Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability

Sergey Samsonov SVSAMSONOV@HSE.RU
HSE University, Moscow, Russia
Daniil Tiapkin DANIIL.TIAPKIN@POLYTECHNIQUE.EDU
Centre de Mathématiques Appliquées – CNRS – École polytechnique – Institut Polytechnique de Paris, France
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, France
Alexey Naumov ANAUMOV@HSE.RU
HSE University, Moscow, Russia
Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russia
Eric Moulines ERIC.MOULINES@POLYTECHNIQUE.EDU
Centre de Mathématiques Appliquées – CNRS – École polytechnique – Institut Polytechnique de Paris, France
Mohamed bin Zayed University of Artificial Intelligence, UAE

Editors: Shipra Agrawal and Aaron Roth

Abstract

In this paper we consider the problem of obtaining sharp bounds for the performance of temporal difference (TD) methods with linear function approximation for policy evaluation in discounted Markov decision processes. We show that a simple algorithm with a universal and instance-independent step size together with Polyak-Ruppert tail averaging is sufficient to obtain near-optimal variance and bias terms. We also provide the respective sample complexity bounds. Our proof technique is based on refined error bounds for linear stochastic approximation together with the novel stability result for the product of random matrices that arise from the TD-type recurrence.

Keywords: Temporal difference learning, stochastic approximation, Polyak-Ruppert averaging

1. Introduction

This paper aims to provide sharp statistical guarantees for the temporal difference (TD) learning algorithms that use a linear function approximation in the on-policy setting. The TD algorithm (Sutton, 1988; Sutton and Barto, 2018) is one of the most fundamental methods for policy evaluation in reinforcement learning (RL), acknowledged for its simplicity and ease of implementation. Theoretical analysis of TD learning in general state space is usually performed in the setting of *linear function approximation* (Bertsekas and Tsitsiklis, 1996). The asymptotic convergence of TD in such a setting was shown in (Tsitsiklis and Van Roy, 1997). At the same time, the current trend in the field of stochastic approximation is to study non-asymptotic properties of the error, and provide high probability error bounds (Mou et al., 2020; Durmus et al., 2024; Huo et al., 2023). However, many of the existing works (Bhandari et al., 2018; Dalal et al., 2018; Lakshminarayanan and Szepesvari, 2018) characterize the convergence guarantees and sample complexity only in terms of the mean-squared error (MSE). Other works (Korda and La, 2015; Patil et al., 2023) study versions of the TD learning algorithm with projections in order to overcome the crucial problem in the analysis related to the stability of the random matrix products (Guo, 1994; Guo and Ljung, 1995). The latter projections onto the feasible set are usually impractical. Other works provide the high-probability bounds (Li et al., 2024), but with the choice of step sizes relying on the (unknown in practice) instance-dependent quantities, related to the problem design, see Section 3 for details.

Contributions. We enhance the existing p-moment and high probability bounds for the iterates of the TD learning procedure. Towards this aim we follow the framework of the linear stochastic approximation (LSA). Our main contributions are as follows:

- We propose a refined high-probability error bound for TD learning with linear function approximation and Polyak-Ruppert averaging with a *universal and instance-independent* step size. We consider both the generative model assumption and trajectory-wise evaluation based on a sequence of observations forming a Markov chain. However, we show that the variance term of the instance-independent TD learning might be suboptimal with respect to its dependence upon the properties of the feature map.
- In order to obtain our results for the particular setting of TD learning, we provide error bounds for the LSA algorithms by directly assuming the core *exponential stability* of the random matrix product (see assumption A 2 and related discussion). We then present a novel proof of exponential stability specifically for the TD(0) algorithm, which quantifies the speed at which the algorithm forgets its initial error. Our bound is tighter than the previously known results in the literature and serves as a pivotal element in eliminating the need for an additional projection step when addressing the high-order moments of the error (Patil et al., 2023). Conventional proofs for the exponential stability of matrix products often impose an unnecessary restriction on the choice of step size by adjusting it to the *minimal* eigenvalue of the design matrix. This limitation explains the need for projections in (Patil et al., 2023) and the instance-dependent step size in (Li et al., 2024). Our approach allows us to mitigate this drawback.

Related works. The number of contributions to the analysis of TD learning is substantial, and we can not hope to comprehensively cover them all. Significant progress has been made in evaluating the effects of tolerance levels and various parameters on the sampling efficiency of TD learning with linear function approximation (Lakshminarayanan and Szepesvari, 2018; Dalal et al., 2018; Bhandari et al., 2018; Srikant and Ying, 2019). However, the minimax-optimal dependence on the tolerance level has only been established in expectation, see (Li et al., 2023). This paper considered the MSE guarantees. The authors in (Khamaru et al., 2021) considered complexity bounds for TD learning for finite state space in terms of ℓ_∞-norm. A recent paper (Duan and Wainwright, 2023) focuses on multi-step ahead TD learning. Among the closest counterparts to our paper, we must mention the following:

- (Li et al., 2023) establishes lower bounds on the mean squared error (MSE) for policy evaluation problems. They also present bounds on the MSE of the variance-reduced TD learning algorithm, which covers both generative model and Markov sampling methods.
- (Li et al., 2024) provides high-probability bounds and sample complexity for the TD(0) learning algorithm and extends these findings to its off-policy counterpart (TDC) under the i.i.d. sampling assumption. Despite not separating the respective error bounds into the deterministic and stochastic components, the authors in (Li et al., 2024) consider the step size that scales with the minimal eigenvalue of the feature matrix (see **TD** 2 for details). Such scaling is not only a drawback for the practical implementation of the algorithm but also inevitably implies a suboptimal rate of forgetting the initial error.
- (Patil et al., 2023) focuses on determining the bounds of the second moment for TD(0) and high-probability bounds for projected TD(0) iterates. However, the established high-probability bounds require a projection procedure that relies on prior knowledge of the true

parameter norm $\|\theta_\star\|$, which is impractical. Despite this limitation, the study shows that this problem can be resolved using the restart technique.

Non-asymptotic results for TD learning could be also derived from the analysis of the general LSA algorithms (Mou et al., 2020; Durmus et al., 2024). However, the respective error bounds are typically loose in terms of problem-dependent quantities, related to the feature mapping considered in TD with linear function approximation. Detailed discussion is provided after A2.

Notations. For the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ we write $a_n \lesssim b_n$ if there exist an absolute constant $c > 0$, such that $a_n \le c b_n$ for any $n \in \mathbb{N}$. We also write that $a_n = \tilde{\mathcal{O}}(b_n)$, if $a_n \le c(\log n)^\kappa b_n$ for some $\kappa > 0$. For the matrix $A \in \mathbb{R}^{d\times d}$, such that $A = A^\top \succeq 0$, and vector $x \in \mathbb{R}^d$ we define the corresponding A-norm of x as $\|x\|_A = \sqrt{x^\top A x}$. In the present text, the following abbreviations are frequently used: "w.r.t." stands for "with respect to", "i.i.d." stands for "independent and identically distributed".

2. General LSA results

We consider the LSA problem, that is, we aim to solve a linear system $\bar{\mathbf{A}}\theta = \bar{\mathbf{b}}$ with a unique solution $\theta_\star$. We do not have access to $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$ but instead we have access to a sequence of observations $\{(\mathbf{A}(Z_n), \mathbf{b}(Z_n))\}_{n\in\mathbb{N}}$, where $(Z_k)_{k\in\mathbb{N}}$ are noise variables taking values in a measurable space $(\mathsf{Z}, \mathcal{Z})$ and $\mathbf{A}\colon \mathsf{Z} \to \mathbb{R}^{d\times d}$, $\mathbf{b}\colon \mathsf{Z} \to \mathbb{R}^d$ are measurable functions. We consider the setting where $(Z_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common distribution μ satisfying

$$\mathbb{E}_\mu[\mathbf{A}(Z_1)] = \bar{\mathbf{A}} \ , \text{ and } \mathbb{E}_\mu[\mathbf{b}(Z_1)] = \bar{\mathbf{b}} \ .$$

For a fixed step size $\alpha > 0$, burn-in period $n_0 \in \mathbb{N}$, and initialization $\theta_0 \in \mathbb{R}^d$, we consider the sequences of LSA iterates $\{\theta_n\}_{n\in\mathbb{N}}$ and its tail-averaged counterpart $\{\bar{\theta}_{n_0,n}\}_{n\ge n_0+1}$ given by

$$\begin{aligned} \theta_k &= \theta_{k-1} - \alpha\{\mathbf{A}(Z_k)\theta_{k-1} - \mathbf{b}(Z_k)\} \ , \quad k \ge 1, \\ \bar{\theta}_{n_0,n} &= (n-n_0)^{-1}\textstyle\sum_{k=n_0}^{n-1}\theta_k \ , \quad n \ge n_0+1 \ . \end{aligned} \tag{1}$$

Unless explicitly stated, we set $n_0 = n/2$ and write $\bar{\theta}_n$ instead of $\bar{\theta}_{n/2,n}$. The sequence $\{\bar{\theta}_n\}$ corresponds to the Polyak-Ruppert averaged iterates; see (Ruppert, 1988; Polyak and Juditsky, 1992). Using the definition (1) and elementary algebra, we obtain

$$\theta_n - \theta_\star = (\mathrm{I} - \alpha\mathbf{A}(Z_n))(\theta_{n-1} - \theta_\star) - \alpha\varepsilon(Z_n) \ , \tag{2}$$

where the noise variable $\varepsilon(\cdot)$ is defined as

$$\varepsilon(z) = \tilde{\mathbf{A}}(z)\theta_\star - \tilde{\mathbf{b}}(z) \ , \ \tilde{\mathbf{A}}(z) = \mathbf{A}(z) - \bar{\mathbf{A}} \ , \ \tilde{\mathbf{b}}(z) = \mathbf{b}(z) - \bar{\mathbf{b}} \ .$$

The quantity $\varepsilon(\cdot)$ is crucial for our analysis, since in controls the noise level measured *at the solution* $\theta_\star$. Note also that $\varepsilon(Z_i)$ are centered and denote by Σ_ε the covariance matrix of $\varepsilon(Z_i)$, that is,

$$\Sigma_\varepsilon = \mathbb{E}[\varepsilon(Z_1)\varepsilon(Z_1)^\top] \ . \tag{3}$$

Running the recurrence (2), we obtain the error decomposition

$$\theta_n - \theta_\star = \underbrace{\Gamma^{(\alpha)}_{1:n}\{\theta_0 - \theta_\star\}}_{\tilde{\theta}^{(\mathsf{tr})}_n} - \underbrace{\alpha\sum_{j=1}^{n}\Gamma^{(\alpha)}_{j+1:n}\varepsilon(Z_j)}_{\tilde{\theta}^{(\mathsf{fl})}_n} \ . \tag{4}$$

In the above formula we introduced the *product of random matrices*

$$\Gamma_{m:n}^{(\alpha)} = \prod_{i=m}^{n}(\mathrm{I} - \alpha \mathbf{A}(Z_i)) \ , \quad m, n \in \mathbb{N}, \quad m \le n \ ,$$

with the convention $\Gamma_{m:n}^{(\alpha)} = \mathrm{I}$ for $m > n$. The error decomposition (4) is essential for the analysis of LSA algorithms since it allows to split the LSA error into two parts; see, among many others, (Aguech et al., 2000; Durmus et al., 2024). The first, $\tilde{\theta}_n^{(\mathsf{tr})}$, reflects the rate at which the initial error of the procedure is forgotten, and the second, $\tilde{\theta}_n^{(\mathsf{fl})}$, is responsible for the fluctuations of the LSA iterates around the solution $\theta_\star$. The analysis of both $\tilde{\theta}_n^{(\mathsf{fl})}$ and $\tilde{\theta}_n^{(\mathsf{fl})}$ crucially relies on the properties of the matrix product $\Gamma_{m:n}^{(\alpha)}$. In what follows, we present a verifiable set of conditions for the general tail-averaged LSA procedure. In Section 3 we then give a recipe for checking these assumptions for the family of TD algorithms. Our first set of assumptions is classical for LSA:

A1 *Random variables $(Z_k)_{k\in\mathbb{N}}$ are i.i.d. taking values in $(\mathsf{Z}, \mathcal{Z})$ with a distribution μ satisfying $\mathbb{E}[\mathbf{A}(Z_1)] = \bar{\mathbf{A}}$ and $\mathbb{E}[\mathbf{b}(Z_1)] = \bar{\mathbf{b}}$. Moreover,*

$$\|\varepsilon\|_\infty = \sup_{z\in\mathsf{Z}} \|\varepsilon(z)\| < \infty \ , \quad \mathrm{C}_{\mathbf{A}} = \sup_{z\in\mathsf{Z}} \|\mathbf{A}(z)\| \vee \sup_{z\in\mathsf{Z}} \|\tilde{\mathbf{A}}(z)\| < \infty \ .$$

Assumption A1 was considered in several papers, e.g. (Srikant and Ying, 2019; Chen et al., 2020). Almost sure bounds for $\|\mathbf{A}(\cdot)\|$ can be replaced by weaker moment-type bounds following the methods described in (Mou et al., 2020; Durmus et al., 2021b). However, the applications of results with unconstrained noise, especially in the Markov noise setting of Section 5, involves additional technical difficulties. For this reason, we refrain from relaxing the boundedness A1. Now we come to the crucial assumption about the matrix product $\Gamma_{1:n}^{(\alpha)}$. Namely, we define the following family of *exponential stability* assumptions for some $p \in [2, \infty)$:

A2 *(p) There exist $a > 0$, $\varkappa_p > 0$, $\alpha_{p,\infty} > 0$ (depending on p), such that $\alpha_{p,\infty} p \le 1/2$, and for any $\alpha \in (0; \alpha_{p,\infty})$, $u \in \mathbb{R}^d$, $n \in \mathbb{N}$,*

$$\mathbb{E}^{1/p}\left[\|\Gamma_{1:n}^{(\alpha)} u\|^p\right] \le \varkappa_p (1 - \alpha a)^n \|u\| \ . \tag{5}$$

Verifying the exponential stability assumption A2 is crucial for studying properties of both $\tilde{\theta}_n^{(\mathsf{fl})}$ and $\tilde{\theta}_n^{(\mathsf{tr})}$, see e.g. (Guo and Ljung, 1995; Priouret and Veretenikov, 1998). For TD algorithms, exponential stability has been verified in Patil et al. (2023) (for $p = 2$) and (Li et al., 2024) (for arbitrary $p > 2$, but with suboptimal $\alpha_{p,\infty}$, see discussion in Section 3). Note that A2 can be verified under the classical stability conditions for linear systems. In particular, it is enough to assume A1 and additionally assume that the system matrix $-\bar{\mathbf{A}}$ is Hurwitz. The Hurwitzness of $-\bar{\mathbf{A}}$ is a necessary and sufficient condition for the exponential stability of the continuous-time ODE system $\dot{\theta}_t = \bar{\mathbf{A}}\theta_t$, see e.g. (Jacob and Zwart, 2012). Therefore, it is a standard condition for the exponential speed of forgetting the initial error $\tilde{\theta}_n^{(\mathsf{tr})}$, see (Mou et al., 2020; Durmus et al., 2021a). However, such an assumption allows to prove a contraction

$$\|\mathrm{I} - \alpha\bar{\mathbf{A}}\|_{\mathrm{Q}}^2 \le 1 - \alpha\tilde{a} \tag{6}$$

only in specific matrix Q-norm, associated with the solution of the Lyapunov equation

$$\bar{\mathbf{A}}^\top \mathrm{Q} + \mathrm{Q}\,\bar{\mathbf{A}} = -P \ .$$

Here the choice of the matrix $P = P^\top \succ 0$ is an additional degree of freedom, with the default choice being $P = \mathrm{I}$. In this case the factor $\tilde{a}$ in (6) might be overly small, yielding suboptimal bias forgetting rate. Detailed discussion is provided in Appendix B.4. In this paper we suggest a different view on the problem: we assume A2 directly and then verify it for the particular example of TD learning with linear function approximation.

2.1. Refined LSA results with i.i.d. noise

In this section we provide general results for the tail-averaged LSA iterates, which can be viewed as simplified versions of (Durmus et al., 2024, Theorem 2) and (Mou et al., 2020, Theorem 3) with explicit dependence on the instance-dependent quantities, such as the contraction rate a. First, we give an elementary statement for the mean square error.

Theorem 1 *Assume A1 and A2(2). Then for any $n \geq 2$, $\alpha \in (0; \alpha_{2,\infty}]$, and $\theta_0 \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}^{1/2}[\|\bar{\mathbf{A}}(\bar{\theta}_n - \theta_\star)\|^2] \lesssim \frac{\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{n^{1/2}} \left(1 + \frac{\varkappa_2 \, \mathrm{C}_{\mathbf{A}} \sqrt{\alpha}}{\sqrt{a}}\right) + \frac{\varkappa_2 \sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{\sqrt{\alpha a n}} + \varkappa_2 (1 - \alpha a)^{n/2} \left(\frac{1}{\alpha n} + \frac{\mathrm{C}_{\mathbf{A}}}{\sqrt{\alpha a n}}\right) \|\theta_0 - \theta_\star\| \, . \quad (7)$$

Proof sketch. The proof relies on the summation by parts applied to the LSA error (2). This approach was previously applied in (Mou et al., 2020; Durmus et al., 2024), and yields

$$\bar{\mathbf{A}} \left(\bar{\theta}_n - \theta_\star\right) = 2(\alpha n)^{-1}(\theta_{n/2} - \theta_n) - 2n^{-1} \textstyle\sum_{t=n/2}^{n-1} e\left(\theta_t, Z_{t+1}\right) \, , \quad (8)$$

where we have defined $e(\theta, z) = \tilde{\mathbf{A}}(z)\theta - \tilde{\mathbf{b}}(z) = \varepsilon(z) + \tilde{\mathbf{A}}(z)(\theta - \theta_\star)$. This transform justifies why it is convenient to state the bounds in terms of $\|\bar{\mathbf{A}}(\bar{\theta}_n - \theta_\star)\|$. The rest of the proof follows from the martingale structure of the term $\sum_{t=n/2}^{n-1} e\left(\theta_t, Z_{t+1}\right)$ w.r.t. filtration $\mathcal{F}_t = \sigma(Z_s, s \leq t)$. We also need to show the last iterate error bound $\mathbb{E}^{1/2}[\|\theta_n - \theta_\star\|^2] = \tilde{\mathcal{O}}(\sqrt{\alpha})$, a standard result that was previously obtained in many papers on LSA, see e.g. (Dalal et al., 2018; Bhandari et al., 2018). This explains the factor $1 + \mathcal{O}(\sqrt{\alpha})$, which affects the leading term in (7). We provide the complete proof in Appendix A, see Theorem 7. □

Now we provide a p-moment bound. We assume that A2(ℓ) is satisfied for any $\ell \geq 2$, however, similar results could be obtained if A2(p) holds only for a fixed parameter $2 \leq p < \infty$.

