{2}\right) = \left(1-\frac{\alpha\tilde{\Delta}_2}{2}\right)^{k-\tau(\alpha)},$$

and

$$\sum_{i=\tau(\alpha)}^{k-1}\tau(\alpha)\alpha^2\prod_{j=i+1}^{k-1}\left(1-\frac{\tilde{\Delta}_2\alpha}{2}\right) = \alpha^2\tau(\alpha)\sum_{i=\tau(\alpha)}^{k-1}\left(1-\frac{\alpha\tilde{\Delta}_2}{2}\right)^{k-i-1} \leq \alpha^2\tau(\alpha)\sum_{i=0}^{\infty}\left(1-\frac{\alpha\tilde{\Delta}_2}{2}\right)^{i} \leq \frac{2\alpha\tau(\alpha)}{\tilde{\Delta}_2}.$$

Using these, for $T \geq \tau(\alpha)$, we get the following bound:

$$\mathbb{E}\big[\|\Theta_k - \Theta^*\|_2^2\big] \leq \xi_1\left(1-\frac{\alpha\tilde{\Delta}}{2}\right)^{k-\tau(\alpha)} + \frac{\xi_2\alpha\tau(\alpha)}{\tilde{\Delta}_2}. \quad \square$$

Theorem C.1 (b). Let's now bound the terms in (48), when $\alpha_i = \frac{\alpha}{i+h}$. First, observe that

$$\begin{aligned}
\prod_{i=k^*}^{k-1}\left(1-\frac{\alpha_i\tilde{\Delta}_2}{2}\right) &= \prod_{i=k^*}^{k-1}\left(1-\frac{\alpha\tilde{\Delta}_2}{2(i+h)}\right)\\
&\leq \prod_{i=k^*}^{k-1} e^{-\frac{\alpha\tilde{\Delta}_2}{2(i+h)}}\\
&= e^{-\frac{\alpha\tilde{\Delta}_2}{2}\sum_{i=k^*}^{k-1}\frac{1}{i+h}}\\
&\leq e^{-\frac{\alpha\tilde{\Delta}_2}{2}\ln\frac{k+h}{k^*+h}}\\
&= \left(\frac{k^*+h}{k+h}\right)^{\frac{\alpha\tilde{\Delta}_2}{2}}. \qquad (49)
\end{aligned}$$

Next, for the other term, since for $k-1 \geq i \geq k^*$,

$$\tau(\alpha_i) \leq \tau(\alpha_k) \leq L\ln\frac{1}{\alpha_k} = L\left(\ln(k+h) - \ln(\alpha)\right), \qquad (50)$$

we have

$$\sum_{i=k^*}^{k-1}\tau(\alpha_i)\alpha_i^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right) \leq L\ln\left(\frac{k+h}{\alpha}\right)\sum_{i=k^*}^{k-1}\alpha_k^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right). \qquad (51)$$

Here, inequalities in (50) follow from (15). Moreover,

$$\begin{aligned}
\prod_{j=i+1}^{k-1}\left(1-\frac{\tilde{\Delta}_2\alpha}{2(j+h)}\right) &\leq e^{-\frac{\tilde{\Delta}_2\alpha}{2}\sum_{j=i+1}^{k-1}\frac{1}{j+h}}\\
&\leq \left(\frac{i+h+1}{k+h}\right)^{\frac{\tilde{\Delta}_2\alpha}{2}}.
\end{aligned}$$

Then, using in (51),

$$\begin{aligned}\sum_{i=k^*}^{k-1}\tau(\alpha_i)\alpha_i^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right) &\le L\ln\left(\frac{k+h}{\alpha}\right)\sum_{i=k^*}^{k-1}\alpha_i^2\left(\frac{i+h+1}{k+h}\right)^{\frac{\tilde{\Delta}_2\alpha}{2}}\\ &= L\ln\left(\frac{k+h}{\alpha}\right)\sum_{i=k^*}^{k-1}\frac{\alpha^2}{(i+h)^2}\left(\frac{i+h+1}{k+h}\right)^{\frac{\tilde{\Delta}_2\alpha}{2}}\\ &= L\ln\left(\frac{k+h}{\alpha}\right)\frac{\alpha^2}{(k+h)^{\frac{\tilde{\Delta}_2\alpha}{2}}}\sum_{i=k^*}^{k-1}\left(\frac{i+h+1}{i+h}\right)^2(i+h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-2}\\ &\le L\ln\left(\frac{k+h}{\alpha}\right)\frac{4\alpha^2}{(k+h)^{\frac{\tilde{\Delta}_2\alpha}{2}}}\sum_{i=k^*}^{k-1}(i+h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-2}. \end{aligned}\tag{52}$$

