NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Applied Mathematics and Informatics"

BACHELOR'S THESIS

Research Project on the Topic:

Optimality of Temporal Difference Learning for policy evaluation in reinforcement learning

Submitted by the Student:

group #БПМИ212, 4th year of study **Horbach Maryna Paulauna**

Approved by the Supervisor:

Samsonov Sergey Vladimirovich
Senior Lecturer, Candidate of Physical and Mathematical Sciences
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Annotation

Temporal Difference Learning (TD) is a widely used method for policy evaluation in reinforcement learning (RL). However, the optimality of TD-based algorithms remains an open problem. Our analysis consider policy evaluation in discounted Markov decision processes as a linear stochastic approximation problem, providing new insights into the theoretical limits of TD learning. We analyse performance and derive improved convergence bounds for TD(0) and its variance-reduced modification with control variables (Variance-reduced Temporal Difference Algorithm, VRTD) under relaxed assumptions – eliminating the need for assumption on bounding relative variance of gradient differences in the original work. We further investigate whether these algorithms can achieve optimal loss rates. Our proof technique is based on decomposition of the error and stability result for the product of random matrices.

Аннотация

Метод временных разностей (Temporal Difference Learning, TD) широко используется для оценки политик в обучении с подкреплением (Reinforcement Learning, RL). Однако оптимальность алгоритмов на основе TD остается открытой проблемой. Наш анализ рассматривает задачу оценки политик в дисконтированных марковских процессах принятия решений в рамках парадигмы линейной стохастической аппроксимации, что позволяет получить новые теоретические результаты TD-алгоритмах. Мы анализируем производительность и выводим улучшенные оценки сходимости для TD(0) и его модификации с использованием контрольных переменных для уменьшения дисперсии (Variance-reduced Temporal Difference Algorithm, VRTD) в ослабленных предположениях – устраняя необходимость в предположении об ограниченности относительной дисперсии разностей градиентов, которое рассматривалось в оригинальной работе. Кроме того, мы исследуем, могут ли эти алгоритмы достигать оптимальных показателей функции потерь. Наша методика доказательства основана на декомпозиции ошибки и результатах устойчивости для произведения случайных матриц.

Keywords

Temporal Difference Learning, Policy Evaluation, Reinforcement Learning, Linear Stochastic Approximation

1 Introduction

Reinforcement Learning algorithms are widely used today in various domains, such as recommendation systems, robotics, natural language processing and others. Applications benefit from RL's ability to deal with complex dynamics and decision-making under uncertainty.

RL problems are usually formulated in terms of Markov decision processes (MDP) [13]. An MDP is defined as a tuple (S, A, P, R, γ), where S represents the state space, A — the action space, $P(s'|s,a)$ — the transition probability function, $R(s,a)$ the reward function, and $\gamma \in [0,1]$ the discount factor. Denote $r(s) := \sum_{s' \in S, a \in A} P(s'|s,a)R(s,a)$ the expected instantaneous reward generated at state s. At each step, an agent chooses an action $a \in A$ according to the current state $s \in S$ of the environment defined by the MDP, which is initially unknown. The agent receives a reward $R(s,a)$, and then the environment moves to a new state s' according to the transition function. The goal is to learn an optimal policy $\pi_* : S \to A$, in other words, to create a mapping from states to actions, which maximizes the expected total discounted reward, represented by value function (1).

$$v_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \Big| s_0 = s, \pi\right] \tag{1}$$

We consider the case where the number of states is finite and equal to D, both the expected reward function r and the value function v are D-dimensional vectors of reals. Then the value function is given by the solution to the Bellman equation

$$v^* = \gamma P v^* + r. \tag{2}$$

From this point, we assume the Markov chain is aperiodic and ergodic, with a unique stationary distribution $\mu := (\mu_1, \ldots, \mu_D)$ where $\mu_i > 0$ for all $i \in [D]$, satisfying $\mu P = \mu$. We define the $D \times D$ diagonal matrix

$$M = \mathrm{diag}(\mu_1, \ldots, \mu_D), \tag{3}$$

whose positive diagonal entries correspond to the stationary probabilities.

Policy evaluation is one of the fundamental challenges in RL, that utilizes Temporal Difference (TD) Learning. It is a necessary component of many algorithms, for example, SARSA [14]. The idea behind the TD algorithm for policy evaluation is to iteratively solve the Bellman equation.

In case of large or continuous state space, storing exact value functions becomes computationally infeasible, although some function approximation techniques can be used. One of the most common approaches is linear function approximation [12], where the value function $v_\pi(s)$ is approximated as

$$v_\pi^\theta(s) = \psi(s)^\top \theta, \tag{4}$$

where $\psi : S \to \mathbb{R}^d$ is a feature mapping. The goal is to estimate an optimal parameter θ^* by optimizing the expected squared error with respect to the invariant state distribution μ

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_\mu \left[(v_\pi^\theta(s) - v_\pi(s))^2 \right]. \tag{5}$$

As the basic TD algorithm is known to be suboptimal, in this work we consider also modifications of it with control variables — Variance-reduced Temporal Difference (VRTD) Algorithm — which was introduced in [5]. This work aims to investigate VRTD algorithm under assumptions from [8] and create mean squared error (MSE) bound. We consider the linear stochastic approximation technique based on the methods introduced in [8], with the goal of studying the feasibility of optimal theoretical bounds.