Theorem 2 *Suppose that assumptions A1 and A2(ℓ) hold for any $\ell \geq 2$. Then, for any $n \in \mathbb{N}$, $p \geq 2$, $\alpha \in [0, \alpha_{p+\log n,\infty})$, $\theta_0 \in \mathbb{R}^d$, we have*

$$\mathbb{E}^{1/p}[\|\bar{\mathbf{A}}(\bar{\theta}_n - \theta_\star)\|^p] \lesssim \frac{p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{n^{1/2}}\left(1 + \mathsf{R}_1(\alpha)\right) + \frac{\varkappa_{p+\log n}(1 + \mathrm{C}_{\mathbf{A}})p\|\varepsilon\|_\infty}{n} + \frac{p\varkappa_{p+\log n}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{n\sqrt{a}}\left(1 + \frac{1}{\sqrt{\alpha p}}\right) + \varkappa_{p+\log n}(1 - \alpha a)^{n/2}\left(\frac{1}{\alpha n} + \frac{p\,\mathrm{C}_{\mathbf{A}}}{\sqrt{\alpha a n}}\right)\|\theta_0 - \theta_\star\| \, , \quad (9)$$

where the term $\mathsf{R}_1(\alpha)$ is given by

$$\mathsf{R}_1(\alpha) = \frac{\varkappa_{p+\log n}\sqrt{\alpha p}\,\mathrm{C}_{\mathbf{A}}}{\sqrt{a}} + \frac{\varkappa_{p+\log n}\,\mathrm{C}_{\mathbf{A}}\,\alpha p\|\varepsilon\|_\infty}{\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}} \, .$$

Table 1: Summary of error bounds for TD(0) algorithm with linear functional approximation.

Paper	Algorithm type	step size schedule	Universal step size	Markovian data	High-order bounds	Not require projections
(Bhandari et al., 2018)[(1)]	Polyak-Ruppert	$1/\sqrt{n}$	✓	✓	✗	✗
(Dalal et al., 2018)[(2)]	Last iterate	$1/k^{\varkappa}$	✓	✗	✓	✓
(Lakshminarayanan and Szepesvari, 2018)	Polyak-Ruppert	constant α	✓	✗	✗	✓
(Patil et al., 2023)[(3)]	Polyak-Ruppert	constant α	✓	✓	✓	✗
(Li et al., 2024)	Polyak-Ruppert	constant α	✗	✗	✓	✓
This paper	Polyak-Ruppert	constant α	✓	✓	✓	✓

[(1)] (Bhandari et al., 2018) considers constant step size $\alpha = 1/\sqrt{n}$ with n being total number of iterations and provide suboptimal MSE bound of order $\tilde{\mathcal{O}}(1/\sqrt{n})$; [(2)] (Dalal et al., 2018) uses last iterate and decreasing step size schedule with $\alpha_k = 1/k^{\varkappa}$. Hence, the corresponding bias forgetting rate is sublinear, and the n-step MSE is of order $\tilde{\mathcal{O}}(1/n^{\kappa})$; [(3)] (Patil et al., 2023) uses projections in order to prove the concentration bounds, moreover, the definition of the projection set involves unknown parameter $\theta_\star$.

Proof sketch. We use the same key decomposition (8) and utilize the martingale structure of $\sum_{t=n/2}^{n-1} e\left(\theta_t, Z_{t+1}\right)$ with Rosenthal's inequality (Pinelis, 1994). This technique requires to handle the (remainder w.r.t. n) term $\mathbb{E}^{1/p}[\max_t \|\tilde{\mathbf{A}}_{t+1}(\theta_t - \theta_\star)\|^p]$, which scales with n as $n^{1/q}$ for any $q \geq p$ and $\alpha \in [0, \alpha_{q,\infty})$. This dependence is removed by setting $q = p + \log(n)$, and $\alpha \in [0, \alpha_{p+\log n,\infty})$. Complete proof is given in Appendix A-Theorem 9. □

The closest counterparts of Theorem 2, (Durmus et al., 2024, Theorem 2), and (Mou et al., 2020, Theorem 3 and 4), are less explicit in terms of dependence of the error terms upon the contraction rate a. Note that for a general SA problem the constant $\varkappa_{p+\log n}$ above might scale polynomially with d, see (Huang et al., 2021). Yet in particular applications $\varkappa_{p+\log n}$ might be *dimension-free* and even independent from p, as we show in Section 3. The bound given in Theorem 2 highlights a remarkable property: the leading term of (9) contains an additional multiplicative factor of

$$1 + \varkappa_{p+\log n}\sqrt{\alpha p}\,\mathrm{C}_{\mathbf{A}}\,/\sqrt{a} + \mathcal{O}(\alpha)\ .$$

If α is chosen so that the ratio $\alpha p/a = o(1)$, we achieve the 'optimal' sub-Gaussian leading term

$$p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}/n^{1/2}\ . \tag{10}$$

Optimality of the presented leading term is discussed in (Fort, 2015; Mou et al., 2020). This is consistent with the findings from decreasing step size, which have ensured the attractiveness of Polyak-Ruppert algorithms, see e.g. (Bhandari et al., 2018). On the other hand, in case of *instance-independent* choice of step size α it is possible that the ratio $\alpha p/a$ is not small. In such a scenario the dominant term in (9) could far exceed the optimum sub-Gaussian leading term (10). Recent studies addressing constant step size SA schemes (Durmus et al., 2024; Mou et al., 2020) circumvent this problem by adjusting the SA step, α, relative to the time horizon n as $\alpha = O(n^{-\kappa})$ for some $\kappa \in (0, 1]$. However, this approach may result in a slower reduction of the initial error $\|\theta_0 - \theta_\star\|$ and suboptimal instance-dependent second-order terms, a phenomenon observed in Khamaru et al. (2021) in case of TD learning.

3. TD learning under i.i.d. noise

In this section we apply results of Section 2 to the TD learning procedure. Namely, we consider a problem of estimating a value of the policy π in a discounted MDP (Markov Decision Process)

given by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ stand, respectively, for state and action spaces, and $\gamma \in (0,1)$ is a discount factor. We assume that $\mathcal{S}$ is a complete metric space equipped with a metric $\mathrm{d}_{\mathcal{S}}$ and Borel σ-algebra $\mathcal{B}(\mathcal{S})$. P stands for the transition kernel $P(B|s,a)$, which determines the probability of moving from state s to a Borel set $B \in \mathcal{B}(\mathcal{S})$ when action a is performed. For simplicity, the reward function $r\colon \mathcal{S} \times \mathcal{A} \to [0,1]$ is assumed to be deterministic. The policy $\pi(\cdot|s)$ is a distribution over the action space $\mathcal{A}$ corresponding to the agent's action preferences in state $s \in \mathcal{S}$. We aim to estimate the agent's *value function*

$$V^{\pi}(s) = \mathbb{E}[\textstyle\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) | s_0 = s] \ ,$$

where $a_k \sim \pi(\cdot|s_k)$, and $s_{k+1} \sim P(\cdot|s_k, a_k)$, for any $k \in \mathbb{N}$. We define the transition kernel

$$P_{\pi}(B|s) = \textstyle\int_{\mathcal{A}} P(B|s,a)\pi(\mathrm{d}a|s) \ , \tag{11}$$

which corresponds to the 1-step transition probability from state s to a set $B \in \mathcal{B}(\mathcal{S})$. The state space is arbitrary: $\mathcal{S}$ may be finite, but with $|\mathcal{S}| \gg 1$, or $\mathcal{S} \subset \mathbb{R}^D$ may be uncountable. In this setting, it is a common option to consider the *linear function approximation* of the value function $V^{\pi}(s)$, defined for $s \in \mathcal{S}$, $\theta \in \mathbb{R}^d$, and feature mapping $\varphi\colon \mathcal{S} \to \mathbb{R}^d$ as

$$V_{\theta}^{\pi}(s) = \varphi^{\top}(s)\theta \ .$$

Here d is the dimension of the feature space. We consider $V_{\theta}^{\pi}(s)$ as an approximation to the true value function $V^{\pi}(s)$, and our goal is to find a parameter $\theta_{\star}$ that defines the best linear approximation of V^{π} (Tsitsiklis and Van Roy, 1997). To properly define what it means, we introduce some notations, following (Li et al., 2024). We denote by μ the invariant distribution over the state space $\mathcal{S}$ induced by the transition kernel $P_{\pi}(\cdot|s)$ in (11). Then we define $\theta_{\star}$ as a solution

$$\theta_{\star} = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mu}\big[(V_{\theta}^{\pi}(s) - V^{\pi}(s))^2\big] \ . \tag{12}$$

We define the *design matrix* Σ_{φ} as

$$\Sigma_{\varphi} = \mathbb{E}_{\mu}[\varphi(s)\varphi(s)^{\top}] \in \mathbb{R}^{d \times d} \ .$$

In the following, we are interested in minimizing the following distance between $\theta \in \mathbb{R}^d$ and $\theta_{\star}$:

$$\|\theta - \theta_{\star}\|_{\Sigma_{\varphi}} = \mathbb{E}_{\mu}^{1/2}\left[(V_{\theta}^{\pi}(s) - V_{\theta_{\star}}^{\pi}(s))^2\right] \ .$$

For the estimator $\hat{\theta}$ of $\theta_{\star}$, our primary concern is to control the error $\|\hat{\theta} - \theta_{\star}\|_{\Sigma_{\varphi}}$ in two ways: firstly, by controlling its second moment $\mathbb{E}[\|\hat{\theta} - \theta_{\star}\|_{\Sigma_{\varphi}}^2]$, and secondly, by giving high-probability bound; available results are summarized in Table 1. We consider the following assumptions on the generative mechanism and on the feature mapping $\varphi(\cdot)$:

TD 1 *Tuples (s, a, s') are generated i.i.d.with $s \sim \mu$, $a \sim \pi(\cdot|s)$, $s' \sim P(\cdot|s,a)$.*

TD 2 *Matrix Σ_{φ} is non-degenerate with the minimal eigenvalue $\lambda_{\min} \equiv \lambda_{\min}(\Sigma_{\varphi})$. Moreover, the feature mapping $\varphi(\cdot)$ satisfies $\sup_{s \in \mathcal{S}} \|\varphi(s)\| \leq 1$.*

Algorithm 1: Temporal difference learning TD(0)
Input : feature mapping $\varphi(\cdot) : \mathcal{S} \to \mathbb{R}^d$, step size α, number of iterations n, behavioral policy π;
for $k = 1, \dots, n$ **do**
  Receive tuple (s_k, a_k, s'_k) following **TD** 1;
  Compute update $\theta_k = \theta_{k-1} - \alpha(\mathbf{A}_k\theta_{k-1} - \mathbf{b}_k)$ based on $\mathbf{A}_k$, $\mathbf{b}_k$ from (15)
end
Output: tail-averaged estimate $\bar{\theta}_n = (2/n)\sum_{k=n/2+1}^{n}\theta_k$;
value function estimate $V^{\pi}_{\bar{\theta}_n}(s) = \varphi^\top(s)\,\bar{\theta}_n$

The generative model assumption **TD** 1 is used in many previous works; see, e.g. (Dalal et al., 2018; Li et al., 2024; Patil et al., 2023). In Section 5 we generalize this assumption to more realistic setting of on-policy evaluation over a single trajectory, where the induced LSA noise is Markovian.

In the setting of linear functional approximation the problem of estimating $V^\pi(s)$ reduces to the problem of estimating $\theta_\star \in \mathbb{R}^d$, which can be done via the LSA procedure. It is known (see e.g. (Tsitsiklis and Van Roy, 1997)), that optimal (in a sense of (12)) parameter $\theta_\star$ is a solution to the deterministic system$\bar{\mathbf{A}}\theta_\star = \bar{\mathbf{b}}$, where

$$\begin{aligned}\bar{\mathbf{A}} &= \mathbb{E}_{s\sim\mu, s'\sim P_\pi(\cdot|s)}[\phi(s)\{\phi(s) - \gamma\phi(s')\}^\top] \\ \bar{\mathbf{b}} &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot|s)}[\phi(s)r(s,a)]\ .\end{aligned} \tag{13}$$

In order to write the instance of the LSA algorithm for the system (13), we introduce the k-th step randomness $Z_k = (s_k, a_k, s'_k)$. With slight abuse of notation, we write $\mathbf{A}_k$ instead of $\mathbf{A}(Z_k)$, and $\mathbf{b}_k$ instead of $\mathbf{b}(Z_k)$. Then the corresponding LSA update equation with step size α writes as

$$\theta_k = \theta_{k-1} - \alpha(\mathbf{A}_k\theta_{k-1} - \mathbf{b}_k)\ , \tag{14}$$

where $\mathbf{A}_k$ and $\mathbf{b}_k$ are given by

$$\begin{aligned}\mathbf{A}_k &= \phi(s_k)\{\phi(s_k) - \gamma\phi(s'_k)\}^\top\ , \\ \mathbf{b}_k &= \phi(s_k)r(s_k, a_k)\ .\end{aligned} \tag{15}$$

We provide the corresponding pseudocode in Algorithm 1. Under listed assumptions, we are able to check A1 and A2(p). We first establish that A1 holds.

Lemma 1. *Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14) under **TD** 1 and **TD** 2. Then this update scheme satisfies assumption A1 with*

$$\mathrm{C}_{\mathbf{A}} = 2(1+\gamma)\ , \quad \|\varepsilon\|_\infty = 2(1+\gamma)(\|\theta_\star\| + 1)\ , \quad \mathrm{Tr}(\Sigma_\varepsilon) \le 2(1+\gamma)^2(\|\theta_\star\|^2_{\Sigma_\varphi} + 1)\ .$$

An elementary proof is given in Appendix B. Checking A2(p) is a more delicate issue. In particular, it is crucial to determine a tight bound on the stability threshold $\alpha_{p,\infty}$. (Patil et al., 2023) contains an instance-independent bound on the maximum step size, which scales only by a factor $1-\gamma$, for the case of 2-nd moment stability. Higher-order moments are analyzed using a modification of TD, with an additional projection. The counterpart of the exponential stability A 2(p) is implicitly obtained in (Li et al., 2024), but in this work the stability bound scales with $\lambda_{\min}$, which is unavailable in practice. To the best of our knowledge, **we provide the first instance-independent stability bound for the TD (0) algorithm beyond the 2-nd moment**:

Lemma 2. *Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14) under **TD** 1 and **TD** 2. Then this update scheme satisfies assumption A2(p) with*

$$a = (1-\gamma)\lambda_{\min}/2\ ,\ \varkappa_p = 1\ ,\ \alpha_{p,\infty} = (1-\gamma)/(128p)\ . \tag{16}$$

Proof of Lemma 2 is provided in Appendix B.2. Note that, in strong contrast with this result, leveraging the matrix stability argument of (Huang et al., 2021) and (Durmus et al., 2021a) yield an instance-dependent stability threshold

$$\alpha_{p,\infty} = (1-\gamma)\lambda_{\min}/(c_0 p) \tag{17}$$

for some absolute constant $c_0 > 0$. A detailed derivation of the bound (17) can be found in Appendix B.4. The same order of magnitude of the step size is predicted in (Li et al., 2024, Theorem 1). Thus, with the result Lemma 2, we can prove the convergence of TD(0) for larger step sizes.

Now we are ready to adapt the conclusions of Section 2.1 to TD learning. To represent the bounds in terms of $\|\cdot\|_{\Sigma_\varphi}$ rather than the norm associated with the system matrix $\bar{\mathbf{A}}$, we can use the lower bound

$$\|\bar{\mathbf{A}}(\bar{\theta}_n - \theta_\star)\|^2 \geq (1-\gamma)^2\lambda_{\min}\|\bar{\theta}_n - \theta_\star\|^2_{\Sigma_\varphi} .$$

The proof of this bound is provided in Lemma 7, and closely follows the idea of (Li et al., 2024, Lemma 5). We begin with bounding the 2-nd moment of the error, and immediately reformulate this result as a sample complexity bound .

Theorem 3 *Assume **TD** 1 and **TD** 2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $n \geq 2$, $\alpha \in \left(0; \frac{1-\gamma}{256}\right]$, and $\theta_0 \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}^{1/2}[\|\bar{\theta}_n - \theta_\star\|^2_{\Sigma_\varphi}] \lesssim \frac{\|\theta_\star\|_{\Sigma_\varphi}+1}{\sqrt{\lambda_{\min} n}(1-\gamma)}\Big(1 + \frac{\sqrt{\alpha}}{\sqrt{(1-\gamma)\lambda_{\min}}}\Big) + \frac{\|\theta_\star\|_{\Sigma_\varphi}+1}{\sqrt{\alpha}(1-\gamma)^{3/2}\lambda_{\min} n}$$
$$+ f_1(\alpha, \lambda_{\min}, n)\Big(1 - \frac{\alpha(1-\gamma)\lambda_{\min}}{2}\Big)^{n/2}\|\theta_0 - \theta_\star\| ,$$

where $f_1(\alpha, \lambda_{\min}, n)$ is a polynomial function in $1/\alpha, 1/\lambda_{\min}, n$ specified in Appendix B.3-(31).

Corollary 1. *Under the assumptions of Theorem 3, to achieve $\mathbb{E}[\|\bar{\theta}_n - \theta_\star\|^2_{\Sigma_\varphi}] \leq \varepsilon^2$ it is enough to use*

$$\tilde{\mathcal{O}}\Bigg(\underbrace{\frac{\|\theta_\star\|^2_{\Sigma_\varphi}+1}{(1-\gamma)^2\lambda_{\min}\varepsilon^2}\left(1 + \frac{\alpha}{\lambda_{\min}(1-\gamma)}\right) + \mathsf{R}_1(1/\varepsilon)}_{\textit{variance term}} + \underbrace{\frac{1}{\alpha\lambda_{\min}(1-\gamma)} \cdot \log\frac{\|\theta_0 - \theta_\star\|}{\varepsilon}}_{\textit{initial error}}\Bigg) .$$

TD(0) updates, where $\mathsf{R}_1(1/\varepsilon) = \frac{\|\theta_\star\|_{\Sigma_\varphi}+1}{\sqrt{\alpha}(1-\gamma)^{3/2}\lambda_{\min}\varepsilon}$.

Comparison to the robust SA approach. Note that the leading term of the bound in Theorem 3 includes factors of $1/\lambda_{\min}$. This dependence is generally unavoidable if one aims to obtain the MSE bound for $\mathbb{E}[\|\bar{\theta}_n - \theta_\star\|^2_{\Sigma_\varphi}]$ that scales as $1/n$. This is due to the fact that the corresponding asymptotic covariance matrix from the central limit theorem (see e.g., (Fort, 2015)) for $\sqrt{n}(\bar{\theta}_n - \theta_\star)$ could scale with $\lambda^{-1}_{\min}$. More details on the asymptotically minimax covariance bounds are provided in Section 4. In contrast, within the basin of robust stochastic approximation (RSA, (Nemirovski et al., 2009)), a convergence rate for $\mathbb{E}[\|\bar{\theta}_n - \theta_\star\|^2_{\Sigma_\varphi}]$ of order $\mathcal{O}(1/\sqrt{n})$ can be derived with the instance-independent choice of step size. Importantly, this rate is not affected by a worst-case factor of $\lambda^{-1}_{\min}$. This result was obtained for the TD algorithm in (Bhandari et al., 2018, Theorem 2).