Choosing

$$\frac{\alpha\tilde{\Delta}_2}{2}>1,$$

we have

$$\begin{aligned}\sum_{i=k^*}^{k-1}(i+h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-2} &\le \int_{i=0}^{k}(i+h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-2}\,dx\\ &= \frac{1}{\frac{\tilde{\Delta}_2\alpha}{2}-1}\left((k+h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-1}-(h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-1}\right)\\ &\le \frac{1}{\frac{\tilde{\Delta}_2\alpha}{2}-1}(k+h+1)^{\frac{\tilde{\Delta}_2\alpha}{2}-1}.\end{aligned}$$

Substituting back in (52), we get

$$\begin{aligned}\sum_{i=k^*}^{k-1}\tau(\alpha_i)\alpha_i^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right) &\le L\ln\left(\frac{k+h}{\alpha}\right)\left(\frac{4\alpha^2}{\frac{\tilde{\Delta}_2\alpha}{2}-1}\right)\frac{1}{k+h+1}\left(\frac{k+h+1}{k+h}\right)^{\frac{\tilde{\Delta}_2\alpha}{2}}\\ &\le L\ln\left(\frac{k+h}{\alpha}\right)\left(\frac{4\alpha^2}{\frac{\tilde{\Delta}_2\alpha}{2}-1}\right)\frac{e^{\frac{\tilde{\Delta}_2\alpha}{2(k+h)}}}{k+h+1}\\ &= L\ln\left(\frac{k+h}{\alpha}\right)\left(\frac{4\alpha^2}{\frac{\tilde{\Delta}_2\alpha}{2}-1}\right)\frac{e^{\frac{\tilde{\Delta}_2\alpha_k}{2}}}{k+h+1}.\end{aligned}$$

Choosing h so that

$$\frac{\tilde{\Delta}_2\alpha_k}{2}\le\frac{\tilde{\Delta}_2\alpha_0}{2}=\frac{\tilde{\Delta}_2\alpha}{2h}<1,$$

i.e.,

$$\frac{\tilde{\Delta}_2\alpha}{2}<h,$$

we have

$$\sum_{i=k^*}^{k-1}\tau(\alpha_i)\alpha_i^2\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_j\tilde{\Delta}_2}{2}\right)\le L\ln\left(\frac{k+h}{\alpha}\right)\left(\frac{4\alpha^2}{\frac{\tilde{\Delta}_2\alpha}{2}-1}\right)\frac{e}{k+h+1}.\tag{53}$$

Using (49) and (53) in (48), we have

$$\mathbb{E}\left[\|\Theta_k-\Theta^*\|_2^2\right]\le\xi_1\left(\frac{k^*+h}{k+h}\right)^{\frac{\alpha\tilde{\Delta}_2}{2}}+\left(\frac{2\alpha^2\xi_2Le}{\frac{\tilde{\Delta}_2\alpha}{2}-1}\right)\frac{\ln(k+h)-\ln\alpha}{k+h+1}.\quad\square$$

C.5. Proof of Proposition 5.3

Since from Theorem 5.2 the proposed algorithm's estimate converges to $\Phi\theta^*$, the approximation error incurred is $\mathcal{E}_{D_\mu,S_\mu}[\Phi\theta^*]$. Following arguments similar to those in (Tsitsiklis & Van Roy, 1999), it can be shown that

$$\mathcal{E}_{D_\mu,S_\mu}[\Phi\theta^*] \leq \frac{\mathcal{E}}{\sqrt{1-\lambda^2}},$$

where $\lambda \in (0,1)$ is a constant. The bound above is a blow-up of the minimum error possible due to the chosen architecture for approximation.

With this, we get the following approximation error for the variance estimate of the proposed algorithm, for any $c \in \mathbb{R}$.

$$\begin{aligned}
|\kappa_\mu - \kappa^*| &= 2\left|\mathbb{E}_{d_\mu}\left[(r(S,A)-J_\mu)\left(Q^*_\mu(S,A)+c-[\Phi\theta^*](S,A)\right)\right]\right| \\
&= 2\left|\left\langle \mathcal{R}-J_\mu\mathbf{e}, Q^*_\mu + c\mathbf{e} - \Phi\theta^*\right\rangle_{D_\mu}\right| \\
&= 2\left|\left\langle D_\mu^{\frac{1}{2}}\left(\mathcal{R}-J_\mu\mathbf{e}\right), D_\mu^{\frac{1}{2}}\left(Q^*_\mu + c\mathbf{e} - \Phi\theta^*\right)\right\rangle\right| \\
&\leq 2\|D_\mu^{\frac{1}{2}}\left(\mathcal{R}-J_\mu\mathbf{e}\right)\|_2 \, \|D_\mu^{\frac{1}{2}}\left(Q^*_\mu + c\mathbf{e} - \Phi\theta^*\right)\|_2 \\
&\leq 4r_{\max}\|Q^*_\mu + c\mathbf{e} - \Phi\theta^*\|_{D_\mu}.
\end{aligned}$$