For VRTD, we select a d-dimensional approximation subspace $\mathbb{S} := \text{span}\{\psi_1, ..., \psi_d\}$, where $\psi_1, \ldots, \psi_d$ are d linearly independent basis vectors. Each state $s \in [D]$ is represented by its feature vector $\psi(s) := [\psi_1(s), \psi_2(s), ..., \psi_d(s)]^\top$.

Let $M_{\mathbb{S}}$ denote the projection operator onto $\mathbb{S}$ with respect to the $\|\cdot\|_M$-norm. The projected fixed-point solution $\bar{v} \in \mathbb{S}$ satisfies

$$\bar{v} = M_{\mathbb{S}}(\gamma P \bar{v} + r). \tag{6}$$

In matrix notation, with $\Psi := [\psi_1, \psi_2, ..., \psi_d]^\top$, any $v^\diamondsuit \in \mathbb{S}$ corresponds to a parameter vector $\theta^\diamondsuit \in \mathbb{R}^d$ via $v^\diamondsuit = \Psi\theta^\diamondsuit$. This allows us to rewrite (6) equivalently as

$$\Psi M \Psi^\top \bar{\theta} = \Psi M \gamma P \Psi^\top \bar{\theta} + \Psi M r. \tag{7}$$

Standard TD solves the original Bellman equation, but VRTD solves the projected version to reduce variance. In this work we aim to establish squared error bounds and study their optimality.

Results of the work are further discussed in Section 4. They includes improved bounds for

TD and VRTD algorithms with Polyak-Ruppert estimation (1, Theorem 2, Theorem 3). Additionally, we provide implementation of VRTD algorithm in Python programming language, which was not previously available in open source.

We present the analysis of both standard TD(0) and its variance-reduced modification (VRTD) with control variables, enhancing the theoretical understanding through several contributions:

- **Relaxed Assumptions:** Our analysis eliminates the restrictive variance bound on gradient differences required in [5], establishing convergence under more practical conditions. This is achieved through an error decomposition that separates the effects of approximation error and stochastic noise.
- **Bound Improvement:** The estimation of the squared error for a basic TD algorithm in our work shows an improvement in minor terms, comparing with [8].
- **Proof Technique**: The theoretical framework combines:
  - An error decomposition isolating initialization, approximation, and stochastic noise terms;
  - Advanced stability analysis for products of random matrices;
  - Martingale difference techniques for handling the temporal correlations.
- **Open-Source Implementation**: We provide the first publicly available Python implementation of VRTD algorithm, enabling reproducible research and practical applications.

2 Literature review

The Temporal Difference learning algorithm was initially introduced in the work of Sutton [9]. While the vanilla TD algorithm does not guarantee optimality in all cases, its asymptotic convergence when combined with linear function approximation was established in [12]. Further classical asymptotic convergence guarantees were later provided by Borkar and his co-authors in [2], who extended the theoretical understanding of the algorithm. Notably, the first finite-time convergence analysis for TD learning in the i.i.d. setting was provided by Sutton and co-authors in [10], making a significant step in the theoretical study of reinforcement learning algorithms.

The finite-time analysis of TD learning under Markovian noise was conducted in [1], using non-smooth analysis for a modified version of TD learning. A limitation of this approach is that

it does not utilize the variance reduction benefits associated with parallel computing. In more recent work [4], there was presented an enhanced analysis of the standard TD algorithm, which effectively addressed this limitation.

The basic TD algorithm can also be effectively combined with the iterate averaging technique. The idea of the method is to compute the result as a weighted arithmetic mean among the fixed number of last iterations. The convergence of this approach was demonstrated in [11], providing a theoretical foundation for its use. Furthermore, the TD algorithm incorporating Polyak-Ruppert iterate averaging has also been proven to converge. This result was achieved by Polyak and Juditsky in their work [6], which highlighted the benefits of this averaging scheme in stabilizing and accelerating the convergence of stochastic approximation algorithms.

The work of [5] establishes lower bounds on the MSE and sample complexity for TD methods. Additionally, the authors develop the variance-reduced temporal difference (VRTD) algorithm, which achieves nearly optimal stochastic error. They also introduce the variance-reduced fast temporal difference (VRFTD) algorithm, which matches both deterministic and stochastic error bounds under certain conditions. While, the VRTD and VRFTD algorithms theoretically achieves nearly optimal bounds presented in [5], in practice, reaching optimality is challenging due to the difficulties in computing necessary constants, which are often unknown. In particular, the step sizes depend on the minimal eigenvalue of the feature matrix, which complicates an optimal implementation. In our work, we analyze the algorithms introduced in [5] and aim to develop bounds for them in a new setting.

The following work, [3] was also a source of inspiration in terms of bounding techniques and algorithms structure. It provides an instance-specific analysis of of the sample complexity in stochastic approximation algorithms for finite state space in terms of the l_∞-norm. The authors show that TD(0) with iterate averaging do not meet these bounds in the non-asymptotic setting. To address this issue, they propose a variance-reduced TD algorithm. Although, this algorithm requires $O(\frac{1}{(1-\gamma)^3})$ samples per epoch, which means it does not achieve optimal sample complexity.