Discussion and comparison. Optimizing the bound of Corollary 1 with respect to the step size α is problematic. Taking the largest possible step size $\alpha \simeq 1-\gamma$ from (16) yields the number of steps to reduce deterministic error of order

$$\tilde{\mathcal{O}}\Big(\frac{1}{(1-\gamma)^2\lambda_{\min}} \cdot \log \frac{\|\theta_0 - \theta_\star\|}{\varepsilon}\Big) ,$$

which was previously reported by (Patil et al., 2023). However, this choice of step size results in the overall sample complexity in Corollary 1 being at least

$$\tilde{\mathcal{O}}\Big(\frac{1}{(1-\gamma)^2\lambda_{\min}} \cdot \log \frac{\|\theta_0 - \theta_\star\|}{\varepsilon} + \frac{1 + \|\theta_\star\|^2_{\Sigma_\varphi}}{(1-\gamma)^2\lambda^2_{\min}\varepsilon^2}\Big) .$$

The $1/\varepsilon^2$ component of this bound is by a factor of $\lambda^{-1}_{\min}$ larger than the one obtained in (Li et al., 2024), albeit it is agrees with the bounds of (Patil et al., 2023, Theorem 1). The reason is that the latter paper uses instance-independent step size $\alpha \simeq (1-\gamma)$, while (Li et al., 2024) adjusts step size with (unknown in practice) quantity $\lambda_{\min}$ as $\alpha^{(\text{small})} \simeq (1-\gamma)\lambda_{\min}$. This choice allows to improve the variance component in Corollary 1, but forgetting the bias would require at least

$$\tilde{\mathcal{O}}\Big(\frac{1}{(1-\gamma)^2\lambda^2_{\min}} \cdot \log \frac{\|\theta_0 - \theta_\star\|}{\varepsilon}\Big)$$

iterations of Algorithm 1. Moreover, the remainder term $\mathsf{R}_1(1/\varepsilon)$ will scale as $(1-\gamma)^{-2}\lambda^{-3/2}_{\min}$. The same phenomenon can be traced in (Li et al., 2024, Theorem 1), albeit the authors do not separate the bias and variance components of the error and assume that the procedure starts at $\theta_0 = 0$. This dilemma is resolved in Li et al. (2023), but for a variance-reduced version of TD learning algorithm.

Instantiating Theorem 2 for TD(0), we can provide the bound on $\mathbb{E}^{1/p}[\|\bar{\theta}_n - \theta_\star\|^p_{\Sigma_\varphi}]$ for $p \geq 2$. For completeness, this result is stated in Appendix B.3. With Markov's inequality applied with $p = \log(1/\delta)$, we can translate it into the sample complexity bound. The corresponding deviation bounds for $\|\bar{\theta}_n - \theta_\star\|_{\Sigma_\varphi}$ are provided in appendix, see Appendix B.3-Corollary 4.

Theorem 4 *Fix $\varepsilon > 0$, $\delta > 0$, assume **TD** 1 and **TD** 2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $n \geq 2$, and step size*

$$\alpha \in \big(0; \frac{1-\gamma}{128\log(n/\delta)}\big]$$

to achieve error $\|(\bar{\theta}_n - \theta_\star)\|_{\Sigma_\varphi} \leq \varepsilon$ with probability at least $1-\delta$ it takes

$$\tilde{\mathcal{O}}\Bigg(\frac{(\|\theta_\star\|^2_{\Sigma_\varphi} + 1)\log(1/\delta)}{(1-\gamma)^2\lambda_{\min}\varepsilon^2}\Big(1 + \frac{\alpha\log(1/\delta)}{(1-\gamma)\lambda_{\min}}\Big) + \mathsf{R}_2(1/\varepsilon,\delta) + \frac{\log\frac{\|\theta_0-\theta_\star\|}{\varepsilon}}{\alpha\lambda_{\min}(1-\gamma)}\Bigg) \tag{18}$$

TD(0) updates, where $\mathsf{R}_2(1/\varepsilon,\delta) = \frac{(\|\theta_\star\|_{\Sigma_\varphi}+1)\log(1/\delta)}{\sqrt{\alpha}(1-\gamma)^{3/2}\lambda_{\min}\varepsilon}$.

Discussion and comparison. Note that in Theorem 4 the symbol $\tilde{\mathcal{O}}$ hides logarithmic dependencies in $\lambda_{\min}, 1-\gamma$, and n, but not in $1/\delta$. Again the direct optimization of the bound Theorem 4 w.r.t. α yield to the same dilemma as in case of 2-nd moment. The stochastic part of the complexity

bound (18) scales inversely proportional to $\lambda_{\min}^2$, which is worse than the scaling of the deterministic component of the error. At the same time, choosing the smaller step size

$$\alpha = \frac{(1-\gamma)\lambda_{\min}(\Sigma_\varphi)}{128(p+\log n)} \,, \tag{19}$$

we retrieve the leading variance term of deviation bound (Li et al., 2024, Theorem 1), while improving the second-order term in $\lambda_{\min}$. Indeed, (Li et al., 2024, Theorem 1) yields the high-probability bound of order

$$\tilde{\mathcal{O}}\left(\frac{(\|\theta_\star\|_{\Sigma_\varphi}^2+1)\log(d/\delta)}{(1-\gamma)^2\lambda_{\min}\varepsilon^2} + \frac{(1+\|\theta_\star\|_{\Sigma_\varphi})\log(nd/\delta)}{(1-\gamma)^2\lambda_{\min}^2\varepsilon}\right)$$

in order to achieve $\|\bar{\theta}_n - \theta_\star\|_{\Sigma_\varphi} \le \varepsilon$ with probability at least $1-\delta$. This result is achieved for the step size α which scales similarly to (19). Also, compared to (Li et al., 2024), we obtain a clear separation between the deterministic and stochastic parts of the error, and remove the explicit dependence upon the feature dimension d. Note, however, that the dependence upon d is hidden implicitly inside $\lambda_{\min}$.

4. On optimality of TD(0) for i.i.d. sampling scheme

In this section we present a version of Theorem 3 with a leading variance term consistent with the minimax lower bound due to (Li et al., 2023, Proposition 1). We first write the TD(0) noise covariance matrix

$$\Sigma_\varepsilon^{(TD)} = \mathbb{E}[\big((\phi(s_k)-\gamma\phi(s_k'))^\top\theta_\star - r_k\big)^2\phi(s_k)\phi(s_k)^\top] \,,$$

which corresponds to the general LSA noise covariance matrix Σ_ε defined in (3). We also define the transformed covariance matrix

$$\Sigma_\varepsilon^{(opt)} = \Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon^{(TD)}\bar{\mathbf{A}}^{-T}\Sigma_\varphi^{1/2} \,,$$

which corresponds to the covariance of modified noise variables $\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-1}\varepsilon$. Now let us introduce the counterpart of Theorem 3 with the modified leading (w.r.t. the sample size n) term.

Theorem 5 *Assume **TD** 1 and **TD** 2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $p \ge 2$, $n \ge 2$, $\alpha \in \big(0; \frac{1-\gamma}{256}\big]$, it holds that*

$$\begin{aligned}\mathbb{E}^{1/2}[\|\bar{\theta}_n-\theta_\star\|_{\Sigma_\varphi}^2] \lesssim \frac{\sqrt{\mathrm{Tr}(\Sigma_\varepsilon^{(opt)})}}{n^{1/2}} + \underbrace{\frac{1+\|\theta_\star\|_{\Sigma_\varphi}}{(1-\gamma)^{3/2}\lambda_{\min}n^{1/2}}\left(\frac{1}{\sqrt{\alpha n}}+\sqrt{\alpha}\right)}_{\mathsf{R}_3(n,\lambda_{\min})} \\ + f_2(\alpha,\lambda_{\min},n)\big(1-\alpha(1-\gamma)\lambda_{\min}\big)^{n/2}\|\theta_0-\theta_\star\| \,,\end{aligned} \tag{20}$$

where $f_2(\alpha,\lambda_{\min},n)$ is a polynomial in $1/\alpha, 1/\lambda_{\min}, n$ specified in Appendix C-(45).

The proof is postponed to Appendix C, along with the analogous p-th moment bound. We highlight the fact that the leading term of (20) scales with the quantity $\mathrm{Tr}(\Sigma_\varepsilon^{(opt)})$ corresponding to the instance optimal variance given in (Li et al., 2023, Section 2) and (Mou et al., 2020). At the same time, with simple algebraic manipulations one can prove an upper bound

$$\mathrm{Tr}(\Sigma_\varepsilon^{(opt)}) \le \frac{\|\theta_\star\|_{\Sigma_\varphi}^2 + 1}{(1-\gamma)^2 \lambda_{\min}} ,$$

thus recovering the result obtained in Theorem 3 before. However, the bound of (20) contains also a term $\mathsf{R}_3(n, \lambda_{\min})$, which scales directly with $\lambda_{\min}^{-1}$. Moreover, even setting α as $n^{-\varkappa}$, $\varkappa \in (0,1)$ would require to use large sample size n in order that the optimal noise term $(\mathrm{Tr}(\Sigma_\varepsilon^{(opt)})/n)^{1/2}$ starts to dominate. This is on par with empirical evaluation from Khamaru et al. (2021). Now we reformulate Theorem 5 as a sample complexity bound.

Corollary 2. *Under the assumptions of Theorem 5, to achieve the weighted MSE* $\mathbb{E}[\|\bar\theta_n - \theta_\star\|_{\Sigma_\varphi}^2] \le \varepsilon^2$ *requires*

$$\tilde{\mathcal{O}}\Big(\underbrace{\frac{\log \frac{\|\theta_0 - \theta_\star\|}{\varepsilon}}{\alpha \lambda_{\min}(1-\gamma)}}_{\textit{initial error}} + \underbrace{\frac{\mathrm{Tr}(\Sigma_\varepsilon^{(opt)})}{\varepsilon^2} + \frac{\alpha(1 + \|\theta_\star\|_{\Sigma_\varphi}^2)}{(1-\gamma)^3 \lambda_{\min}^2 \varepsilon^2} + \mathsf{R}_4(1/\varepsilon)}_{\textit{variance term}} \Big) ,$$

TD(0) updates, where $\mathsf{R}_4(1/\varepsilon)$ *scales linearly with* $1/\varepsilon$.

5. TD learning under Markov noise

Here we present an extension of the results of Section 3 under Markovian sampling. The corresponding results generalize the high probability bounds of Corollary 4 and Theorem 4. We start with the following assumption:

TD 3 *Training tuples* (s_k, a_k, s_{k+1}) *are generated sequentially following the generative model* $a_k \sim \pi(\cdot|s_k)$, $s_{k+1} \sim P(\cdot|s_k, a_k)$.

Note that the assumption **TD** 3 yields that the sequence $\{s_k\}_{k\in\mathbb{N}}$ is a Markov chain with the Markov kernel $P_\pi(\cdot|s)$ defined in (11), that corresponds to a classical problem of on-policy evaluation. However, since we are using a single chain for policy evaluation, our subsequent analysis requires to impose ergodicity constraints on $P_\pi(\cdot|s)$.

TD 4 *The Markov kernel* P_π *admits a unique invariant distribution* μ *and is uniformly geometrically ergodic, that is, there exist* $t_{\mathrm{mix}} \in \mathbb{N}$, *such that for any* $k \in \mathbb{N}$, *it holds that*

$$\sup_{s,s' \in \mathcal{S}} (1/2) \|P_\pi^k(\cdot|s) - P_\pi^k(\cdot|s')\|_{\mathsf{TV}} \le (1/4)^{\lfloor k/t_{\mathrm{mix}} \rfloor} . \tag{21}$$

We note that **TD** 4 is widely used in theoretical RL and stochastic optimization, see, e.g. (Bhandari et al., 2018; Nagaraj et al., 2020; Dorfman and Levy, 2022; Patil et al., 2023). The parameter t_{mix} is the *mixing time*, see e.g. (Paulin, 2015). The constant $1/4$ in (21) can be changed to arbitrary constant in $[0,1)$ with proper rescaling of t_{mix}.

The algorithm that we analyze in the Markovian setting is not a standard version of TD(0), but its modification with data-drop. It is summarized in Algorithm 2 and has additional parameter $q \in \mathbb{N}$. We take every q-th tuple from the trajectory $\{(s_k, a_k, s_{k+1})\}_{k\in\mathbb{N}}$. Parameter q here needs

Algorithm 2: Temporal difference learning TD(0) with data drop

Input : features $\varphi(\cdot) : \mathcal{S} \to \mathbb{R}^d$, step size α, number of iterations n, burn-in size n_0, behavioral policy π, time window $q \in \mathbb{N}$

Compute number of blocks $m = \lfloor n/q \rfloor$

for $k = 1, \dots, n$ **do**

  Receive tuple (s_k, a_k, s_{k+1}) following **TD** 4

  if $k = qj, j \in \mathbb{N}$ **then**

    Compute update $\tilde{\theta}_j = \tilde{\theta}_{j-1} - \alpha(\mathbf{A}_k \tilde{\theta}_{j-1} - \mathbf{b}_k)$ based on $\mathbf{A}_k$, $\mathbf{b}_k$ from (15)

  else

    skip current learning tuple

  end

end

Output: tail-averaged estimate $\bar{\theta}_n = (2/m) \sum_{k=m/2+1}^{m} \tilde{\theta}_k$

value function estimate $V^{\pi}_{\bar{\theta}_n}(s) = \varphi^{\top}(s)\, \bar{\theta}_n$

to be properly adjusted with t_{mix}, see Theorem 6 below. The data-drop approach was previously explored in (Nagaraj et al., 2020) for the general least-squares problems. The authors of (Nagaraj et al., 2020) further established that this strategy is optimal in a sense that required number of samples scales linearly with $\mathrm{t}_{\mathrm{mix}}$, and this dependence is worst-case optimal. In the context of TD(0) algorithm the same approach was suggested and studied by (Patil et al., 2023), with the restriction to finite state space setting. Now we are ready to state and prove the counterpart of Theorem 4 for the case of TD(0) updates generated by Algorithm 2.

Theorem 6 *Assume **TD** 2, **TD** 3, and **TD** 4. Fix $\delta \in (0, 1/3)$ and let $\bar{\theta}_n$ be a tail-averaged estimate generated by Algorithm 2 run with parameters*

$$\alpha = \frac{1-\gamma}{128 \log(n/\delta)}, \quad q = \left\lceil \frac{t_{\mathrm{mix}} \log(n/\delta)}{\log 4} \right\rceil,$$

given that the sample size n satisfies $n \geq \frac{\log(1/\delta)}{(1-\gamma)^2} \vee \frac{2 t_{\mathrm{mix}} \log(4/\delta)}{\log 4}$. Then it holds with probability at least $1 - 3\delta$ that

$$\|\bar{\theta}_n - \theta_\star\|_{\Sigma_\varphi} \lesssim \frac{(\|\theta_\star\|_{\Sigma_\varphi} + 1) t_{\mathrm{mix}}^{1/2} \log(n/\delta)}{n^{1/2}(1-\gamma)\lambda_{\min}} + \exp\left\{ -\frac{(1-\gamma)^2 \lambda_{\min} n}{128 t_{\mathrm{mix}} \log^2(n/\delta)} \right\} \frac{\|\theta_0 - \theta_\star\| t_{\mathrm{mix}} \log^2(n/\delta)}{(1-\gamma)^2 \lambda_{\min} n}. \quad (22)$$

The proof is postponed to Appendix D and is based on Berbee's coupling lemma, see (Berbee, 1979). Note that the result of Theorem 6 is slightly suboptimal compared to Corollary 4. Indeed, the leading term with respect to n of the bound (22) scales with $\log(1/\delta)$ instead of $\sqrt{\log(1/\delta)}$ in the i.i.d. counterpart. That is, the leading term of (22) exhibits subexponential behaviour instead of sub-Gaussian. This behaviour is a result of using Berbee's coupling lemma.

Discussion The practical application of data-drop approach is limited, since the gap size q in Algorithm 2 should scale with unknown in practice parameter $\mathrm{t}_{\mathrm{mix}}$. It is possible to analyze under

TD 3 and **TD** 4 a direct counterpart of Algorithm 1 without data-drop. The key difficulty of such analysis is to verify an exponential stability assumption A2. This is done, for example, in (Durmus et al., 2024, Proposition 7) for the general LSA problem. However, the respective stability threshold $\alpha_{p,\infty}$ scales as t_{mix}^{-1}. This means, that from theoretical perspective we still observe the following dilemma - either we run data-drop algorithm with the number of dropped observations, which scales with t_{mix}, or we run Algorithm 1 without data-drop, but the step size α has to be adjusted with t_{mix}^{-1}.

Similarly to the i.i.d. setting, we can rewrite Theorem 6 as a sample complexity bound.

Corollary 3. *Under assumptions of Theorem 6 in order to achieve $\|\bar{\theta}_n - \theta_\star\|_{\Sigma_\varphi} \leq \varepsilon$ with probability at least $1 - 3\delta$ it requires*

$$\tilde{\mathcal{O}}\left(\frac{t_{\mathrm{mix}}(\|\theta_\star\|_{\Sigma_\varphi}^2 + 1)\log(1/\delta)}{(1-\gamma)^2\lambda_{\min}^2\varepsilon^2} + \frac{t_{\mathrm{mix}}\log^2(1/\delta)}{\lambda_{\min}(1-\gamma)^2}\log\frac{\|\theta_0 - \theta_\star\|}{\varepsilon}\right)$$

observation used in Algorithm 2.

Note that in Corollary 3 the symbol $\tilde{\mathcal{O}}$ hides logarithmic dependencies in $\lambda_{\min}, 1-\gamma$, and n, but not in $1/\delta$. The sample complexity bounds of Corollary 3 matches the ones coming from Theorem 4 up to an additional t_{mix} factor and extra factor of $\sqrt{\log(1/\delta)}$. We believe that a factor of $\sqrt{\log(1/\delta)}$ can be removed using the analysis based on Algorithm 1 without data drop and appropriate versions of Rosenthal inequality for Markov chains and leave it as a direction for further work.