Since the above inequality is true for all $c \in \mathbb{R}$, we have that the approximation error for κ_μ is bounded as below:

$$\begin{aligned}
(\kappa_\mu - \kappa^*)^2 &\leq 16r_{\max}^2 \inf_{c\in\mathbb{R}} \|\Phi\theta^* - Q^*_\mu - c\mathbf{e}\|^2_{D_\mu} \\
&= 16r_{\max}^2 \inf_{Q\in S_\mu} \|\Phi\theta^* - Q\|^2_{D_\mu} \\
&= 16r_{\max}^2 \left(\mathcal{E}_{D_\mu,S_\mu}[\Phi\theta^*]\right)^2 \\
&\leq \inf_{\lambda\in[0,1)} \frac{16r_{\max}^2}{1-\lambda^2}\,\mathcal{E}^2. \quad \square
\end{aligned}$$

D. Auxiliary Technical Lemmas

Lemma D.1. *For $b \geq 0$, and $c \geq 0$,*

$$\min_{x\in[0,c]}\left\{ax - b\sqrt{cx-x^2}\right\} = \frac{c}{2}\left(a - \sqrt{a^2+b^2}\right).$$

Proof. Let $f(x) := ax - b\sqrt{cx-x^2}$. Let $f'(x)$ denote the derivative of $f(x)$ and $f''(x)$ denote the corresponding second derivative, both evaluated at x. Then,

$$f'(x) = a - \frac{b(c-2x)}{2\sqrt{cx-x^2}},$$

and $f''(x) > 0$ for $b \geq 0$. Solving for x such that $f'(x) = 0$, we get x_1 defined below as the optimizer.

$$x_1 := \frac{c}{2}\left(1 - \frac{a}{\sqrt{a^2+b^2}}\right),$$

and $f(x_1)$ is the desired optimal value. $\square$

E. Estimating Variance: IID Setting

Consider $X_1, X_2, \ldots,$ independent samples from a distribution with mean 0 and an unknown variance σ^2. Further assume that the sampling distribution has bounded 4^{th} moment. The goal is to estimate σ^2 using these samples.

Since, in this setting, $\mathrm{Var}[X] = \mathbb{E}[X^2]$, let

$$\hat{\sigma}^2_n := \frac{1}{n}\sum_{j=1}^n X_j^2$$

denote the estimate for σ^2 using n samples. Clearly, $\mathbb{E}\big[\hat{\sigma}_n^2\big] = \sigma^2$, i.e., $\hat{\sigma}_n^2$ is an unbiased estimator.

Next, consider the following

$$\begin{aligned}
\mathbb{E}\big[(\hat{\sigma}_n^2 - \sigma^2)^2\big] &= \mathbb{E}\big[(\hat{\sigma}_n^2)^2 + \sigma^4 - 2\sigma^2\hat{\sigma}_n^2\big] \\
&= \mathbb{E}\big[(\hat{\sigma}_n^2)^2\big] - \sigma^4 \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{j=1}^{n} X_j^4 + \sum_{i\neq j} X_i^2 X_j^2\right] - \sigma^4 \\
&= \frac{\mathbb{E}\big[X^4\big]}{n} + \frac{n(n-1)\sigma^4}{n^2} - \sigma^4 \\
&= \frac{1}{n}\left(\mathbb{E}\big[X^4\big] - \sigma^4\right) \\
&= \frac{c_1}{n},
\end{aligned}$$

where $c_1 = \mathbb{E}\big[X^4\big] - (\mathbb{E}\big[X^2\big])^2$. Thus, we have that the mean-squared estimation error in this setting is exactly $O(\frac{1}{n})$. In fact, Cramér Rao lower bound (Nielsen, 2013) for the mean-squared error in estimating σ^2 using an unbiased estimators is $O(1/n)$, establishing that this rate cannot be improved in certain settings.

For $\sigma > 0$, let $\mathcal{N}(0, \sigma^2)$ denote the Gaussian distribution with variance σ^2. If $X \sim \mathcal{N}(0, \sigma^2)$, c_1 equals $3\sigma^2$, and in this case we have that

$$\mathbb{E}\Big[\big(\hat{\sigma}_n^2 - \sigma^2\big)^2\Big] = \frac{3\sigma^2}{n}, \quad \text{for } c > 0.$$