The setting and assumptions on the generative model were adopted from [8] for our research. There, the authors establish a sharper high-probability error bound for TD learning with linear stochastic approximation and Polyak-Ruppert averaging, using an instance-independent step size. However, they also demonstrate that this approach may result in suboptimal variance. Another significant contribution of this work is a novel proof of exponential stability for the TD algorithm, which provides deeper insights into its behavior.

3 Notations and definitions

Define $\|x\|_2 = \sqrt{\sum_{i=1}^m x_{(i)}^2}$ as l_2-norm and $\|x\|_\infty := \max_{i\in[m]} |x_{(i)}|$ as l_∞-norm. Also let $\|A\|_2$ be a spectral norm of given matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ — the smallest and largest eigenvalue. For a symmetric positive definite matrix A, define $\langle x, y\rangle_A = x^\top Ay$ and $\|x\|_A = \sqrt{x^\top Ax}$. The last expression is denoted as l_A-norm.

In this work "i.i.d." stands for ""independent and identically distributed"

4 Main results

4.1 Bounds for TD(0) learning algorithm

In this section we analyse the the canonical Temporal Difference learning algorithm under a constant learning rate. Our framework incorporates Polyak-Ruppert averaging as a final stage of the algorithm to improve convergence rate.

Algorithm 1 Temporal difference learning TD(0)

Input: n – number of iterations, η – step size, $\psi(\cdot) : S \to \mathbb{R}^d$ – feature mapping, π – behavioral policy;

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

Receive tuple (s_t, a_t, s'_t) according to **TD**1;

Update the parameter $\theta_t = \theta_{t-1} - \eta(A_t\theta_{t-1} - b_t)$ based on A_t, b_t from (8)

end for

Output: Averaged estimation $\widehat{\theta}_n = \frac{1}{n}\sum_{t=1}^n \theta_t$

We begin by deriving error bounds for the TD(0) algorithm (Algorithm 1). To facilitate our analysis, we introduce the following notation

$$\begin{aligned} A_t &= \psi(s_t)\{\psi(s_t) - \gamma\psi(s'_t)\}^\top, \\ b_t &= \psi(s_t)r(s_t, a_t) \\ \bar{A} &= \mathbb{E}_{s\sim\mu, s'\sim P_\pi(\cdot|s)}[\psi(s)\{\psi(s) - \gamma\psi(s')\}^\top] \\ \bar{b} &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot|s)}[\psi(s)r(s, a)] \end{aligned} \tag{8}$$

In order to write the instance of the LSA algorithm for the system (8), we also present the t-th step randomness $Z_t = (s_t, a_t, s'_t)$. For convenience, define $A_t = A(Z_t)$, and $b_t = b(Z_t)$. Then

the corresponding LSA update equation with step size η rewrites as

$$\theta_t = \theta_{t-1} - \eta(A_t\theta_{t-1} - b_t)\,, \tag{9}$$

where A_t and b_t are given by (8). Using the update rule in equation (9), we derive an expression for the approximation error 10. The analysis will focus on bounding the norm of this error term through the following steps.

$$\begin{aligned}
&\theta_{t+1} - \theta_* = \theta_t - \theta_* - \eta(A_{t+1}\theta_t - b_{t+1}) = (I - \eta A_{t+1})(\theta_t - \theta_*) - \eta\varepsilon_{t+1}, \\
&\text{where we have set} \\
&\varepsilon_t = (A_t - \bar{A})\theta_* - (b_t - \bar{b})
\end{aligned} \tag{10}$$

To simplify further computations, introduce additional notations, which also will be used in other sections of the work.

$$\Sigma_\varepsilon = \mathbb{E}[\varepsilon_i \varepsilon_i^T] \tag{11}$$

$$\Gamma_{k:t} = \prod_{i=k}^{t} (I - \eta A_i) \tag{12}$$

$$G_{t+1:i}^{(\eta)} = \mathbb{E}[\Gamma_{t+1:i}^{(\eta)}] = (I - \eta\bar{A})^{i-t} \tag{13}$$

$$Q_t = \eta \sum_{i=t}^{n} G_{t+1:i}^{(\eta)} \tag{14}$$

$$\Sigma_n = \frac{1}{n} \sum_{t=1}^{n} Q_t \Sigma_\varepsilon Q_t^T \tag{15}$$

$$V_A = \mathbb{E}\left\|(A_i - \bar{A})\right\|^2 \tag{16}$$

In this work, we analyze Temporal Difference (TD) algorithms under a standard set of assumptions, as introduced in [7]. We subsequently verify some of these assumptions for TD learning through several propositions.