6. Conclusion

In this paper we presented a refined analysis of linear stochastic approximation algorithms and provide high-probability and sample complexity bounds for the TD(0) algorithm via the exponential stability argument. Our approach allows to obtain high-probability bounds without requiring projections or instance-dependent step size. Further research directions include generalizing the high-order error bounds to the Markov setting for versions of the TD learning algorithm that do not use the data drop modification, while maintaining the precise variance from the corresponding central limit theorem. Second, our version of Algorithm 2 requires knowledge of t_{mix}, which is a common drawback shared by the versions of SGD with data drop algorithm (Nagaraj et al., 2020). To the best of our knowledge, it is an open problem to develop a version of this algorithm which would be oblivious to t_{mix}.

Acknowledgments

The work of Sergey Samsonov and Alexey Naumov was prepared within the framework of the HSE University Basic Research Program. The work of D. Tiapkin has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF. The work by Eric Moulines is partially funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work was partially conducted under the auspices of the Lagrange Mathematics and Computing Research Center.

References

Rafik Aguech, Eric Moulines, and Pierre Priouret. On a perturbation approach for the analysis of stochastic tracking algorithms. *SIAM Journal on Control and Optimization*, 39(3):872–899, 2000.

H.C.P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*. Mathematical Centre tracts. Centrum Voor Wiskunde en Informatica, 1979. ISBN 9789061961826.

Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

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

G. Dalal, Balázs Szörényi, G. Thoppe, and S. Mannor. Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jérôme Dedecker and Sana Louhichi. Maximal inequalities and empirical central limit theorems. In *Empirical process techniques for dependent data*, pages 137–159. Springer, 2002.

Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with Markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018. ISBN 978-3-319-97703-4.

Yaqi Duan and Martin J. Wainwright. A finite-sample analysis of multi-step temporal difference estimates. In Nikolai Matni, Manfred Morari, and George J. Pappas, editors, *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pages 612–624. PMLR, 15–16 Jun 2023.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed stepsize. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, and Y. Dauphin, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 30063–30074. Curran Associates, Inc., 2021a.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with Markovian noise: Application to linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR, 2021b.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for Polyak–Ruppert averaged iterates of linear stochastic approximation. *Mathematics of Operations Research*, 2024.

G. Fort. Central limit theorems for stochastic approximation with controlled Markov chain dynamics. *ESAIM: PS*, 19:60–80, 2015.

L. Guo. Stability of recursive stochastic tracking algorithms. *SIAM Journal on Control and Optimization*, 32(5):1195–1225, 1994.

L. Guo and L. Ljung. Exponential stability of general tracking algorithms. *IEEE Transactions on Automatic Control*, 40(8):1376–1387, 1995.

De Huang, Jonathan Niles-Weed, Joel A Tropp, and Rachel Ward. Matrix concentration for products. *Foundations of Computational Mathematics*, pages 1–33, 2021.

Dongyan Huo, Yudong Chen, and Qiaomin Xie. 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*, pages 81–82, 2023.

Birgit Jacob and Hans J Zwart. *Linear port-Hamiltonian systems on infinite-dimensional spaces*, volume 223. Springer Science & Business Media, 2012.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040, 2021.

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

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

Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. High-probability sample complexities for policy evaluation with linear function approximation. *IEEE Transactions on Information Theory*, 2024.

Tianjiao Li, Guanghui Lan, and Ashwin Pananjady. Accelerated and instance-optimal policy evaluation with linear function approximation. *SIAM Journal on Mathematics of Data Science*, 5(1): 174–200, 2023. doi: 10.1137/21M1468668.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

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

A. Osekowski. *Sharp Martingale and Semimartingale Inequalities*. Monografie Matematyczne 72. Birkhäuser Basel, 1 edition, 2012. ISBN 3034803699,9783034803694.

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

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015. doi: 10.1214/EJP.v20-4039.

Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

P. Priouret and A. Veretenikov. A remark on the stability of the LMS tracking algorithm. *Stochastic analysis and applications*, 16(1):119–129, 1998.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

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

R. S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997. ISSN 2334-3303. doi: 10.1109/9.580874.

Appendix A. Proofs of LSA error bounds presented in Section 2

Recall that we consider a sequence of LSA estimates $\{\theta_n\}_{n\in\mathbb{N}}$ given by the recurrence

$$\theta_n = \theta_{n-1} - \alpha\{\mathbf{A}_n\theta_{n-1} - \mathbf{b}_n\} \ , \ \ n \geq 1 \ . \tag{23}$$

In the formula above we use $\mathbf{A}_n$ and $\mathbf{b}_n$ as a shorthand notations for $\mathbf{A}(Z_n)$ and $\mathbf{b}(Z_n)$, respectively. We use the same convention throughout the appendix section. Our analysis relies heavily on the stability assumption A2 for the matrix products of the form

$$\Gamma_{1:n}^{(\alpha)} = \prod_{i=1}^{n}(\mathrm{I} - \alpha\mathbf{A}_i) \ .$$

We obtain the following refined bound on the last iterate error of the procedure (23):

Theorem 7

(i) Assume A1 and A2(2). Then, for any $\alpha \in (0; \alpha_{2,\infty})$ and $n \in \mathbb{N}$, it holds that

$$\mathbb{E}^{1/2}[\|\theta_n - \theta_\star\|^2] \leq \varkappa_2(1-\alpha a)^n\|\theta_0 - \theta_\star\| + \frac{\varkappa_2\sqrt{\alpha\,\mathrm{Tr}(\Sigma_\varepsilon)}}{\sqrt{a}} \ . \tag{24}$$

(ii) Let $p \geq 2$. Assume A1 and A2(p). Then, for any $\alpha \in (0; \alpha_{p,\infty})$ and $n \in \mathbb{N}$, it holds that

$$\mathbb{E}^{1/p}[\|\theta_n - \theta_\star\|^p] \leq \varkappa_p(1-\alpha a)^n\|\theta_0 - \theta_\star\| + \frac{\varkappa_p p\sqrt{\alpha}}{\sqrt{a}}\|\varepsilon\|_\infty \ . \tag{25}$$

(iii) Grant A1 and A2(ℓ) for any $\ell \geq 2$. Then for any $p \geq 2$, $n \geq 2$ and $\alpha \in [0; \alpha_{p+\log n,\infty})$, it holds that

$$\begin{aligned}\mathbb{E}^{1/p}[\|\theta_n - \theta_\star\|^p] \leq \varkappa_{p+\log n}(1-\alpha a)^n\|\theta_0 - \theta_\star\| + \mathrm{C}_{\mathsf{Rm},\mathbf{1}}\, p^{1/2}\frac{\varkappa_{p+\log n}\sqrt{\alpha\,\mathrm{Tr}(\Sigma_\varepsilon)}}{\sqrt{a}} \\ + \mathrm{C}_{\mathsf{Rm},\mathbf{2}}\,\mathrm{e}\alpha p\varkappa_{p+\log n}\|\varepsilon\|_\infty \ ,\end{aligned} \tag{26}$$

where $\mathrm{C}_{\mathsf{Rm},\mathbf{1}} = 60$ and $\mathrm{C}_{\mathsf{Rm},\mathbf{2}} = 60\mathrm{e}$ are constants from the martingale version of Rosenthal's inequality (Pinelis, 1994, Theorem 4.1).

Proof Using the error expansion technique from Aguech et al. (2000) (see also Durmus et al. (2024)), we decompose θ_n into a transient and fluctuation terms

$$\theta_n - \theta_\star = \tilde{\theta}_n^{(\mathsf{tr})} + \tilde{\theta}_n^{(\mathsf{fl})} \ ,$$

where we have defined the quantities

$$\tilde{\theta}_n^{(\mathsf{tr})} = \Gamma_{1:n}^{(\alpha)}\{\theta_0 - \theta_\star\} \ , \quad \tilde{\theta}_n^{(\mathsf{fl})} = -\alpha\textstyle\sum_{j=1}^{n}\Gamma_{j+1:n}^{(\alpha)}\varepsilon_j \ . \tag{27}$$

The first term $\tilde{\theta}_n^{(\mathsf{tr})}$ in the error decomposition (27) is transient and reflects the forgetting of the initial error of the LSA. It can be directly controlled using the assumption A2(p), $p \geq 2$:

$$\mathbb{E}^{1/p}[\|\Gamma_{1:n}^{(\alpha)}\{\theta_0 - \theta_\star\}\|^p] \leq \varkappa_p(1-\alpha a)^n\|\theta_0 - \theta_\star\| \ .$$

In order to control the fluctuation term $\tilde{\theta}_n^{(\mathsf{fl})}$, we note that it is a reverse martingale w.r.t. filtration $\mathcal{F}_k = \sigma(Z_j, j \geq k)$. Thus, applying the Burkholder inequality (Osekowski, 2012, Theorem 8.6), we obtain that, assuming A2(p)

$$\begin{aligned}
\mathbb{E}^{1/p}\left[\left\|\alpha \sum_{j=1}^n \Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\right\|^p\right] &\leq \alpha p \left(\mathbb{E}^{2/p}\left[\left(\sum_{j=1}^n \|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^2\right)^{p/2}\right]\right)^{1/2} \\
&\leq \alpha p \left(\sum_{j=1}^n \mathbb{E}^{2/p}\left[\|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^p\right]\right)^{1/2} \\
&\leq \alpha p \varkappa_p \big(\mathbb{E}^{2/p}\left[\|\varepsilon_1\|^p\right] \sum_{j=1}^n (1-\alpha a)^{2(n-j)}\big)^{1/2} \\
&\leq \frac{\varkappa_p p \sqrt{\alpha}}{\sqrt{a}} \|\varepsilon\|_\infty \,,
\end{aligned}$$

where for the last bound we additionally used that $\alpha a \leq 1/2$. Substituting the bounds above into (27) completed the proof. Obtaining the second moment bound (24) follows the same lines as above using the martingale structure of $\tilde{\theta}_n^{(\mathsf{fl})}$, that is,

$$\begin{aligned}
\mathbb{E}^{1/2}\left[\left\|\alpha \sum_{j=1}^n \Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\right\|^2\right] &\leq \alpha \left(\sum_{j=1}^n \mathbb{E}\left[\|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^2\right]\right)^{1/2} \\
&\leq \alpha \varkappa_{p+\log n} \left(\sum_{j=1}^n (1-\alpha a)^{2(n-j)} \operatorname{Tr}(\Sigma_\varepsilon)\right)^{1/2} \\
&\leq \frac{\varkappa_{p+\log n} \sqrt{\alpha}}{\sqrt{a}} \{\operatorname{Tr}(\Sigma_\varepsilon)\}^{1/2} \,.
\end{aligned}$$

Now we aim to obtain the refined bound (26). For $k \in \{1, \ldots, n\}$, we set $\mathcal{F}_k = \sigma(Z_s \,:\, s \leq k)$, and $\mathcal{F}_0 = \{\emptyset, \mathsf{Z}\}$. Then it is easy to see that $\mathbb{E}^{\mathcal{F}_{j-1}}\left[\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\right] = 0$ for any $j = 1, \ldots, n$. Hence, applying the Pinelis version of Rosenthal inequality (Pinelis, 1994, Theorem 4.1), we obtain that

$$\begin{aligned}
\mathbb{E}^{1/p}\left[\left\|\alpha \sum_{j=1}^n \Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\right\|^p\right] \leq \alpha\, \mathrm{C}_{\mathsf{Rm},\mathbf{1}}\, p^{1/2} \mathbb{E}^{1/p}\left[\left(\sum_{j=1}^n \mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^2\right]\right)^{p/2}\right] \\
+ \alpha p\, \mathrm{C}_{\mathsf{Rm},\mathbf{2}}\, \mathbb{E}^{1/p}\left[\max_j \|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^p\right] \,. \quad (28)
\end{aligned}$$

Since ε_j is independent of $\Gamma_{j+1:n}^{(\alpha)}$, it is easy to see that

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^2\right] &= \mathbb{E}^{\mathcal{F}_{j-1}}\left[\mathbb{E}^{\mathcal{F}_j}\left[\|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^2\right]\right] \leq \varkappa_{p+\log n}^2 (1-\alpha a)^{2(n-j)} \mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\varepsilon_j\|^2\right] \\
&= \varkappa_{p+\log n}^2 (1-\alpha a)^{2(n-j)} \operatorname{Tr}(\Sigma_\varepsilon) \,.
\end{aligned}$$

Thus, with simple algebra and using that $\alpha a \leq 1/2$, we get that

$$\alpha\, \mathrm{C}_{\mathsf{Rm},\mathbf{1}}\, p^{1/2} \mathbb{E}^{1/p}\left[\left(\sum_{j=1}^n \mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\Gamma_{j+1:n}^{(\alpha)} \varepsilon_j\|^2\right]\right)^{p/2}\right] \leq \mathrm{C}_{\mathsf{Rm},\mathbf{1}}\, p^{1/2} \frac{\varkappa_{p+\log n} \sqrt{\alpha \operatorname{Tr}(\Sigma_\varepsilon)}}{\sqrt{a}} \,.$$

In order to control the remainder term in Rosenthal's inequality (28), we note that, with $q = p + \log n$, it holds

$$\mathbb{E}^{1/p}\big[\max_j \|\Gamma^{(\alpha)}_{j+1:n}\varepsilon_j\|^p\big] \le \mathbb{E}^{1/q}\big[\max_j \|\Gamma^{(\alpha)}_{j+1:n}\varepsilon_j\|^q\big] \le \Big(\sum\nolimits_{j=1}^n \mathbb{E}[\|\Gamma^{(\alpha)}_{j+1:n}\varepsilon_j\|^q]\Big)^{1/q}$$
$$\le \varkappa_{p+\log n} n^{1/q}\|\varepsilon\|_\infty \le \mathrm{e}\varkappa_{p+\log n}\|\varepsilon\|_\infty \,.$$

Now it remains to combine the bounds above in (28), and the result of (26) follows. ■

Note that Theorem 7 provides 2 bounds for the last LSA iterate error, (25) and (26). The second one might provide an improvement, since $\|\varepsilon\|_\infty \ge \sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}$. If we aim to obtain bounds in terms of solely the noise variance $\mathrm{Tr}(\Sigma_\varepsilon)$, we need that the reverse inequality holds, that is,

$$\|\varepsilon\|_\infty \le c_1\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}$$

for some appropriate constant $c_1 > 0$. The problem is that the scaling of c_1 with instance-dependent quantities of Section 3 might be pessimistic. That is why it is desirable to have this dependence coming with additional α factor, instead of just $\sqrt{\alpha}$ in (25).

Now we state and proof the similar results for the Polyak-Ruppert averaged estimator $\bar\theta_{n_0,n}$. We use the following decomposition based on the summation by parts formula:

$$\bar{\mathbf{A}}\left(\bar\theta_{n_0,n} - \theta_\star\right) = \frac{\theta_{n_0} - \theta_n}{\alpha(n-n_0)} - \frac{\sum_{t=n_0}^{n-1} e\left(\theta_t, Z_{t+1}\right)}{n - n_0}\,, \tag{29}$$

where we have defined

$$e(\theta, z) = \tilde{\mathbf{A}}(z)\theta - \tilde{\mathbf{b}}(z) = \varepsilon(z) + \tilde{\mathbf{A}}(z)(\theta - \theta_\star)\,. \tag{30}$$

The decomposition above is nothing but summation by parts formula used in Mou et al. (2020), yet it can be traced to the preceding papers. Recall also that we have set the notation $\bar\theta_n$ as an alias for $\bar\theta_{n_0,n}$ used with $n_0 = n/2$. Before we proceed to the proof of Theorem 2, we first provide a simpler statement regarding the 2-nd moment of the PR-averaged error.

Theorem 8 *Assume A1 and A2(2). Then for any $n \ge 2$, $\alpha \in (0; \alpha_{2,\infty}]$, it holds that*

$$\mathbb{E}[\|\bar{\mathbf{A}}(\bar\theta_n - \theta_\star)\|^2] \lesssim \varkappa_2^2(1-\alpha a)^n\left(\frac{1}{\alpha^2 n^2} + \frac{\mathrm{C}_{\mathbf{A}}^2}{\alpha a n^2}\right)\|\theta_0 - \theta_\star\|^2$$
$$+ \frac{\mathrm{Tr}(\Sigma_\varepsilon)}{n}\left(1 + \frac{\varkappa_2^2\,\mathrm{C}_{\mathbf{A}}^2\,\alpha}{a}\right) + \frac{\varkappa_2^2\,\mathrm{Tr}(\Sigma_\varepsilon)}{\alpha a n^2}\,. \tag{31}$$

Proof Our proof is essentially a version of (Durmus et al., 2024, Proposition 5) with tighter instance-dependent bound on the last LSA iterate error provided by Theorem 7. We leverage the error decomposition (29). Then we get

$$\mathbb{E}[\|\bar{\mathbf{A}}(\bar\theta_n - \theta_\star)\|^2] \le \underbrace{\frac{8\mathbb{E}[\|\theta_{n/2} - \theta_n\|^2]}{\alpha^2 n^2}}_{T_1} + \underbrace{\frac{8\mathbb{E}[\|\sum_{t=n/2}^{n-1} e\left(\theta_t, Z_{t+1}\right)\|^2]}{n^2}}_{T_2}\,,$$

and estimate the terms T_1 and T_2 separately. Applying the bounds of Theorem 7, we get first that

$$T_1 \lesssim \frac{\varkappa_2^2(1-\alpha a)^n\|\theta_0-\theta_\star\|^2}{\alpha^2 n^2} + \frac{\varkappa_2^2\,\mathrm{Tr}(\Sigma_\varepsilon)}{\alpha a n^2}\,.$$

Similarly, since $e\,(\theta_t, Z_{t+1})$ is a martingale-difference sequence w.r.t. filtration $\mathcal{F}_k = \sigma(Z_j, j \le k)$, we get the following bound for T_2:

$$T_2 \le n^{-2}\sum_{t=n/2}^{n-1}\mathbb{E}[\|e\,(\theta_t, Z_{t+1})\,\|^2] \lesssim \frac{\mathrm{Tr}(\Sigma_\varepsilon)}{n} + \frac{\varkappa_2^2\,\mathrm{C}_{\mathbf{A}}^2(1-\alpha a)^n\|\theta_0-\theta_\star\|^2}{\alpha a n^2} + \frac{\varkappa_2^2\,\mathrm{C}_{\mathbf{A}}^2\,\alpha\,\mathrm{Tr}(\Sigma_\varepsilon)}{an}\,,$$

and it remains to combine the above bounds. ■

Now we are ready to proceed with the main result of this section, that is, with the p-moment error bound Theorem 2.

Theorem 9 *Assume A1 and A2(∞). Then for any $p \ge 2$, $n \ge 2$, $\alpha \in [0; \alpha_{p+\log n,\infty})$, it holds that*

$$\begin{aligned}
\mathbb{E}^{1/p}[\|\bar{\mathbf{A}}(\bar\theta_n-\theta_\star)\|^p] &\lesssim \frac{p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{n^{1/2}}\left(1 + \frac{\varkappa_{p+\log n}\sqrt{\alpha p}\,\mathrm{C}_{\mathbf{A}}}{\sqrt{a}} + \frac{\varkappa_{p+\log n}\,\mathrm{C}_{\mathbf{A}}\,\alpha p\|\varepsilon\|_\infty}{\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}\right)\\
&+ \frac{\varkappa_{p+\log n} p\|\varepsilon\|_\infty}{n}\left(1+\mathrm{C}_{\mathbf{A}}\,\alpha(p+\log n)\right)\\
&+ \frac{\varkappa_{p+\log n}p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{\sqrt{a}n}\left[\frac{1}{\sqrt{\alpha}} + p^{1/2}\,\mathrm{C}_{\mathbf{A}}\,\sqrt{\alpha(p+\log n)}\right]\\
&+ \varkappa_{p+\log n}(1-\alpha a)^{n/2}\left(\frac{1}{\alpha n} + \frac{p\,\mathrm{C}_{\mathbf{A}}}{\sqrt{\alpha a}n}\right)\|\theta_0-\theta_\star\|\,.
\end{aligned}$$

Proof The proof is also based on the expansion formula (29). We recall that we set $n_0 = n/2$. Then, with the direct application of Minkowski's inequality, we obtain

$$\mathbb{E}^{1/p}\left[\|\bar{\mathbf{A}}\left(\bar\theta_n-\theta_\star\right)\|^p\right] \le \underbrace{\frac{\mathbb{E}^{1/p}[\|\theta_{n/2}-\theta_n\|^p]}{\alpha n}}_{T_1} + \underbrace{\frac{\mathbb{E}^{1/p}\left[\|\sum_{t=n/2}^{n-1}\mathrm{e}(\theta_t, Z_{t+1})\|^p\right]}{n}}_{T_2}\,,$$

and bound T_1, T_2 separately. Note that T_1 is a remainder term (w.r.t. sample size n), and thus we can control it using a simple bound on the last iterate error provided in Theorem 7-(26). Proceeding this way, we obtain

$$T_1 \lesssim \frac{\varkappa_{p+\log n}(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\|}{\alpha n} + \frac{\varkappa_{p+\log n}p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)}}{\sqrt{\alpha a}n} + \frac{p\varkappa_{p+\log n}\|\varepsilon\|_\infty}{n}\,.$$

Now we proceed with bounding T_2. Using again Minkowski's inequality, we get

$$T_2 \le n^{-1}\,\mathbb{E}^{1/p}\left[\|\textstyle\sum_{t=n/2}^{n-1}\varepsilon_{t+1}\|^p\right] + n^{-1}\,\mathbb{E}^{1/p}[\|\textstyle\sum_{t=n/2}^{n-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p]\,.$$

The first term of the above sum can be controlled by directly applying Pinelis' version of Rosenthal's inequality (Pinelis, 1994, Theorem 4.3):

$$\mathbb{E}^{1/p}\left[\|\textstyle\sum_{t=n/2}^{n-1}\varepsilon_{t+1}\|^p\right] \lesssim p^{1/2}n^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon)} + p\|\varepsilon\|_\infty\,.$$

It remains to bound $\mathbb{E}^{1/p}[\|\sum_{t=n/2}^{n-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p]$. Note that the sequence $\{\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\}$ is a martingale-difference w.r.t. $\mathcal{F}_t=\sigma(Z_k,k\le t)$. A further application of Rosenthal's inequality thus shows that

$$\begin{aligned}
&\mathbb{E}^{1/p}\left[\left\|\sum_{t=n/2}^{n-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\right\|^p\right]\\
&\lesssim p^{1/2}\mathbb{E}^{1/p}\left[\left(\sum_{t=n/2}^{n-1}\mathbb{E}^{\mathcal{F}_t}\left[\|\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^2\right]\right)^{p/2}\right]+p\,\mathbb{E}^{1/p}\left[\max_t\|\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p\right]\\
&\le p^{1/2}\,\mathrm{C}_{\mathbf{A}}\left(\sum_{t=n/2}^{n-1}\mathbb{E}^{2/p}\left[\|\theta_t-\theta_\star\|^p\right]\right)^{1/2}+p\,\mathrm{C}_{\mathbf{A}}\,\mathbb{E}^{1/p}\left[\max_t\|\theta_t-\theta_\star\|^p\right]\,.
\end{aligned}$$

Now, applying the last iterate bound Theorem 7-(26), and using that $\alpha a\le 1/2$, we get

$$\begin{aligned}
p^{1/2}\,\mathrm{C}_{\mathbf{A}}\left(\sum_{t=n/2}^{n-1}\mathbb{E}^{2/p}[\|\theta_t-\theta_\star\|^p]\right)^{1/2}&\lesssim\frac{\varkappa_{p+\log n}p^{1/2}\,\mathrm{C}_{\mathbf{A}}(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\|}{\sqrt{\alpha a}}\\
&+\frac{\varkappa_{p+\log n}\,\mathrm{C}_{\mathbf{A}}\,p\sqrt{\alpha n\,\mathrm{Tr}(\Sigma_\varepsilon)}}{\sqrt{a}}+\varkappa_{p+\log n}\,\mathrm{C}_{\mathbf{A}}\,\alpha p^{3/2}n^{1/2}\|\varepsilon\|_\infty\,.
\end{aligned}$$

Moreover, a further application of Theorem 7-(26) together with $n^{1/\log n}\le \mathrm{e}$ yield

$$\begin{aligned}
p\,\mathrm{C}_{\mathbf{A}}\,\mathbb{E}^{1/p}[\max_t\|\theta_t-\theta_\star\|^p]&\le p\,\mathrm{C}_{\mathbf{A}}\left(\sum_{t=n/2}^{n}\mathbb{E}[\|\theta_t-\theta_\star\|^{p+\log n}]\right)^{1/(p+\log n)}\\
&\lesssim p\,\mathrm{C}_{\mathbf{A}}\,n^{1/(p+\log n)}\max_{n/2\le t<n}\mathbb{E}^{1/(p+\log n)}[\|\theta_t-\theta_\star\|^{p+\log n}]\\
&\lesssim\varkappa_{p+\log n}p\,\mathrm{C}_{\mathbf{A}}(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\|+\varkappa_{p+\log n}p\,\mathrm{C}_{\mathbf{A}}\,\frac{\sqrt{\alpha(p+\log n)\,\mathrm{Tr}(\Sigma_\varepsilon)}}{a}\\
&\quad+\varkappa_{p+\log n}p\,\mathrm{C}_{\mathbf{A}}\,\alpha(p+\log n)\|\varepsilon\|_\infty\,.
\end{aligned}$$

Now it remains to combine the obtained bounds, and the statement follows. The result of Theorem 2 follows from a simple observation that $\alpha(p+\log n)\le 1/2$ under A 2$(p+\log n)$ for $\alpha\in(0;\alpha_{p+\log n,\infty}]$. ∎

Appendix B. Proofs of TD learning of Section 3

B.1. Proof of Lemma 1

Proof Under **TD** 2, it is easily seen that $\|\mathbf{A}_1\|\le(1+\gamma)$ almost surely, which implies $\|\bar{\mathbf{A}}\|\le(1+\gamma)$. The remaining bounds follow from

$$\begin{aligned}
\|\varepsilon\|_\infty&=\sup_{z\in\mathsf{Z}}\|\varepsilon(z)\|=\sup_{z=(s,s')}\|(\mathbf{A}(z)-\bar{\mathbf{A}})\theta_\star-(\mathbf{b}(z)-\bar{\mathbf{b}})\|\le 2(1+\gamma)(\|\theta_\star\|+1)\,,\\
\mathrm{Tr}(\Sigma_\varepsilon)&=\mathbb{E}[\|(\mathbf{A}_1-\bar{\mathbf{A}})\theta_\star-(\mathbf{b}_1-\bar{\mathbf{b}})\|^2]\le 2\theta_\star^\top\mathbb{E}[\mathbf{A}_0^\top\mathbf{A}_0]\theta_\star+2\mathbb{E}[r^2(s_0)\,\mathrm{Tr}(\varphi(s_0)\varphi^\top(s_0))]\\
&\le 2(1+\gamma)^2\theta_\star^\top\Sigma_\varphi\theta_\star+2\le 2(1+\gamma)^2\left(\|\theta_\star\|^2_{\Sigma_\varphi}+1\right)\,,
\end{aligned}$$

and the statement follows. ■

B.2. Proof of Lemma 2

In this subsection we obtain a new, refined bounds on the transient term $\tilde{\theta}_n^{(\mathsf{tr})} = \Gamma_{1:n}^{(\alpha)}\{\theta_0 - \theta_\star\}$ appearing in the error decomposition (27) in case of TD(0) algorithm. Recall that in case of the general LSA algorithm we have to refer to the matrix product stability result of Theorem 13, which is based on the framework suggested by Huang et al. (2021). The interplay between step size $\alpha_{\infty,p}$ and maximal controlled moment p (which roughly can be written as $\alpha_{\infty,p} \lesssim 1/p$) is in general unavoidable. The respective $1-$dimensional counterexample is provided in (Durmus et al., 2021a, Example 1). At the same time, the general L_p-stability of the random matrix product appears to induce some undesirable phenomenons. First, it induces the additional $d^{1/p}$ factor in the r.h.s. of the bound (5). Such a dependence requires to introduce additional (logarithmic) dependence of the dimension d in the step size α in order to remove the $d^{1/p}$ factor in the r.h.s..

Second, and more important, the trade-off between $\|\mathrm{I} - \alpha\bar{\mathbf{A}}\| \le 1 - \alpha a$ and upper bounds for fluctuation term $\alpha\mathbb{E}^{1/p}[\|\mathbf{A} - \bar{\mathbf{A}}\|^p]$ requires that the step size α scales with some instance-dependent quantities, related with the matrix $\bar{\mathbf{A}}$. Typically this means that the resulting rate-optimal algorithm is not really implementable, as $\bar{\mathbf{A}}$ is not accessible in practice.

This drawback is shared by most of the recent papers on the subject, see e.g. (Li et al., 2024, Theorem 1), where the maximal allowed step size α scales with $\lambda_{\min}(\Sigma)$. Our subsequent analysis allows us to eliminate this drawback. Recall that for any $1 \le j \le n$ we set $\mathcal{F}_j = \sigma(Z_i, 1 \le i \le j)$ and $\mathcal{F}_0 = \emptyset$. Then the exponential stability property of Lemma 2 will follow from the following general result:

Theorem 10 *Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14) under **TD 1** and **TD 2**. Then, for any $n \in \mathbb{N}$, $1 \le j \le n$, $p \ge 2$, step size $\alpha \in (0; \frac{1-\gamma}{128p}]$, and any ξ_{j-1} being a d-dimensional $\mathcal{F}_{j-1}$-measurable random vector, it holds $\mathbb{P}$-a.s. that*

$$\mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\Gamma_{j:n}^{(\alpha)}\xi_{j-1}\|^p\right] \le (1 - \alpha p(1-\gamma)\lambda_{\min}/2)^{n-j}\|\xi_{j-1}\|^p \,. \tag{32}$$

In particular, for any $\theta_0 \in \mathbb{R}^d$ we obtain that

$$\mathbb{E}^{1/p}[\|\Gamma_{1:n}^{(\alpha)}(\theta_0 - \theta_\star)\|^p] \le (1 - \alpha(1-\gamma)\lambda_{\min}/2)^{n-j}\|\theta_0 - \theta_\star\| \,. \tag{33}$$

Proof Note that it is enough to prove the bound (32) for $p = 2^s$, $s \in \mathbb{N}$, since otherwise we can find the nearest dyadic power $q \ge p$ and use the Lyapunov inequality. Note that we increase the power of p by no more than a factor of 2 in such a case.

Now we consider the case $p = 2^s$, $s \in \mathbb{N}$. Then, expanding the p-power of the norm, we get

$$\|\Gamma_{j:n}^{(\alpha)}\xi_{j-1}\|^p = \left(\xi_{j-1}^\top\{\Gamma_{j:n}^{(\alpha)}\}^\top\Gamma_{j:n}^{(\alpha)}\xi_{j-1}\right)^{p/2} = (\eta_{n-1}^\top(\mathrm{I} - \alpha\mathbf{A}_n)^\top(\mathrm{I} - \alpha\mathbf{A}_n)\eta_{n-1})^{p/2} \,,$$

where we have introduced a vector $\eta_{n-1} = \Gamma_{j:n-1}^{(\alpha)}\xi_{j-1}$. Note that a vector η_{n-1} is $\mathcal{F}_{n-1}$-measurable, and thus, combining Lemma 3 and Lemma 4, we get

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\Gamma_{j:n}^{(\alpha)}\xi_{j-1}\|^p\right] &= \mathbb{E}^{\mathcal{F}_{j-1}}\left[\mathbb{E}^{\mathcal{F}_{n-1}}\left[(\eta_{n-1}^\top(\mathrm{I} - \alpha\mathbf{A}_n)^\top(\mathrm{I} - \alpha\mathbf{A}_n)\eta_{n-1})^{p/2}\right]\right] \\ &\le (1 - \alpha p(1-\gamma)\lambda_{\min}/2)\mathbb{E}^{\mathcal{F}_{j-1}}\left[\|\Gamma_{j:n-1}^{(\alpha)}\xi_{j-1}\|^p\right] \,,\end{aligned}$$

and the bound (32) follows by backward induction in n. In order to get the bound (33), it remains to combine (32) together with the fact that $g(x,p) = (1-px)^{1/p}$ monotonically decreases in p for $p \geq 1$ and $0 < x < 1$. ■

The stability result of Theorem 10 favorably compares to the one of Theorem 13. First, we removed an artificial $d^{1/p}$ factor in the r.h.s. of the bound. Second, new stability threshold for α is *computable* and does not contain any instance-independent quantities.

Below we provide some useful auxiliary technical lemmas required for the proof of Theorem 10.

Lemma 3. *Let $B = B^\top \geq 0$, $B \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix and $u \in \mathbb{R}^d$ be some vector. Then, for any $s \in \mathbb{N}$ and $p = 2^s$, it holds that*

$$\left(u^\top B u\right)^p \leq \|u\|^{2p-2} u^\top B^p u \,. \tag{34}$$

Proof We will proof the statement by induction in $s \in \mathbb{N}$. The statement obviously holds for $s = 0$. For $s = 1$ (resp., $p = 2$), we aim to prove that

$$\left(u^\top B u\right)^2 = u^\top B u u^\top B u \leq \|u\|^2 u^\top B^2 u \,, \tag{35}$$

and the statement follows from the bound $Buu^\top B \leq \|u\|^2 B^2$. Let us provide the detailed proof of last inequality. We aim to check that for any $y \in \mathbb{R}^d$ it holds that

$$y^\top B u u^\top B y \leq \|u\|^2 y^\top B^2 y \,.$$

Note that, since B is symmetric and positive definite, $B = U\Lambda U^\top$ with diagonal matrix $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_d\}$ and orthogonal matrix U. Hence, the previous inequality is equivalent to

$$y^\top U\Lambda U^\top u u^\top U\Lambda U^\top y \leq \|u\|^2 y^\top U\Lambda^2 U^\top y \,.$$

Setting $z = U^\top y$ and $v = U^\top u$, we have from the previous bound

$$z^\top \Lambda v v^\top \Lambda z \leq \|v\|^2 z^\top \Lambda^2 z \,.$$

Writing the previous bound in a coordinate form, we obtain that

$$\left(\sum_{i=1}^d \lambda_i z_i v_i\right)^2 \leq \left(\sum_{i=1}^d v_i^2\right)\left(\sum_{i=1}^d \lambda_i^2 z_i^2\right) \,,$$

which holds due to Cauchy-Schwartz inequality, and (35) holds.

Suppose now that the inequality (34) holds for some $p = 2^s$. Then

$$\left(u^\top B u\right)^{2p} = \left(u^\top B u\right)^p \left(u^\top B u\right)^p \leq \|u\|^{4p-4} u^\top B^p u u^\top B^p u \leq \|u\|^{4p-2} u^\top B^{2p} u \,,$$

and the statement follows. ■

Now we provide a key statement on the in-expectation contraction of 1-step-ahead random matrix $\mathbf{A}$ corresponding to the TD(0) algorithm.

Lemma 4. *Let $\mathbf{A} = \varphi(s)\{\varphi(s) - \gamma\varphi(s')\}^\top$ be a random TD update matrix defined in (15), where $s' \sim P^\pi(\cdot|s)$, and $s \sim \mu$. Then, for any $p \in \mathbb{N}$ and $\alpha \in (0; \frac{1-\gamma}{64p}]$, it holds that*

$$\mathbb{E}\left[\{(\mathrm{I} - \alpha\mathbf{A})^\top (\mathrm{I} - \alpha\mathbf{A})\}^p\right] \preceq \mathrm{I} - (1/2)\alpha p(1-\gamma)\Sigma_\varphi \,.$$

Proof Consider the (random) matrix $(\mathrm{I}-\alpha\mathbf{A})^\top(\mathrm{I}-\alpha\mathbf{A})$. Note that it is symmetric, and, introducing matrix $\mathbf{B} = \mathbf{A} + \mathbf{A}^\top - \alpha\mathbf{A}^\top\mathbf{A}$, we get that

$$(\mathrm{I} - \alpha\mathbf{A})^\top(\mathrm{I} - \alpha\mathbf{A}) = \mathrm{I} - \alpha\mathbf{B} \,.$$

Using Lemma 5, it holds that for any $k \in \mathbb{N}$,

$$\mathbb{E}[\mathbf{B}] \succeq (1-\gamma)\Sigma_\varphi \,, \quad \mathbb{E}[\mathbf{B}^k] \preceq \frac{13}{12}\cdot 4^k\Sigma_\varphi \,.$$

Thus, expanding the brackets, we get

$$\mathbb{E}[(\mathrm{I}-\alpha\mathbf{B})^p] \preceq \mathrm{I} - \alpha p\mathbb{E}[\mathbf{B}] + \sum_{k=2}^{p}\alpha^k\binom{p}{k}\mathbb{E}[\mathbf{B}^k] \preceq \mathrm{I} - \alpha p(1-\gamma)\Sigma_\varphi + \frac{13}{12}\cdot\left(\sum_{k=2}^{p}(4\alpha)^k\binom{p}{k}\right)\Sigma_\varphi \,.$$

Since we know that $\alpha p \le (1-\gamma)/64$, we can bound

$$\sum_{k=2}^{p}(4\alpha)^k\binom{p}{k} \le \sum_{k=2}^{p}(4\alpha p)^k \le \frac{16\alpha^2p^2}{1-4\alpha p} \le \frac{16}{15}\cdot 16\alpha^2p^2 \le \frac{16}{15}\cdot\alpha p(1-\gamma)/4 \,.$$

Thus the combination of above bounds together with $\frac{16}{15}\cdot\frac{13}{12} < 2$ imply that

$$\mathbb{E}[(\mathrm{I} - \alpha\mathbf{B})^p] \preceq \mathrm{I} - (1/2)\alpha p(1-\gamma)\Sigma_\varphi \,,$$

and the statement follows. ■

Now we provide a technical lemma on the behaviour of the symmetrized random matrix update $(\mathrm{I} - \alpha\mathbf{A})^\top(\mathrm{I} - \alpha\mathbf{A})$, where $\mathbf{A}$ is defined in (15). This lemma generalize the results presented in (Patil et al., 2023, Lemma 5).