A 1. *Sequence $\{Z_k\}_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution μ.*

A 2. *$\int_Z A(z)\mathrm{d}\pi(z) = \bar{A}$ and $\int_Z b(Z)\mathrm{d}\pi(z) = \bar{b}$, with the matrix $-\bar{A}$ being Hurwitz. Moreover, $\|\varepsilon\|_\infty = \sup_{z\in Z} \|\varepsilon(z)\| < +\infty$, and the mapping $z \to A(z)$ is bounded, that is,*

$$C_A = \sup_{z\in Z} \|A(z)\| \vee \sup_{z\in Z} \left\|\tilde{A}(z)\right\| < \infty\,. \tag{17}$$

Moreover, for the noise covariance matrix, which is defined by the following expression

$$\Sigma_\varepsilon = \int_Z \varepsilon(z)\varepsilon(z)^\top \mathrm{d}\pi(z) \tag{18}$$

it holds that its minimal eigenvalue is bounded by 0*. In other words,*

$$\lambda_{\min} = \lambda_{\min}(\Sigma_\varepsilon) > 0\,. \tag{19}$$

A 3. *(p) There exist* $a > 0$, $\varkappa_p > 0$, $\eta_{p,\infty} > 0$ *(depending on* p*), such that* $\eta_{p,\infty} p \le 1/2$*, and for any* $\eta \in (0; \eta_{p,\infty})$, $u \in \mathbb{R}^d$, $n \in \mathbb{N}$,

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

We also consider the following assumptions on the generative mechanism.

TD1. *The tuples* (s, a, s') *are generated independently and identically distributed (i.i.d.) according to:*

$$s \sim \mu, \quad a \sim \pi(\cdot|s), \quad s' \sim P(\cdot|s,a), \tag{20}$$

where:

- μ *is the stationary state distribution*
- $\pi(\cdot|s)$ *is the policy distribution at state* s
- $P(\cdot|s,a)$ *is the transition dynamics*

TD2. *The feature covariance matrix* Σ_ψ *is non-degenerate, with minimal eigenvalue* $\lambda_{\min}(\Sigma_\psi) > 0$*. Furthermore, the feature mapping* $\psi : \mathcal{S} \to \mathbb{R}^d$ *is uniformly bounded*

$$\sup_{s \in \mathcal{S}} \|\psi(s)\| \le 1 \tag{21}$$

From now on we will assume that $\theta_0 = \theta_*$ in order to avoid problems with transient term completely. Under this condition, Theorem 1 establishes improved error bounds for the TD(0) algorithm, advancing previous convergence rate estimates presented in [8].

Theorem 1. *Under assumptions A1, A2, A3, TD1, TD2 the error of algorithm 1 satisfy the next*

bounding for any $n > 0$

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_n - \theta_*\right\|^2 \le \frac{\sqrt{\mathrm{Tr}\,(\Sigma_\infty)}}{\sqrt{n}} + \frac{\|\Sigma_\infty\|\, dC(n,\eta)}{2n^{3/2}\sqrt{\mathrm{Tr}\,(\Sigma_\infty)}} + \eta^{1/2}\frac{2\sqrt{\mathrm{Tr}\,(\Sigma_\varepsilon)}V_A^{1/2}}{a^{3/2}\sqrt{n}} + \eta^{3/2}\frac{V_A\sqrt{\mathrm{Tr}\,(\Sigma_\varepsilon)}}{a^{3/2}}\left(1 + \frac{V_A^{1/2}}{a}\right)$$

where d *is dimension of* θ_* *and*

$$C(n,\eta) = 2\frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}} + \frac{e^{-a\eta}(1 - e^{-an\eta})}{1 - e^{-a\eta}}.$$

Further in this section we introduce all necessary notations and provide a proof for the Theorem 1. We start with the following propositions from [7, Proposition 2] and [8, Lemma 2], which verifies the assumptions for TD algorithms and present some useful results.

Proposition 1. *Let* $\{\theta\}_{k\in\mathbb{N}}$ *be a sequence of TD updates generated by* (9) *under* ***TD** 1 and **TD** 2.*

- *Then the update scheme satisfies assumption* ***A** 2 with*

$$\bar{C}_A = 2(1+\gamma)\,, \quad \|\varepsilon\|_\infty = 2(1+\gamma)(\|\theta^*\| + 1)\,.$$

- *Moreover,* $\left\|\mathrm{I} - \eta\bar{A}\right\|^2 \le 1 - \eta a$ *for*

$$a = (1-\gamma)\lambda_{\min}(\Sigma_\psi)\,, \quad \eta_\infty = (1-\gamma)/(1+\gamma)^2\,.$$

Proposition 2. *Let* $\{\theta_k\}_{k\in\mathbb{N}}$ *be a sequence of TD(0) updates generated by* (10) *under* ***TD** 1 and **TD** 2. Then this update scheme satisfies assumption A3(p) with*

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

To prove Theorem 1, we employ a classical approach by decomposing the error term in (10) into components and then estimating them separately. This method is particularly convenient in such cases due to the martingale structures described in Appendix A.1.

$$\theta_t - \theta_* = J_t^{(0)} + H_t^{(0)} \tag{22}$$

The terms mentioned above are defined by the following pair of recursive relations

$$\begin{aligned} J_t^{(0)} &= (I - \eta\bar{A})J_{t-1}^{(0)} - \eta\varepsilon_t\,, & J_0^{(0)} &= 0\,, \\ H_t^{(0)} &= (I - \eta A_t)\,H_{t-1}^{(0)} - \eta(A_t - \bar{A})J_{t-1}^{(0)}\,, & H_0^{(0)} &= 0\,. \end{aligned} \tag{23}$$

Proceeding iteratively, we apply the same decomposition to $H_n^{(0)}$ and its subsequent terms, obtaining the definitions for all $L \geq 1$

$$\begin{aligned} J_t^{(l)} &= \left(I - \eta\bar{A}\right)J_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l-1)}\,, & J_0^{(l)} &= 0\,, \\ H_t^{(l)} &= (I - \eta A_t)\,H_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l)}\,, & H_0^{(l)} &= 0\,. \end{aligned} \tag{24}$$

Theoretical analysis of the sequences $\{J_t^{(l)}\}_{l\geq 1}$ and $\{H_t^{(l)}\}_{l\geq 1}$, including detailed proofs of their properties, can be found in Appendix A.1. With these foundations established, we now proceed to the proof of Theorem 1.