Lemma 5. *For the random matrix $\mathbf{A}$ defined in (15) and $\mathbf{B} = \mathbf{A} + \mathbf{A}^\top - \alpha\mathbf{A}^\top\mathbf{A}$, $p \in \mathbb{N}$ and step size $\alpha \in (0; \frac{1-\gamma}{(1+\gamma)^2}]$ it holds that*

$$\begin{aligned}\mathbb{E}[\mathbf{B}] &\succeq (1-\gamma)\Sigma_\varphi \,,\\ \mathbb{E}[\mathbf{B}^p] &\preceq \frac{13}{12}\cdot 4^p\Sigma_\varphi \,.\end{aligned} \tag{36}$$

Proof With the definition of $\mathbf{A}$, we get that

$$\begin{aligned}\mathbf{A} + \mathbf{A}^\top &= \varphi(s)\{\varphi(s) - \gamma\varphi(s')\}^\top + \{\varphi(s) - \gamma\varphi(s')\}\varphi(s)^\top\\ &= 2\varphi(s)\varphi(s)^\top - \gamma\{\varphi(s)\varphi(s')^\top + \varphi(s')\varphi(s)^\top\}\\ &\succeq (2-\gamma)\varphi(s)\varphi(s)^\top - \gamma\varphi(s')\varphi(s')^\top \,,\end{aligned} \tag{37}$$

where we used an elementary inequality $uv^\top + vu^\top \preceq (uu^\top + vv^\top)$ valid for any $u, v \in \mathbb{R}^d$. Similarly, with elementary algebra, we obtain

$$\begin{aligned}\mathbf{A}^\top\mathbf{A} &= \{\varphi(s) - \gamma\varphi(s')\}\varphi(s)^\top\varphi(s)\{\varphi(s) - \gamma\varphi(s')\}^\top\\ &= \|\varphi(s)\|^2\{\varphi(s) - \gamma\varphi(s')\}\{\varphi(s) - \gamma\varphi(s')\}^\top\\ &= \|\varphi(s)\|^2\{\varphi(s)\varphi(s)^\top + \gamma^2\varphi(s')\varphi(s')^\top - \gamma(\varphi(s)\varphi(s')^\top + \varphi(s')\varphi(s)^\top)\}\\ &\overset{(a)}{\preceq} (1+\gamma)\varphi(s)\varphi(s)^\top + \gamma(1+\gamma)\varphi(s')\varphi(s')^\top \,,\end{aligned} \tag{38}$$

where in (a) we additionally used that $\|\varphi(s)\| \leq 1$ and

$$-(uu^\top + vv^\top) \preceq uv^\top + vu^\top \preceq (uu^\top + vv^\top)$$

for any $u, v \in \mathbb{R}^d$. Combining the bounds above yields that for $0 \leq \alpha \leq \frac{1-\gamma}{(1+\gamma)^2}$ it holds that

$$\mathbb{E}[\mathbf{B}] \succeq 2(1-\gamma)\Sigma_\varphi - \alpha(1+\gamma)^2\Sigma_\varphi \succeq (1-\gamma)\Sigma_\varphi \,,$$

and the first part of (36) is proved. To prove the second part it remains to notice that, for $p \in \mathbb{N}$, and $0 \leq \alpha \leq \frac{1-\gamma}{(1+\gamma)^2}$, it holds that

$$\mathbf{B}^p = \mathbf{B}^{p-2}\mathbf{B}^2 \preceq \|\mathbf{B}\|^{p-2}\mathbf{B}^2 \,,$$

and

$$\|\mathbf{B}\| = \|\mathbf{A} + \mathbf{A}^\top - \alpha\mathbf{A}^\top\mathbf{A}\| \leq 2(1+\gamma) + \alpha(1+\gamma)^2 \leq 3 + \gamma \leq 4 \,.$$

Now it remains to analyze the expectation of the matrix $\mathbf{B}^2$, which is symmetric and positive semi-definite:

$$\begin{aligned}\mathbf{B}^2 &= (\mathbf{A} + \mathbf{A}^\top - \alpha\mathbf{A}^\top\mathbf{A})(\mathbf{A} + \mathbf{A}^\top - \alpha\mathbf{A}^\top\mathbf{A}) \\ &= (\mathbf{A} + \mathbf{A}^\top)^2 - \alpha\left[(\mathbf{A} + \mathbf{A}^\top)\mathbf{A}^\top\mathbf{A} + \mathbf{A}^\top\mathbf{A}(\mathbf{A} + \mathbf{A}^\top)\right] + \alpha^2(\mathbf{A}^\top\mathbf{A})^2.\end{aligned}$$

We start from the first term. With the simple algebra and the definition of $\mathbf{A}$, we obtain that

$$\begin{aligned}\mathbf{A}^2 &= \varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top\varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top = \langle\varphi(s), \varphi(s) - \gamma\varphi(s')\rangle\mathbf{A} \,, \\ (\mathbf{A}^\top)^2 &= (\varphi(s) - \gamma\varphi(s'))\varphi(s)^\top(\varphi(s) - \gamma\varphi(s'))\varphi(s)^\top = \langle\varphi(s), \varphi(s) - \gamma\varphi(s')\rangle\mathbf{A}^\top \,, \\ \mathbf{A}\mathbf{A}^\top &= \varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top(\varphi(s) - \gamma\varphi(s'))\varphi(s)^\top = \|\varphi(s) - \gamma\varphi(s')\|^2\phi(s)\phi(s)^\top \,, \\ \mathbf{A}^\top\mathbf{A} &= (\varphi(s) - \gamma\varphi(s'))\varphi(s)^\top\varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top \\ &= \|\varphi(s)\|^2(\varphi(s) - \gamma\varphi(s'))(\varphi(s) - \gamma\varphi(s'))^\top \,.\end{aligned}$$

Additionally, in expectation we have the following relations, that follows from (37) and (38)

$$\mathbb{E}\left[\mathbf{A} + \mathbf{A}^\top\right] \preceq 2(1+\gamma)\Sigma_\varphi, \quad \mathbb{E}\left[\mathbf{A}^\top\mathbf{A}\right] \leq (1+\gamma)^2\Sigma_\varphi.$$

Using this relations, and using the fact that

$$(\mathbf{A} + \mathbf{A}^\top)^2 = \mathbf{A}^2 + \mathbf{A}\mathbf{A}^\top + \mathbf{A}^\top\mathbf{A} + (\mathbf{A}^\top)^2 \,,$$

we obtain that

$$\mathbb{E}\left[(\mathbf{A} + \mathbf{A}^\top)^2\right] \preceq 4(1+\gamma)^2\Sigma_\varphi.$$

For the term $(\mathbf{A}^\top\mathbf{A})^2$ we obtain that

$$\begin{aligned}(\mathbf{A}^\top\mathbf{A})^2 = \mathbf{A}^\top\left(\mathbf{A}\mathbf{A}^\top\right)\mathbf{A} &= \|\varphi(s) - \gamma\varphi(s')\|^2\mathbf{A}^\top\phi(s)\phi(s)^\top\mathbf{A} \\ &\overset{(a)}{=} \|\varphi(s) - \gamma\varphi(s')\|^2\|\varphi(s)\|^2\mathbf{A}^\top\mathbf{A} \,.\end{aligned}$$

Here the last identity (a) follows from the particular form of the matrix $\mathbf{A} = \varphi(s)\{\varphi(s) - \gamma\varphi(s')\}^\top$. Therefore, using the representation (38), we obtain that

$$\mathbb{E}\left[(\mathbf{A}^\top\mathbf{A})^2\right] \preceq (1+\gamma)^4\Sigma_\varphi \,.$$

The above bounds together with the Cauchy-Schwartz inequality imply that, for any $\epsilon > 0$, we have

$$\mathbf{B}^2 \preceq (1+\epsilon)(\mathbf{A}+\mathbf{A}^\top)^2 + (1+1/\epsilon)\alpha^2(\mathbf{A}^\top\mathbf{A})^2 \,,$$

which implies that

$$\mathbb{E}[\mathbf{B}^2] \preceq 4(1+\epsilon)(1+\gamma)^2\Sigma_\varphi + (1+1/\epsilon)\alpha^2(1+\gamma)^4\Sigma_\varphi \,.$$

Thus, setting $\epsilon = 1/12$, and using that $\gamma \in [0;1]$, we get that

$$\mathbb{E}[\mathbf{B}^2] \preceq (52/3)\Sigma_\varphi = (13/12)16\Sigma_\varphi \,.$$

As a result

$$\mathbb{E}[\mathbf{B}^2] \preceq (13/12)16\Sigma_\varphi \Rightarrow \mathbb{E}[\mathbf{B}^p] \preceq 4^{p-2}\mathbb{E}[\mathbf{B}^2] \preceq (13/12)4^p\Sigma_\varphi.$$

■

B.3. Missing results from Section 3

We begin this section from instantiating Theorem 3 for the sequence $\{\theta_k\}$ which corresponds to TD(0) algorithm. We use that $\varkappa_2 = 1$, $\mathrm{Tr}(\Sigma_\varepsilon) \le 1 + \|\theta_\star\|^2_{\Sigma_\varphi}$, $a = \lambda_{\min}(1-\gamma)/2$. Then we get

Theorem 11 *Assume **TD** 1 and **TD** 2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $n \ge 2$, $\alpha \in \left(0; \frac{1-\gamma}{256}\right]$, and $\theta_0 \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}^{1/2}[\|\bar\theta_n - \theta_\star\|^2_{\Sigma_\varphi}] \lesssim \frac{\|\theta_\star\|_{\Sigma_\varphi}+1}{\sqrt{\lambda_{\min} n(1-\gamma)}}\left(1 + \frac{\sqrt{\alpha}}{\sqrt{(1-\gamma)\lambda_{\min}}}\right) + \frac{\|\theta_\star\|_{\Sigma_\varphi}+1}{\sqrt{\alpha}(1-\gamma)^{3/2}\lambda_{\min} n}$$
$$+ \left(\frac{1}{\alpha n(1-\gamma)\lambda_{\min}^{1/2}} + \frac{1}{\sqrt{\alpha} n(1-\gamma)^{3/2}\lambda_{\min}}\right)\left(1 - \frac{\alpha(1-\gamma)\lambda_{\min}}{2}\right)^{n/2}\|\theta_0 - \theta_\star\| \,.$$

Similarly to the discussion above, we can state the respective p-moment bound for the case of TD(0) algorithm. This theorem is an adaptation of Theorem 2 (see also Theorem 9).

Theorem 12 *Assume **TD** 1 and **TD** 2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $p \ge 2$, $n \ge 2$, and step size*

$$\alpha \in \left(0; \frac{1-\gamma}{128(p+\log n)}\right] \,,$$

we have that

$$\mathbb{E}^{1/p}[\|(\bar\theta_n - \theta_\star)\|^p_{\Sigma_\varphi}] \lesssim \frac{p^{1/2}(\|\theta_\star\|_{\Sigma_\varphi}+1)}{n^{1/2}(1-\gamma)\lambda_{\min}^{1/2}}\left(1 + \frac{\sqrt{\alpha p}+\alpha p}{\sqrt{(1-\gamma)\lambda_{\min}}}\right) + \frac{p(\|\theta_\star\|_{\Sigma_\varphi}+1)}{n(1-\gamma)^{3/2}\lambda_{\min}}\left(1 + \frac{1}{\sqrt{\alpha p}}\right)$$
$$+ \left(1 - \frac{\alpha(1-\gamma)\lambda_{\min}}{2}\right)^{n/2}\left((p+\log(n))^{1/2} + \frac{p}{\sqrt{\lambda_{\min}}}\right)\frac{\{p+\log(n)\}^{1/2}}{(1-\gamma)^2\sqrt{\lambda_{\min} n}}\|\theta_0 - \theta_\star\| \,.$$

The respective high-probability bound can be written as follows:

Corollary 4. *Assume **TD** 1 and **TD** 2. Let $\{\theta_k\}$ be a sequence of TD(0) updates generated by (14). Fix $\delta \in (0; 1/\mathrm{e})$. Then, for the step- and sample size*

$$\alpha = \frac{1-\gamma}{128 \log(n/\delta)}, \quad n \geq \frac{\log(1/\delta)}{(1-\gamma)^2}$$

it holds with probability at least $1-\delta$ that

$$\|\bar{\theta}_n - \theta_\star\|_{\Sigma_\varphi} \lesssim \frac{(\|\theta_\star\|_{\Sigma_\varphi} + 1)\sqrt{\log(1/\delta)}}{n^{1/2}(1-\gamma)\lambda_{\min}} + \big(1 - \frac{(1-\gamma)^2\lambda_{\min}}{128\log(n/\delta)}\big)^{n/2} \frac{\|\theta_0 - \theta_\star\| \log^{3/2}(n/\delta)}{(1-\gamma)^2\lambda_{\min} n} .$$

B.4. Proof of stability bound (17) based on matrix stability argument

In the previous section we have presented a stability result Theorem 10, which allows for maximal step size in the constant-step size $TD(0)$ algorithm $\alpha_{\infty,p}$ of the form

$$\alpha_{\infty,p} = \frac{1-\gamma}{128p} .$$

In this subsection we show that such type of result can not be readily obtained from existing results on the stability of random matrix product Huang et al. (2021).

We first introduce some matrix notations. For the matrix $B \in \mathbb{R}^{d\times d}$ we denote by $(\sigma_\ell(B))_{\ell=1}^d$ its singular values. For $p \geq 1$, the Shatten p-norm is denoted by $\|B\|_p = \{\sum_{\ell=1}^d \sigma_\ell^p(B)\}^{1/p}$. For $p, q \geq 1$ and a random matrix $\mathbf{X}$ we write $\|\mathbf{X}\|_{p,q} = \{\mathbb{E}[\|\mathbf{X}\|_p^q]\}^{1/q}$. Then it is easily seen that

$$\mathbb{E}^{1/q}[\|\mathbf{X}\|^q] \leq \|\mathbf{X}\|_{q,p} ,$$

and one can control an operator norm of the matrix with its Shatten norm of an appropriate order. Now we state the following result from (Durmus et al., 2021a, Proposition 2).

Proposition 13 *Let $\{\mathbf{Y}_\ell\}_{\ell\in\mathbb{N}}$ be a sequence on independent matrices, $\mathbf{Y}_\ell \in \mathbb{R}^{d\times d}$ and Q be a positive definite matrix. Assume that for each $\ell \in \mathbb{N}$ there exist $m_\ell \in (0,1)$ and $\sigma_\ell > 0$ such that $\|\mathbb{E}[\mathbf{Y}_\ell]\|_{\mathrm{Q}}^2 \leq 1 - m_\ell$ and $\|\mathbf{Y}_\ell - \mathbb{E}[\mathbf{Y}_\ell]\|_{\mathrm{Q}} \leq \sigma_\ell$ almost surely. Define $\mathbf{Z}_n = \prod_{\ell=0}^n \mathbf{Y}_\ell = \mathbf{Y}_n\mathbf{Z}_{n-1}$, for $n \geq 1$ with some (deterministic) matrix $\mathbf{Z}_0 \in \mathbb{R}^{d\times d}$. Then, for any $2 \leq q \leq p$ and $n \geq 1$,*

$$\|\mathbf{Z}_n\|_{p,q}^2 \leq \kappa_{\mathrm{Q}} \prod_{\ell=1}^n (1 - m_\ell + (p-1)\sigma_\ell^2) \| \mathrm{Q}^{1/2} \mathbf{Z}_0 \mathrm{Q}^{-1/2} \|_{p,q}^2 , \tag{39}$$

where $\kappa_{\mathrm{Q}} = \lambda_{min}^{-1}(\mathrm{Q})\lambda_{max}(\mathrm{Q})$.

Note that the result of Theorem 13 is generic in a sense that it allows us an additional degree of freedom in the choice of the contracting matrix norm $\|\cdot\|_{\mathrm{Q}}$. An almost sure bound on $\|\mathbf{Y}_\ell - \mathbb{E}[\mathbf{Y}_\ell]\|_{\mathrm{Q}}$ can be generalized to a moment-type bound, with the same shape of the bound in (39). The main drawback of this technique is an inevitable trade-off between m_ℓ and $(p-1)\sigma_\ell^2$ factors, which directly influences the speed with which $\|\mathbf{Z}_n\|_{p,q}^2$ decays to 0.

Now we aim to apply Theorem 13 to check the assumption A2 for the TD(0) algorithm.

Lemma 6. *Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14) under **TD 1** and **TD 2**. Then this update scheme satisfies assumption A2(p) with*

$$a = \frac{(1-\gamma)\lambda_{\min}(\Sigma_\varphi)}{2}\,,\ \alpha_{p,\infty} = \frac{1-\gamma}{128p} \wedge \frac{(1-\gamma)\lambda_{\min}(\Sigma_\varphi)}{64p}\,,\quad \varkappa_p = d^{1/p}\,.$$

Proof We aim to apply here the result of Theorem 13 with $\mathbf{Y}_\ell = \mathrm{I} - \alpha\mathbf{A}_\ell$ and $\mathbf{Z}_n = \Gamma^{(\alpha)}_{1:n}$. Towards this aim, note that Lemma 4 implies that, with $\mathbf{A} = \varphi(s)\{\varphi(s) - \gamma\varphi(s')\}^\top$ being a random TD update matrix defined in (15), we have

$$\begin{aligned}\|\mathrm{I} - \alpha\bar{\mathbf{A}}\| = \|\mathbb{E}[\mathrm{I} - \alpha\mathbf{A}]\| &\le \sqrt{\|\mathbb{E}[(\mathrm{I} - \alpha\mathbf{A})^\top(\mathrm{I} - \alpha\mathbf{A})\}]\|} \\ &\le \sqrt{1 - \alpha(1-\gamma)\lambda_{\min}(\Sigma_\varphi)} \\ &\le 1 - (1/2)\alpha(1-\gamma)\lambda_{\min}(\Sigma_\varphi)\,,\end{aligned}$$

which holds for $\alpha \in (0; \frac{1-\gamma}{128})$. Moreover,

$$\|\alpha(\mathbf{A} - \bar{\mathbf{A}})\| \le \alpha\|\varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top\| + \alpha\|\mathbb{E}\left[\varphi(s)(\varphi(s) - \gamma\varphi(s'))^\top\right]\| \le 2(1+\gamma)\alpha \le 4\alpha\,.$$

Hence, setting $a = (1-\gamma)\lambda_{\min}(\Sigma_\varphi)$, the assumptions of Lemma 4 are satisfied with

$$\sigma_\ell = 4\alpha\,, m_\ell = \alpha a/2\,.$$

Hence, applying the result of Lemma 4 with $Q = \mathrm{I}$, $\mathbf{Z}_0 = \mathrm{I}$, we get

$$\mathbb{E}^{1/q}\left[\|\Gamma^{(\alpha)}_{1:n}\|^q\right] \le \|\Gamma^{(\alpha)}_{1:n}\|_{p,q} \le d^{1/p}(1 - \alpha a + 16(p-1)\alpha^2)^{n/2}\,.$$

Now we have to balance the terms $\alpha a/2$ and $16(p-1)\alpha^2$, which yields the scaling of α with a (and, hence, with $\lambda_{\min}(\Sigma_\varphi)$). In particular, setting $\alpha = \frac{a}{32p}$, we get the statement of the Lemma. ■

Appendix C. Proofs of Section 4

In this section we need to introduce an additional assumptions which relates matrices $\mathbf{G}$, $\bar{\mathbf{A}}$, and (random) matrices $\mathbf{A}_i$ for $i \in \{1, \dots, n\}$.