Proof. We start by taking the average of (22). Applying Lemma 1 to this averaged expression yields the following representation for $H_t^{(0)}$

$$H_t^{(0)} = \sum_{l=1}^{L} J_t^{(l)} + H_t^{(L)}\,.$$

In particular, for the case $L = 1$, we obtain the following decomposition

$$\hat{\theta}_n - \theta_* = \frac{1}{n}\sum_{t=1}^{n}(\theta_t - \theta_*) = \frac{1}{n}\sum_{t=1}^{n} J_t^{(0)} + \frac{1}{n}\sum_{t=1}^{n} H_t^{(0)} = \frac{1}{n}\sum_{t=1}^{n} J_t^{(0)} + \frac{1}{n}\sum_{t=1}^{n} J_t^{(1)} + \frac{1}{n}\sum_{t=1}^{n} H_t^{(1)}$$

Lemma 3 directly implies the following result

$$\mathsf{Cov}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} J_t^{(0)}\right) = \Sigma_n$$

Applying the Minkowski inequality to bound the L_2-norm of this expression and combining this with Lemma 3, we obtain the following estimate

$$\begin{aligned} \mathbb{E}^{1/2}\left\|\hat{\theta}_n - \theta_*\right\|^2 &\leq \frac{1}{\sqrt{n}}\mathbb{E}^{1/2}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{n} J_t^{(0)}\right\|^2 + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} J_t^{(1)}\right\|^2}{n} + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} H_t^{(1)}\right\|^2}{n} \\ &\leq \sqrt{\frac{\mathrm{Tr}\,(\Sigma_n)}{n}} + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} J_t^{(1)}\right\|^2}{n} + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} H_t^{(1)}\right\|^2}{n} \end{aligned} \tag{25}$$

We now bound each term in the above expression to derive our final error estimate. The second term's bound follows directly from Lemma 5, which establishes that

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_t^{(1)}\right\|^2 \leq \eta \frac{4n\mathrm{Tr}\left(\Sigma_\varepsilon\right) V_A}{a^3}$$

The bound for the third term is provided in Lemma 6, which states that

$$\mathbb{E}^{1/2}\left\|H_t^{(1)}\right\|^2 \leq \eta^{3/2} \frac{V_A \sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}} \left(1 + \frac{V_A^{1/2}}{a}\right)$$

The final step is to estimate $\mathrm{Tr}\left(\Sigma_n\right)$ through Σ_∞. To this end, we utilize representations for $Q_t - \bar{A}^{-1}$, $\sum_{t=1}^{n}(Q_t - \bar{A}^{-1})$ and $\Sigma_n - \Sigma_\infty$ analogous to those established in [13]. The detailed proof is provided in Appendix B.

$$\Sigma_n - \Sigma_\infty = \underbrace{\frac{1}{n}\sum_{t=1}^{n}(Q_t - \bar{A}^{-1})\Sigma_\varepsilon \bar{A}^{-\top} + \frac{1}{n}\sum_{t=1}^{n}\bar{A}^{-1}\Sigma_\varepsilon(Q_t - \bar{A}^{-1})^\top}_{D_1} +$$
$$+ \underbrace{\frac{1}{n}\sum_{t=1}^{n}(Q_t - \bar{A}^{-1})\Sigma_\varepsilon(Q_t - \bar{A}^{-1})^\top}_{D_2}$$

To analyze both terms, we apply Lemma 13, which derives the following bound

$$\|\Sigma_n - \Sigma_\infty\| \leq \frac{2\,\|\Sigma_\infty\|}{n} \frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}} + \frac{\|\Sigma_\infty\|}{n} \frac{e^{-a\eta}(1 - e^{-an\eta})}{1 - e^{-a\eta}}$$

For notational convenience, we define

$$C(n,\eta) = 2\frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}} + \frac{e^{-a\eta}(1 - e^{-an\eta})}{1 - e^{-a\eta}}$$

To facilitate subsequent analysis, we reformulate the inequality as follows

$$\|\Sigma_n - \Sigma_\infty\| \leq \frac{\|\Sigma_\infty\|\, C(n,\eta)}{n}$$

At this stage, applying the elementary inequality $|\mathrm{Tr}\left(\Sigma_n\right) - \mathrm{Tr}\left(\Sigma_\infty\right)| \leq d\,\|\Sigma_n - \Sigma_\infty\|$,

where d denotes the dimension of the covariance matrices, we obtain the bound

$$\begin{aligned}\sqrt{\mathrm{Tr}\left(\Sigma_n\right)} &\leq \sqrt{\mathrm{Tr}\left(\Sigma_\infty\right) + \left|\mathrm{Tr}\left(\Sigma_n\right) - \mathrm{Tr}\left(\Sigma_\infty\right)\right|} = \sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}\sqrt{1 + \left|\mathrm{Tr}\left(\Sigma_n\right) - \mathrm{Tr}\left(\Sigma_\infty\right)\right| / \mathrm{Tr}\left(\Sigma_\infty\right)} \\ &\leq \sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}\left(1 + \frac{\left|\mathrm{Tr}\left(\Sigma_n\right) - \mathrm{Tr}\left(\Sigma_\infty\right)\right|}{2\mathrm{Tr}\left(\Sigma_\infty\right)}\right) \leq \sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)} + \frac{\|\Sigma_\infty\| \, dC(n,\eta)}{2n\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}}.\end{aligned} \tag{26}$$

To complete the proof, we combine all preceding estimates and substitute them into (25), yielding the final result.

$$\begin{aligned}\mathbb{E}^{1/2}\left\|\hat{\theta}_n - \theta_*\right\|^2 \leq{}& \frac{\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}}{\sqrt{n}} + \frac{\|\Sigma_\infty\| \, dC(n,\eta)}{2n^{3/2}\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}} \\ &+ \eta^{1/2}\frac{2\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}V_A^{1/2}}{a^{3/2}\sqrt{n}} + \eta^{3/2}\frac{V_A\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}}\left(1 + \frac{V_A^{1/2}}{a}\right)\end{aligned}$$

□

We compare our derived bound with that established in [8]. With an appropriate choice of the step size η, this analysis reveals improvements in the higher-order terms. Building on the framework developed in Theorem 1, we subsequently analyze the VRTD algorithm under a more sophisticated structural configuration.

4.2 Bounds for VRTD algorithm with control variables

In this section we study VRTD algorithm, presented in [5]. The algorithms aim to reduce the variance compared to the simple TD(0). Below we introduce notations from the original paper.

Following [5], we define the operator for solving equation (7)

$$g(\theta) = \Psi M(\Psi^\top\theta - r - \gamma P\Psi^\top\theta), \quad \text{for } \theta \in \mathbb{R}^d \tag{27}$$

We begin by noting that by definition θ_* is the solution to $g(\theta) = 0$. The corresponding stochastic operator from sample ξ_i is defined as follows

$$\widetilde{g}(\theta, \xi_i) = \left(\langle\psi(s_i), \theta\rangle - r(s_i, s_i') - \gamma\langle\psi(s_i'), \theta\rangle\right)\psi(s_i) \tag{28}$$

Additionally, $\mathbb{E}_{s_i\sim\pi, s_i'\sim\mathsf{P}(\cdot|s_i)}[\widetilde{g}(\theta, \xi_i)] = g(\theta)$. Furthermore, from (2) and (6), it follows that $g(\theta_*) = 0$. Using the matrix definitions in (8), we can equivalently express the stochastic operator

as

$$\widetilde{g}(\theta, \xi_i) = A_i\theta - b_i \tag{29}$$

To support our algorithmic construction, we define the control variable

$$\widehat{g}(\theta) = \frac{1}{N_k}\sum_{i=1}^{N_k}\widetilde{g}(\theta, \xi_i^k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\{A_i^{(k)}\theta - b_i^{(k)}\} \tag{30}$$

Consider algorithm 2 from [5] under assumptions TD1 and TD2. The key idea is as follows: at the start of epoch k, we collect N_k samples to compute $\widehat{g}(\theta_{k,0})$, which serves as a control variate for variance reduction. We assume $N_k = N$ for all k in this analysis.

While VRTD in [5] implements weighted averaging at epoch ends, our analysis uses simple averaging (equal coefficients).

Algorithm 2 Variance Reduced Temporal Difference Algorithm from [5]

Input: $\widehat{\theta}_0 \in \mathbb{R}^d$, $\eta > 0$ and $N \in \mathbb{Z}_+$.

for $k = 1, \ldots, K$ **do**

Set $\theta_{k,0} = \widehat{\theta}_{k-1}$. Collect N samples $\xi_i^k = (s_i^k, s_i'^k, r(s_i^k, s_i'^k))$ from the i.i.d. model.

Calculate $\widehat{g}(\theta_{k,0})$.

for $t = 1, \ldots, n$ **do**

Collect a sample $\xi_{k,t} = (s_{k,t}, s'_{k,t}, R(s_{k,t}, s'_{k,t}))$ from the i.i.d. observation model and compute

$$\theta_{k,t} = \theta_{k,t-1} - \eta\left(\widetilde{g}(\theta_{k,t-1}, \xi_{k,t}) - \widetilde{g}(\theta_{k,0}, \xi_{k,t}) + \widehat{g}(\theta_{k,0})\right).$$

end for

Output of the epoch:

$$\widehat{\theta}_k = \frac{\sum_{t=1}^{n+1}\theta_{k,t}}{(n+1)}. \tag{31}$$

end for

Then the corresponding LSA update equation with step size η writes as

$$\theta_{k,t+1} = \theta_{k,t} - \eta\left(\widetilde{g}(\theta_{k,t}, \xi_{k,t+1}) - \widetilde{g}(\theta_{k,0}, \xi_{k,t+1}) + \widehat{g}(\theta_{k,0})\right) \tag{32}$$