C 1 *There exist such symmetric positive-definite matrix $\mathbf{G} = \mathbf{G}^\top > 0$ and constants $\mathrm{g} > 0$, $\omega > 0$, $\varrho > 0$, such that*

(i) for the system matrix $\bar{\mathbf{A}}$ it holds that

$$\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-\top}\mathbf{G}\bar{\mathbf{A}}^{-1}\mathbf{G}^{1/2} \preceq \mathrm{g}^2\mathrm{I}\,;$$

(ii) for the random matrix $\mathbf{A}_1$ it holds that

$$\mathbb{E}[\mathbf{A}_1^\top\mathbf{G}^{-1}\mathbf{A}_1] \preceq \omega^2\mathbf{G}\,;$$

(iii) for the matrix Σ_ε defined in (3) it holds that

$$\operatorname{Tr}(\Sigma_\varepsilon) \leq \varrho^2 \operatorname{Tr}(\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon\bar{\mathbf{A}}^{-T}\mathbf{G}^{1/2}) ;$$

Under Assumption **C** 1 we introduce a new notation

$$\Sigma_\varepsilon^{(tr)} = \mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon\bar{\mathbf{A}}^{-T}\mathbf{G}^{1/2} .$$

Our proof in this section follows the general procedure introduced for the Polyak-Ruppert estimator $\bar{\theta}_{n_0,n}$ in (29). Recall that with summation by parts we obtain the following

$$\bar{\mathbf{A}}\left(\bar{\theta}_{n_0,n} - \theta_\star\right) = \frac{\theta_{n_0} - \theta_n}{\alpha(n-n_0)} - \frac{\sum_{t=n_0}^{n-1} e\left(\theta_t, Z_{t+1}\right)}{n-n_0} ,$$

where the quantities $e\left(\theta_t, Z_{t+1}\right)$ are defined in (30). Since we assume that $\bar{\mathbf{A}}$ is non-degenerate, for symmetric positive-definite matrix $\mathbf{G} = \mathbf{G}^\top > 0$ from **C** 1, we get from the previous inequality that

$$\mathbf{G}^{1/2}\left(\bar{\theta}_{n_0,n} - \theta_\star\right) = \frac{\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}(\theta_{n_0} - \theta_n)}{\alpha(n-n_0)} - \frac{\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\sum_{t=n_0}^{n-1} e\left(\theta_t, Z_{t+1}\right)}{n-n_0} . \tag{40}$$

Based on the above identity, we prove the following counterpart of the 2-nd-moment bound Theorem 1 for the general LSA problem.

Theorem 14 *Assume A1, A2(2), and **C** 1. Then for any $n \geq 2$, $\alpha \in (0; \alpha_{2,\infty}]$, it holds that*

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_n - \theta_\star\|_{\mathbf{G}}^2] \lesssim \frac{\operatorname{Tr}(\Sigma_\varepsilon^{(tr)})}{n} + \frac{\varkappa_2^2 \mathsf{g}^2 \operatorname{Tr}(\Sigma_\varepsilon)}{an}\left(\frac{\|\mathbf{G}^{-1/2}\|^2}{\alpha n} + \omega^2\|\mathbf{G}^{1/2}\|^2\alpha\right) \\ + \varkappa_2^2\mathsf{g}^2(1-\alpha a)^n\left(\frac{\|\mathbf{G}^{-1/2}\|^2}{\alpha^2 n^2} + \frac{\omega^2\|\mathbf{G}^{1/2}\|^2}{\alpha a n^2}\right)\|\theta_0 - \theta_\star\|^2\end{aligned}$$

Proof Following the pipeline of Theorem 1 and using (40), we get

$$\mathbb{E}[\|\bar{\theta}_n - \theta_\star\|_{\mathbf{G}}^2] \lesssim \underbrace{\frac{\mathbb{E}[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}(\theta_{n/2} - \theta_n)\|^2]}{\alpha^2 n^2}}_{T_1} + \underbrace{\frac{\mathbb{E}[\|\sum_{t=n/2}^{n-1}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}e\left(\theta_t, Z_{t+1}\right)\|^2]}{n^2}}_{T_2} ,$$

and estimate the terms T_1 and T_2 separately. Applying the bounds of Theorem 7, we get first that

$$T_1 \lesssim \mathsf{g}^2\|\mathbf{G}^{-1/2}\|^2\varkappa_2^2\left[\frac{(1-\alpha a)^n\|\theta_0 - \theta_\star\|^2}{\alpha^2 n^2} + \frac{\operatorname{Tr}(\Sigma_\varepsilon)}{\alpha a n^2}\right] .$$

Here we additionally used an upper bound

$$\begin{aligned}\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}u\|^2 &= u^\top\bar{\mathbf{A}}^{-\top}\mathbf{G}\bar{\mathbf{A}}^{-1}u \\ &= u^\top\mathbf{G}^{-1/2}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-\top}\mathbf{G}\bar{\mathbf{A}}^{-1}\mathbf{G}^{1/2}\mathbf{G}^{-1/2}u \\ &\leq \mathsf{g}^2 u^\top\mathbf{G}^{-1}u \\ &\leq \mathsf{g}^2\|\mathbf{G}^{-1/2}\|^2\|u\|^2 ,\end{aligned} \tag{41}$$

which is valid for any $u \in \mathbb{R}^d$. Similarly, since $\{\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}e\left(\theta_t, Z_{t+1}\right)\}_{t\in\mathbb{N}}$ is a martingale-difference sequence w.r.t. filtration $\mathcal{F}_k = \sigma(Z_j, j \le k)$, we get the following bound for T_2:

$$
\begin{aligned}
T_2 &\lesssim n^{-2}\sum_{t=n/2}^{n-1}\mathbb{E}[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}e\left(\theta_t, Z_{t+1}\right)\|^2] \\
&\lesssim \frac{\operatorname{Tr}(\Sigma_\varepsilon^{(tr)})}{n} + \varkappa_2^2\omega^2\mathrm{g}^2\,\|\mathbf{G}^{1/2}\|^2\left[\frac{(1-\alpha a)^n\|\theta_0-\theta_\star\|^2}{\alpha a n^2} + \frac{\alpha\operatorname{Tr}(\Sigma_\varepsilon)}{an}\right].
\end{aligned}
$$

In particular, to bound the first term we use the bound

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{F}_t}\left[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^2\right] \qquad (42)\\
&\quad= \mathbb{E}^{\mathcal{F}_t}\left[(\theta_t-\theta_\star)^\top\tilde{\mathbf{A}}_{t+1}^\top\bar{\mathbf{A}}^{-\top}\mathbf{G}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\right]\\
&\quad= \mathbb{E}^{\mathcal{F}_t}\left[(\theta_t-\theta_\star)^\top\mathbf{A}_{t+1}^\top\mathbf{G}^{-1/2}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-\top}\mathbf{G}\bar{\mathbf{A}}^{-1}\mathbf{G}^{1/2}\mathbf{G}^{-1/2}\mathbf{A}_{t+1}(\theta_t-\theta_\star)\right]\\
&\qquad\qquad - \mathbb{E}^{\mathcal{F}_t}\left[(\theta_t-\theta_\star)^\top\mathbf{G}(\theta_t-\theta_\star)\right]\\
&\quad\le \mathrm{g}^2\mathbb{E}^{\mathcal{F}_t}\left[(\theta_t-\theta_\star)^\top\mathbf{A}_{t+1}^\top\mathbf{G}^{-1}\mathbf{A}_{t+1}(\theta_t-\theta_\star)\right]\\
&\quad\le \omega^2\mathrm{g}^2(\theta_t-\theta_\star)^\top\mathbf{G}(\theta_t-\theta_\star)\\
&\quad\le \omega^2\mathrm{g}^2\,\|\mathbf{G}^{1/2}\|^2\,\|\theta_t-\theta_\star\|^2\,.
\end{aligned}
$$

■

Now we trace Theorem 16 in the case of TD (0) updates. First we check whether the assumption **C** 1 holds.

Lemma 7. *Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14) under **TD** 1 and **TD** 2. Then this update scheme satisfies assumption **C** 1 with*

$$
\mathbf{G} = \Sigma_\varphi\,, \quad \mathrm{g} = 1/(1-\gamma)\,, \quad \omega = (1+\gamma)\lambda_{\min}^{-1/2}\,, \quad \varrho = 1+\gamma\,.
$$

Moreover, it holds that

$$
\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2} \succeq (1-\gamma)^2\mathrm{I}\,. \qquad (43)
$$

Proof In order to prove that

$$
\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-\top}\Sigma_\varphi\bar{\mathbf{A}}^{-1}\Sigma_\varphi^{1/2} \preceq \frac{1}{(1-\gamma)^2}\mathrm{I}\,,
$$

it is enough to show the lower bound (43). For the finite state space $\mathcal{S}$ this follows from (Li et al., 2024, Lemma 5), we provide a slightly modified argument for completeness. Indeed, for any $x \in \mathbb{R}^d$, using that $\bar{\mathbf{A}} = \Sigma_\varphi - \gamma\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]$, we have

$$
\begin{aligned}
x^\top\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}x &= \|\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}x\|^2 = \|\big(\mathrm{I}-\gamma\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]\Sigma_\varphi^{-1/2}\big)x\|^2\\
&\ge \left(\|x\| - \gamma\|\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]\Sigma_\varphi^{-1/2}x\|\right)^2\,,
\end{aligned}
$$

and to complete the proof it is enough to show that $\|\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]\Sigma_\varphi^{-1/2}\| \leq 1$. In order to do it, note that

$$\begin{aligned}
\|\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]\Sigma_\varphi^{-1/2}\| &= \sup_{\|x\|=1,\|y\|=1} x^\top \Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]\Sigma_\varphi^{-1/2} y \\
&= \sup_{\|x\|=1,\|y\|=1} \mathbb{E}\left[\left([\Sigma_\varphi^{-1/2}x]^\top\varphi(s_1)\right)\left(\varphi(s_1')^\top\Sigma_\varphi^{-1/2}y\right)\right] \\
&\leq \sup_{\|x\|=1,\|y\|=1} \mathbb{E}\left[\frac{1}{2}\left([\Sigma_\varphi^{-1/2}x]^\top\varphi(s_1)\right)^2 + \frac{1}{2}\left(\varphi(s_1')^\top\Sigma_\varphi^{-1/2}y\right)^2\right] \\
&= \sup_{\|x\|=1,\|y\|=1}\Bigg[\frac{1}{2}x^\top\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1)^\top]\Sigma_\varphi^{-1/2}x \\
&\qquad + \frac{1}{2}y^\top\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1')\varphi(s_1')^\top]\Sigma_\varphi^{-1/2}y\Bigg] = 1.
\end{aligned}$$

where we used the fact that a distribution μ is the invariant. Hence, we get

$$x^\top\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}x \geq (1-\gamma)^2\|x\|^2 ,$$

and the bound (43) is proved. To prove the second part of the bound, we use (38) and obtain that

$$\begin{aligned}
\mathbb{E}[\mathbf{A}_1^\top\Sigma_\varphi^{-1}\mathbf{A}_1] &= \mathbb{E}[\big(\varphi(s_1)-\gamma\varphi(s_1')\big)\varphi(s_1)^\top\Sigma_\varphi^{-1}\varphi(s_1)\big(\varphi(s_1)-\gamma\varphi(s_1')\big)^\top] \\
&\preceq \lambda_{\min}^{-1}\mathbb{E}[\big(\varphi(s_1)-\gamma\varphi(s_1')\big)\big(\varphi(s_1)-\gamma\varphi(s_1')\big)^\top] \preceq \lambda_{\min}^{-1}(1+\gamma)^2\Sigma_\varphi ,
\end{aligned}$$

where the last inequality follows (38) in the proof of Lemma 5. To check the last one, note that

$$\begin{aligned}
\mathrm{Tr}(\Sigma_\varepsilon) &= \mathrm{Tr}(\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1/2}\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-\top}\Sigma_\varepsilon\bar{\mathbf{A}}^{-1}\Sigma_\varphi^{1/2}\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}) \\
&\overset{(a)}{=} \mathrm{Tr}(\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}\Sigma_\varphi\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1/2}\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-\top}\Sigma_\varepsilon\bar{\mathbf{A}}^{-1}\Sigma_\varphi^{1/2}) \\
&\overset{(b)}{\leq} \|\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}\Sigma_\varphi\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1/2}\|\,\mathrm{Tr}(\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-\top}\Sigma_\varepsilon\bar{\mathbf{A}}^{-1}\Sigma_\varphi^{1/2}) .
\end{aligned}$$

Note that the identity (a) above follows from the cyclic property of trace, and the inequality (b) is due to $\mathrm{Tr}(CD) \leq \|C\|\,\mathrm{Tr}(D)$, which is valid for symmetric positive semi-definite matrices C, D. In the bound above it remains to estimate

$$\|\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}\Sigma_\varphi\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1/2}\| \leq \|\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}\|^2\|\Sigma_\varphi\| . \tag{44}$$

Consider now the operator norm of the matrix $\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}$. Note that $\bar{\mathbf{A}} = \Sigma_\varphi - \gamma\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top]$. Thus, we get

$$\begin{aligned}
&\|\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}\| = \sup_{\|x\|=1,\|y\|=1} x^\top\Sigma_\varphi^{-1/2}(\Sigma_\varphi - \gamma\mathbb{E}[\varphi(s_1)\varphi(s_1')^\top])\Sigma_\varphi^{-1/2}y \\
&\leq 1 + \gamma \sup_{\|x\|=1,\|y\|=1}\left[\frac{x^\top\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1)\varphi(s_1)^\top]\Sigma_\varphi^{-1/2}x}{2} + \frac{y^\top\Sigma_\varphi^{-1/2}\mathbb{E}[\varphi(s_1')\varphi(s_1')^\top]\Sigma_\varphi^{-1/2}y}{2}\right] \\
&= 1+\gamma .
\end{aligned}$$

Plugging this inequality into (44), we get

$$\|\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}\Sigma_\varphi^{-1/2}\Sigma_\varphi\Sigma_\varphi^{-1/2}\bar{\mathbf{A}}^\top\Sigma_\varphi^{-1/2}\| \le (1+\gamma)^2 \,.$$

In the last bound we additionally used that $\|\Sigma_\varphi\| \le 1$ under **TD** 2. ■

Now a simple combination of the above bounds allows us to prove the following result:

Theorem 15 *Assume **TD** 1 and **TD** 2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $p \ge 2$, $n \ge 2$, $\alpha \in \left(0; \frac{1-\gamma}{256}\right]$, it holds that*

$$\begin{aligned}\mathbb{E}[\|\bar\theta_n - \theta_\star\|^2_{\Sigma_\varphi}] \lesssim \frac{\mathrm{Tr}(\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon^{(TD)}\bar{\mathbf{A}}^{-T}\Sigma_\varphi^{1/2})}{n} + \frac{1+\|\theta_\star\|^2_{\Sigma_\varphi}}{(1-\gamma)^3\lambda_{\min}^2 n}\left(\frac{1}{\alpha n}+\alpha\right) \\ + \frac{(1-\alpha(1-\gamma)\lambda_{\min}/2)^n}{\lambda_{\min}(1-\gamma)^2}\left(\frac{1}{\alpha^2 n^2} + \frac{1}{\alpha(1-\gamma)\lambda_{\min} n^2}\right)\|\theta_0-\theta_\star\|^2\end{aligned} \tag{45}$$

Based on the identity above, we can prove the following counterpart of the result Theorem 9 for the general LSA problem.