To prove the theorem analogous to Theorem 1, we adapt the technique from [5]. For each

epoch k, we introduce an auxiliary parameter θ'_k defined as

$$g(\theta'_k) = g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}) \tag{33}$$

Using these notations, we reformulate the update equation (32) as

$$\begin{aligned}
&\theta_{k,t} - \theta'_k = \theta_{k,t-1} - \theta'_k - \eta\widetilde{g}(\theta_{k,t-1}, \xi_{k,t}) + \eta\widetilde{g}(\theta_{k,0}, \xi_{k,t}) - \eta(g(\theta_{k,0}) - g(\theta'_k)) \\
&= \theta_{k,t-1} - \theta'_k - \eta(g(\theta_{k,t-1}) - g(\theta'_k)) + \eta(g(\theta_{k,t-1}) - \widetilde{g}(\theta_{k,t-1}, \xi_{k,t}) - g(\theta_{k,0}) + \widetilde{g}(\theta_{k,0}, \xi_{k,t})) \\
&= (\theta_{k,t-1} - \theta'_k) - \eta(\bar{A}\theta_{k,t-1} - \bar{b} - \bar{A}\theta'_k + \bar{b}) \\
&+ \eta(\bar{A}\theta_{k,t-1} - \bar{b} - A_{k,t}\theta_{k,t-1} + b_{k,t} - \bar{A}\theta_{k,0} + \bar{b} + A_{k,t}\theta_{k,0} - b_{k,t}) \\
&= (I - \eta\bar{A})(\theta_{k,t-1} - \theta'_k) + \eta(\bar{A} - A_{k,t})(\theta_{k,t-1} - \theta_{k,0}) \\
&= (I - \eta\bar{A})(\theta_{k,t-1} - \theta'_k) + \eta(\bar{A} - A_{k,t})(\theta_{k,t-1} - \theta'_k) + \eta(\bar{A} - A_{k,t})(\theta'_k - \theta_{k,0}) \\
&= (I - \eta A_{k,t})(\theta_{k,t-1} - \theta'_k) + \eta(\bar{A} - A_{k,t})(\theta'_k - \theta_{k,0})
\end{aligned}$$

We now decompose the error term analogously to the approach in Subsection 4.1, as follows

$$\theta_{k,t} - \theta'_k = J^{(0)}_{k,t} + H^{(0)}_{k,t} + F^{(0)}_{k,t}\,, \tag{34}$$

where the terms are defined by recursions bellow, with $\bar{A} = \mathbb{E}[A_{k,i}]$

$$\begin{aligned}
J^{(0)}_{k,t} &= (I - \eta\bar{A})J^{(0)}_{k,t-1} + \eta(\bar{A} - A_{k,t})(\theta'_k - \theta_{k,0})\,, \quad J^{(0)}_{k,0} = 0\,, \\
H^{(0)}_{k,t} &= (I - \eta A_{k,t})\, H^{(0)}_{k,t-1} - \eta(A_{k,t} - \bar{A})J^{(0)}_{k,t-1}\,, \quad H^{(0)}_{k,0} = 0\,, \\
F^{(0)}_{k,t} &= (I - \eta A_{k,t})\, F^{(0)}_{k,t-1}\,, \qquad\qquad\qquad\qquad F^{(0)}_{k,0} = (\theta_{k,0} - \theta'_k)\,.
\end{aligned} \tag{35}$$

Extending this decomposition to $H^{(0)}_n$, we obtain for all $l \geq 1$ the following

$$\begin{aligned}
J^{(l)}_{k,t} &= \left(I - \eta\bar{A}\right) J^{(l)}_{k,t-1} - \eta(A_{k,t} - \bar{A})J^{(l-1)}_{k,t-1}\,, \qquad J^{(l)}_{k,0} = 0\,, \\
H^{(l)}_{k,t} &= (I - \eta A_{k,t})\, H^{(l)}_{k,t-1} - \eta(A_{k,t} - \bar{A})J^{(l)}_{k,t-1}\,, \quad H^{(l)}_{k,0} = 0\,.
\end{aligned} \tag{36}$$

Observe that the identity $\bar{A}\theta_* = \bar{b}$ implies $\widetilde{g}(\theta_*, \xi_i) = \varepsilon_i$. Consequently, we can reformulate

$$\mathsf{Cov}(\widehat{g}(\theta_*)) = \mathsf{Cov}\left(\frac{1}{N}\sum_{i=1}^{N}\widetilde{g}(\theta_*, \xi_i)\right) = \frac{1}{N}\mathsf{Cov}\left(\widetilde{g}(\theta_*, \xi_i)\right) = \frac{1}{N}\Sigma_\varepsilon \tag{37}$$

In this section, we establish an analogue of Theorem 1 that bounds the approximation error for a single epoch of the VRTD algorithm.