Theorem 16 *Assume A1, A2(∞), and **C** 1. Then for any $p \ge 2$, $n \ge 2$, $\alpha \in [0; \alpha_{p+\log n,\infty})$, it holds that*

$$\begin{aligned}\mathbb{E}^{1/p}[\|\bar\theta_n - \theta_\star\|^p_{\mathbf{G}}] \lesssim \frac{p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})}}{n^{1/2}}\left(1 + \frac{\sqrt{\alpha p}\varkappa_\infty \mathrm{g}\varrho\|\mathbf{G}^{1/2}\|\,\mathrm{C}_{\mathbf{A}}}{\sqrt{a}} + \frac{\alpha p\varkappa_\infty \mathrm{g}\|\mathbf{G}^{1/2}\|\|\varepsilon\|_\infty}{\sqrt{\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})}}\right) \\ + \frac{p^{1/2}\mathrm{g}\varrho\|\mathbf{G}^{-1/2}\|\varkappa_\infty\sqrt{\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})}}{\sqrt{a}n}\left[\frac{1}{\sqrt{\alpha}} + p^{1/2}\,\mathrm{C}_{\mathbf{A}}\sqrt{\alpha(p+\log n)}\right] \\ + \frac{p\mathrm{g}\|\mathbf{G}^{-1/2}\|\varkappa_\infty\|\varepsilon\|_\infty}{n}\left(1 + \mathrm{C}_{\mathbf{A}}\,\alpha(p+\log n)\right) \\ + \mathrm{g}\|\mathbf{G}^{-1/2}\|\varkappa_\infty(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\|\left(\frac{1}{\alpha n} + \frac{p\,\mathrm{C}_{\mathbf{A}}}{n} + \frac{p^{1/2}\omega}{\sqrt{\alpha a}n}\right).\end{aligned}$$

Proof The proof follows the general scheme of Theorem 9. Setting $n_0 = n/2$ and using Minkowski's inequality, we obtain from (40) that

$$\begin{aligned}\mathbb{E}^{1/p}\left[\|\bar\theta_n - \theta_\star\|^p_{\mathbf{G}}\right] \le \underbrace{\frac{\mathbb{E}^{1/p}[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}(\theta_{n/2}-\theta_n)\|^p]}{\alpha n}}_{T_1} + \\ \underbrace{\frac{\mathbb{E}^{1/p}\left[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\sum_{t=n/2}^{n-1}\mathrm{e}(\theta_t, Z_{t+1})\|^p\right]}{n}}_{T_2},\end{aligned} \tag{46}$$

and bound T_1, T_2 separately. We begin with bounding the term T_1, which is a remainder term (w.r.t. sample size n). With Theorem 7-(26), **C** 1, and (41), we obtain

$$T_1 \lesssim \mathrm{g}\|\mathbf{G}^{-1/2}\|\varkappa_\infty\left[\frac{(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\|}{\alpha n} + \frac{p^{1/2}\varrho\sqrt{\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})}}{\sqrt{\alpha a}n} + \frac{p\|\varepsilon\|_\infty}{n}\right].$$

Now we bound T_2. Using again Minkowski's inequality, we get

$$T_2 \leq n^{-1}\,\mathbb{E}^{1/p}\big[\|\textstyle\sum_{t=n/2}^{n-1}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\varepsilon_{t+1}\|^p\big] + n^{-1}\,\mathbb{E}^{1/p}[\|\textstyle\sum_{t=n/2}^{n-1}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p]\,.$$

The first term of the above sum can be controlled by directly applying Pinelis' version of Rosenthal's inequality (Pinelis, 1994, Theorem 4.3):

$$\begin{aligned}\mathbb{E}^{1/p}\left[\left\|\sum_{t=n/2}^{n-1}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\varepsilon_{t+1}\right\|^p\right] &\lesssim p^{1/2}n^{1/2}\sqrt{\mathrm{Tr}(\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon\bar{\mathbf{A}}^{-T}\mathbf{G}^{1/2})} + p\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\varepsilon\|_\infty \\ &\overset{(a)}{\lesssim} p^{1/2}n^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})} + p\mathsf{g}\|\mathbf{G}^{-1/2}\|\,\|\varepsilon\|_\infty\,.\end{aligned}$$

In order to prove the step (a) above we used the bound (41). Hence it remains to bound the quantity

$$\mathbb{E}^{1/p}[\|\sum_{t=n/2}^{n-1}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p]\,.$$

Note that $\{\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\}$ is a martingale-difference w.r.t. $\mathcal{F}_t=\sigma(Z_k,k\leq t)$. A further application of Rosenthal's inequality thus shows that

$$\begin{aligned}\mathbb{E}^{1/p}\left[\left\|\sum_{t=n/2}^{n-1}\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\right\|^p\right] &\lesssim p^{1/2}\mathbb{E}^{1/p}\left[\Big(\sum_{t=n/2}^{n-1}\mathbb{E}^{\mathcal{F}_t}\left[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^2\right]\Big)^{p/2}\right] \\ &\quad + p\,\mathbb{E}^{1/p}\left[\max_t\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p\right]\,. \quad (47)\end{aligned}$$

Now we upper bound both terms in the r.h.s. separately. Using the bound (42), for the first term in r.h.s. of (47) we have, using Theorem 7-(26) and **C** 1, that

$$\begin{aligned}&p^{1/2}\mathbb{E}^{1/p}\left[\Big(\sum_{t=n/2}^{n-1}\mathbb{E}^{\mathcal{F}_t}\left[\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^2\right]\Big)^{p/2}\right] \leq p^{1/2}\omega\mathsf{g}\|\mathbf{G}^{1/2}\|\left[\sum_{t=n/2}^{n-1}\mathbb{E}^{2/p}\left[\|\theta_t-\theta_\star\|^p\right]\right]^{1/2} \\ &\quad\lesssim \varkappa_\infty p^{1/2}\omega\mathsf{g}\|\mathbf{G}^{1/2}\|\cdot\left[\frac{(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\|}{\sqrt{\alpha a}} + \frac{p^{1/2}\varrho\sqrt{\alpha n\,\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})}}{\sqrt{a}} + \alpha p n^{1/2}\|\varepsilon\|_\infty\right]\,.\end{aligned}$$

For the second term in (47) we have, applying Theorem 7-(26) and using $n^{1/\log n}\leq \mathrm{e}$, that

$$\begin{aligned}&p\,\mathbb{E}^{1/p}[\max_t\|\mathbf{G}^{1/2}\bar{\mathbf{A}}^{-1}\tilde{\mathbf{A}}_{t+1}(\theta_t-\theta_\star)\|^p] \\ &\quad\lesssim p\,\mathrm{C}_{\mathbf{A}}\,\mathsf{g}\|\mathbf{G}^{-1/2}\|n^{1/(p+\log n)}\max_{n/2\leq t<n}\mathbb{E}^{1/(p+\log n)}[\|\theta_t-\theta_\star\|^{p+\log n}] \\ &\quad\lesssim \varkappa_\infty p\,\mathrm{C}_{\mathbf{A}}\,\mathsf{g}\|\mathbf{G}^{-1/2}\|\cdot\left[(1-\alpha a)^{n/2}\|\theta_0-\theta_\star\| + \frac{\sqrt{\alpha(p+\log n)\,\mathrm{Tr}(\Sigma_\varepsilon^{(tr)})}}{a} + \alpha(p+\log n)\|\varepsilon\|_\infty\right]\,.\end{aligned}$$

Now the statement follows from combining the above estimates in (46). ■

Now a simple combination of the above bounds allows us to prove the following bound:

Theorem 17 *Assume* ***TD*** *1 and* ***TD*** *2. Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (14). Then for any $p \geq 2$, $n \geq 2$, $\alpha \in \big(0; \frac{1-\gamma}{128(p+\log n)}\big]$, it holds that*

$$\mathbb{E}^{1/p}[\|(\bar{\theta}_n - \theta_\star)\|^p_{\Sigma_\varphi}] \lesssim \frac{p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon^{(TD)}\bar{\mathbf{A}}^{-T}\Sigma_\varphi^{1/2})}}{n^{1/2}}\left(1 + \frac{\sqrt{\alpha p}}{(1-\gamma)^{3/2}\lambda_{\min}}\right) + \frac{\alpha p^{3/2}(1+\|\theta_\star\|)}{(1-\gamma)\lambda_{\min}^{1/2}n^{1/2}}$$
$$+ \frac{p^{1/2}\sqrt{\mathrm{Tr}(\Sigma_\varphi^{1/2}\bar{\mathbf{A}}^{-1}\Sigma_\varepsilon^{(TD)}\bar{\mathbf{A}}^{-T}\Sigma_\varphi^{1/2})}}{(1-\gamma)^{3/2}\lambda_{\min}n}\left[\frac{1}{\sqrt{\alpha}} + p^{1/2}\sqrt{\alpha(p+\log n)}\right] + \frac{p(1+\|\theta_\star\|)}{(1-\gamma)\lambda_{\min}n}\left[1 + \alpha(p+\log n)\right]$$
$$+ \frac{1}{(1-\gamma)\lambda_{\min}^{1/2}}\Big(1-\alpha(1-\gamma)\lambda_{\min}\Big)^{n/2}\|\theta_0 - \theta_\star\|\left(\frac{1}{\alpha n} + \frac{p}{n} + \frac{p^{1/2}}{\sqrt{\alpha}(1-\gamma)\lambda_{\min}^{1/2}n}\right) .$$

Appendix D. Berbee's lemma and coupling inequalities for Markov chains

We preface this section with some definitions essential for the Berbee's lemma construction. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with σ-fields $\mathfrak{F}$ and $\mathfrak{G}$ such that $\mathfrak{F} \subseteq \mathcal{F}$, $\mathfrak{G} \subseteq \mathcal{F}$. Then the β-mixing coefficient of $\mathfrak{F}$ and $\mathfrak{G}$ is defined as

$$\beta(\mathfrak{F}, \mathfrak{G}) = (1/2)\sup\sum_{i\in\mathsf{I}}\sum_{j\in\mathsf{J}}|\mathbb{P}(\mathsf{A}_i \cap \mathsf{B}_j) - \mathbb{P}(\mathsf{A}_i)\mathbb{P}(\mathsf{B}_j)| ,$$

and the supremum is taken over all pairs of partitions $\{\mathsf{A}_i\}_{i\in\mathsf{I}} \in \mathfrak{F}^{\mathsf{I}}$ and $\{\mathsf{B}_j\}_{j\in\mathsf{J}} \in \mathfrak{G}^{\mathsf{J}}$ of $\tilde{\mathsf{Z}}_{\mathbb{N}}$ with finite I and J.

Now let $(\mathsf{Z}, \mathsf{d}_\mathsf{Z})$ be a Polish space endowed with its Borel σ-field, denoted by $\mathcal{Z}$, and let $(\mathsf{Z}^{\mathbb{N}}, \mathcal{Z}^{\otimes\mathbb{N}})$ be the corresponding canonical space. Consider a Markov kernel Q on $\mathsf{Z} \times \mathcal{Z}$ and denote by $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ the corresponding probability distribution and expectation with initial distribution ξ. Without loss of generality, we assume that $(Z_k)_{k\in\mathbb{N}}$ is the associated canonical process. By construction, for any $\mathsf{A} \in \mathcal{Z}$, $\mathbb{P}_\xi(Z_k \in \mathsf{A} \,|\, Z_{k-1}) = \mathrm{Q}(Z_{k-1}, \mathsf{A})$, $\mathbb{P}_\xi$-a.s. In the case $\xi = \delta_z$, $z \in \mathsf{Z}$, $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ are denoted by $\mathbb{P}_z$ and $\mathbb{E}_z$, respectively. We now make an assumption about the mixing properties of Q, which essentially reflects **TD** 4.

UGE 1 *The Markov kernel Q admits μ as an invariant distribution and is uniformly geometrically ergodic, that is, there exists $t_{\mathrm{mix}} \in \mathbb{N}$ such that for all $k \in \mathbb{N}$,*

$$\Delta(\mathrm{Q}^k) = \sup_{z,z'\in\mathsf{Z}}(1/2)\|\mathrm{Q}^k(z,\cdot) - \mathrm{Q}^k(z',\cdot)\|_{\mathsf{TV}} \leq (1/4)^{\lfloor k/t_{\mathrm{mix}}\rfloor} .$$

For $q \in \mathbb{N}$, $k \in \mathbb{N}$, and the Markov chain $\{Z_n\}_{n\in\mathbb{N}}$ satisfying the uniform geometric ergodicity condition **UGE** 1, we define the σ-algebras $\mathcal{F}_k = \sigma(Z_\ell, \ell \leq k)$ and $\mathcal{F}^+_{k+q} = \sigma(Z_\ell, \ell \geq k+q)$. In such a scenario, using (Douc et al., 2018, Theorem 3.3), the respective β-mixing coefficient of $\mathcal{F}_k$ and $\mathcal{F}^+_{k+q}$ is bounded by

$$\beta(q) \equiv \beta(\mathcal{F}_k, \mathcal{F}^+_{k+q}) \leq (1/4)^{\lfloor q/t_{\mathrm{mix}}\rfloor} .$$

In this chapter we rely on the following useful version of Berbee's coupling lemma Berbee (1979), which is due to (Dedecker and Louhichi, 2002, Lemma 4.1):

Lemma 8. *(Lemma 4.1 in Dedecker and Louhichi (2002)) Let X and Y be two random variables taking their values in Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let U be a random variable with uniform distribution on $[0;1]$ that is independent of (X,Y). There exists a random variable $Y^\star = f(X,Y,U)$ where f is a measurable function from $\mathcal{X}\times\mathcal{Y}\times[0,1]$ to $\mathcal{Y}$, such that:*

1. *$Y^\star$ is independent of X and has the same distribution as Y;*

2. $\mathbb{P}(Y^\star \neq Y) = \beta(\sigma(X), \sigma(Y))$.

Let us now consider the extended measurable space $\tilde{\mathsf{Z}}_{\mathbb{N}} = \mathsf{Z}^{\mathbb{N}} \times [0,1]$, equipped with the σ-field $\tilde{\mathcal{Z}}_{\mathbb{N}} = \mathcal{Z}^{\otimes \mathbb{N}} \otimes \mathcal{B}([0,1])$. For each probability measure ξ on $(\mathsf{Z}, \mathcal{Z})$, we consider the probability measure $\tilde{\mathbb{P}}_\xi = \mathbb{P}_\xi \otimes \mathbf{Unif}([0,1])$ and denote by $\tilde{\mathbb{E}}_\xi$ the corresponding expected value. Finally, we denote by $(\tilde{Z}_k)_{k\in\mathbb{N}}$ the canonical process $\tilde{Z}_k \colon ((z_i)_{i\in\mathbb{N}}, u) \in \tilde{\mathsf{Z}}_{\mathbb{N}} \mapsto z_k$ and $U \colon ((z_i)_{i\in\mathbb{N}}, u) \in \tilde{\mathsf{Z}}_{\mathbb{N}} \mapsto u$. Under $\tilde{\mathbb{P}}_\xi$, $\{\tilde{Z}_k\}_{k\in\mathbb{N}}$ is by construction a Markov chain with initial distribution ξ and Markov kernel Q independent of U. Moreover, the distribution of U under $\tilde{\mathbb{P}}_\xi$ is uniform over $[0,1]$. Using the above construction, we obtain a useful blocking lemma, which is also stated in Dedecker and Louhichi (2002).

Lemma 9. *Assume **UGE** 1, let $q \in \mathbb{N}$ and ξ be a probability measure on $(\mathsf{Z}, \mathcal{Z})$. Then, there exists a random process $(\tilde{Z}^\star_k)_{k\in\mathbb{N}}$ defined on $(\tilde{\mathsf{Z}}_{\mathbb{N}}, \tilde{\mathcal{Z}}_{\mathbb{N}}, \tilde{\mathbb{P}}_\xi)$ such that for any $k \in \mathbb{N}$,*

(a) For any i, vector $V_i^\star = (\tilde{Z}^\star_{iq+1}, \ldots, \tilde{Z}^\star_{iq+q})$ has the same distribution as $V_i = (Z_{iq+1}, \ldots, Z_{iq+q})$ under $\tilde{\mathbb{P}}_\xi$;

(b) The sequences $(V^\star_{2i})_{i\geq 0}$ and $(V^\star_{2i+1})_{i\geq 0}$ are i.i.d. ;

(c) For any i, $\tilde{\mathbb{P}}_\xi(V_i \neq V_i^\star) \leq \beta(q)$;

Proof The proof follows from Lemma 8 and the relations between **UGE** 1 and β-mixing coefficient, see (Douc et al., 2018, Theorem 3.3). ■

D.1. Proof of Theorem 6

We aim to reduce the proof of the given bound to that of Corollary 4. Since the initial distribution of the sequence of states is $s_0 \sim \nu$, we must first remove the dependence on the initial condition. Indeed, using (Douc et al., 2018, Lemma 19.3.6 and Theorem 19.3.9) for any two probabilities ν and $\tilde{\nu}$ on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ there is a *maximal exact coupling* $(\Omega, \mathcal{F}, \tilde{\mathbb{P}}_{\nu,\tilde{\nu}}, s, \tilde{s}, T)$ of $\mathbb{P}_\nu$ and $\mathbb{P}_{\tilde{\nu}}$, that is,

$$\|\nu P_\pi^n - \tilde{\nu} P_\pi^n\|_{\mathrm{TV}} = 2\tilde{\mathbb{P}}_{\nu,\tilde{\nu}}(T > n) \, .$$

Under $\tilde{\mathbb{P}}_{\nu,\tilde{\nu}}$, the sequences $\{s_k\}_{k\in\mathbb{N}}$ and $\{\tilde{s}_k\}_{k\in\mathbb{N}}$ are Markov chains with initial distributions ν and $\tilde{\nu}$, respectively. We write $\tilde{\mathbb{E}}_{\nu,\tilde{\nu}}$ for the expectation with respect to $\tilde{\mathbb{P}}_{\nu,\tilde{\nu}}$. T is the coupling time, which is defined as

$$T = \inf_{k\in\mathbb{N}} \{s_k = \tilde{s}_k\} \, .$$

Let us now fix $\tilde{\nu} = \mu$ and for $n \in \mathbb{N}$ define an event $A_n = \{T > n/2\}$. Under **TD** 4, we can bound its probability as

$$\tilde{\mathbb{P}}_{\nu,\mu}(A_n) = \tilde{\mathbb{P}}_{\nu,\mu}(T > n/2) \leq (1/4)^{\lfloor n/(2t_{\mathrm{mix}})\rfloor} \, .$$

Thus, for a fixed $\delta \in (0; 1/3)$ we can achieve $\tilde{\mathbb{P}}_{\nu,\mu}(A_n) \leq \delta$ as soon as

$$n \geq \frac{2t_{\mathrm{mix}} \log(4/\delta)}{\log 4} \, .$$

Hence, starting from this point we work on the event $\Omega \setminus A_n$ which has probability at least $1-\delta$. On this event $\{s_k\}_{k\geq n/2}$ coincides with $\{\tilde{s}_k\}_{k\geq n/2}$, which is a stationary Markov chain with initial distribution μ. Assume now that the sample size n satisfies

$$n/2 = 2qm + k \;, \quad 0 \leq k < 2q \;, \tag{48}$$

where $q \in \mathbb{N}$ is a parameter that will be determined later. Using the construction of Lemma 9, we then construct a sequence of random variables$\{\tilde{s}^\star_{n/2+2jq}\}_{j=1,\ldots,m}$, which are i.i.d. with law μ under $\tilde{\mathbb{P}}_{\nu,\mu}$. Moreover, with a union bound,

$$\tilde{\mathbb{P}}_{\nu,\mu}(\exists j \in \{1,\ldots,m\} : \tilde{s}^\star_{n/2+2jq} \neq \tilde{s}_{n/2+2jq}) \leq m(1/4)^{\lfloor q/t_{\rm mix}\rfloor} \;.$$

The bound (48) implies that $m \leq n/(4q)$. Thus in order to achieve that $\tilde{\mathbb{P}}_{\nu,\mu}(\exists j \in \{1,\ldots,m\} : \tilde{s}^\star_{n/2+2jq} \neq \tilde{s}_{n/2+2jq}) \leq \delta$ it is enough to ensure that

$$m(1/4)^{\lfloor q/t_{\rm mix}\rfloor} \leq 4m(1/4)^{q/t_{\rm mix}} \leq (n/q)(1/4)^{q/t_{\rm mix}} \leq \delta \;.$$

In order to satisfy this constraint for fixed $\delta \in (0,1)$, we choose

$$q = \left\lceil \frac{t_{\rm mix}\log(n/\delta)}{\log 4} \right\rceil \;. \tag{49}$$

Thus, setting the block size q as in (49), we get that for sample size n satisfying (48), with probability at least $1-2\delta$ the results of Algorithm 2 are indistinguishable from the result of Algorithm 1 under the generative model assumption **TD** 1 applied with sample size

$$n/(4q) - 1 \leq m \leq n/(4q) \;.$$

Hence, the rest of the proof follows directly from the results of Corollary 4 applied with sample size m.