Theorem 2. *Under the assumptions A1, A2, A3, TD1, TD2 the error of algorithm 2 satisfies the next bounding for any $n > 0$*

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_k - \theta_*\right\|^2 \leq \frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{\sqrt{N}} + \frac{\left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_* - \theta_{k,0}\right\|}{\sqrt{N}}$$
$$+ \frac{\sqrt{2V_A}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}}{\sqrt{nN}} + \frac{\sqrt{V_A}\left\|\Sigma_\infty\right\| dC(n,\eta)\left\|\bar{A}^{-1}\right\|}{n^{3/2}\sqrt{2N}\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}} + \frac{\sqrt{8}V_A\left\|\bar{A}^{-1}\right\|\left\|\theta_* - \theta_{k,0}\right\|}{\sqrt{nN}a}$$
$$+ \frac{\left\|\theta_{k,0} - \theta_*\right\|}{\eta a n} + \frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)} + \left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_* - \theta_{k,0}\right\|}{\eta a\sqrt{N}n}$$
$$+ \frac{2\sqrt{2\eta}V_A\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{nN}} + \frac{2\sqrt{2\eta}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\left\|\theta_* - \theta_{k,0}\right\|}{a^{3/2}\sqrt{nN}}$$
$$+ \frac{\sqrt{2}\eta^{3/2}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{N}}\left(I + \frac{V_A^{1/2}}{a}\right) + \frac{\sqrt{2}\eta^{3/2}V_A^2\left\|\bar{A}^{-1}\right\|\left\|\theta_* - \theta_{k,0}\right\|}{a^{3/2}\sqrt{N}}\left(I + \frac{V_A^{1/2}}{a}\right)$$

where d is dimension of θ_* and

$$C(n,\eta) = 2\frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}} + \frac{e^{-a\eta}(1 - e^{-an\eta})}{1 - e^{-a\eta}}.$$

Proof. We first clarify the relationship between θ_* and θ'_k through definitions (33) and (29).

$$\begin{aligned}
\mathbb{E}^{1/2}\left\|\theta'_k - \theta_*\right\|^2 &= \mathbb{E}^{1/2}\left\|\bar{A}^{-1}(g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}) - g(\theta_*))\right\|^2 \\
&= \mathbb{E}^{1/2}\left\|-\bar{A}^{-1}\widehat{g}(\theta_*) + \bar{A}^{-1}(\widehat{g}(\theta_*) - g(\theta_*) + g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}))\right\|^2 \\
&\leq \mathbb{E}^{1/2}\left\|\bar{A}^{-1}\widehat{g}(\theta_*)\right\|^2 + \mathbb{E}^{1/2}\left\|\bar{A}^{-1}(\widehat{g}(\theta_*) - g(\theta_*) + g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}))\right\|^2 \\
&\leq \frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{\sqrt{N}} + \frac{\left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_* - \theta_{k,0}\right\|}{\sqrt{N}}
\end{aligned}$$

Building on these results, we derive the following bound via the Minkowski inequality

$$\begin{aligned}
\mathbb{E}^{1/2}\left\|\hat{\theta}_k - \theta_*\right\|^2 &\leq \mathbb{E}^{1/2}\left\|\hat{\theta}_k - \theta'_k\right\|^2 + \mathbb{E}^{1/2}\left\|\theta'_k - \theta_*\right\|^2 \\
&\leq \mathbb{E}^{1/2}\left\|\hat{\theta}_k - \theta'_k\right\|^2 + \frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{\sqrt{N}} + \frac{\left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_* - \theta_{k,0}\right\|}{\sqrt{N}}
\end{aligned} \tag{38}$$

We now focus on bounding the first term. Beginning with averaging (34), we observe that Lemma 1 remains valid for systems with control variables. Applying this lemma with parameter

$L = 1$ yields

$$\begin{aligned}\hat{\theta}_k - \theta'_k &= \frac{1}{n}\sum_{t=1}^{n}(\theta_{k,t} - \theta'_k) = \frac{1}{n}\sum_{t=1}^{n} J_{k,t}^{(0)} + \frac{1}{n}\sum_{t=1}^{n} H_{k,t}^{(0)} + \frac{1}{n}\sum_{t=1}^{n} F_{k,t}^{(0)} \\ &= \frac{1}{n}\sum_{t=1}^{n} J_{k,t}^{(0)} + \frac{1}{n}\sum_{t=1}^{n} F_{k,t}^{(0)} + \frac{1}{n}\sum_{t=1}^{n} J_{k,t}^{(1)} + \frac{1}{n}\sum_{t=1}^{n} H_{k,t}^{(1)}\end{aligned}$$

We bound the L_2-norm of this expression via Minkowski's inequality, which enables separate estimation of each term to derive the final bound.

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_k - \theta'_k\right\|^2 \leq \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} J_{k,t}^{(0)}\right\|^2}{n} + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} F_{k,t}^{(0)}\right\|^2}{n} + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} J_{k,t}^{(1)}\right\|^2}{n} + \frac{\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} H_{k,t}^{(1)}\right\|^2}{n} \tag{39}$$

To bound the first term, we invoke Lemma 8, which establishes the following

$$\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} J_{k,t}^{(0)}\right\|^2 \leq \frac{\sqrt{2nV_A}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_n\right)}}{\sqrt{N}} + \frac{\sqrt{8n}V_A\left\|\bar{A}^{-1}\right\|\left\|\theta_* - \theta_{k,0}\right\|}{\sqrt{N}a}$$

According the equation (35), the second term is expressed as

$$F_{k,t}^{(0)} = \left(I - \eta A_{k,t}\right)F_{k,t-1}^{(0)} = \left(I - \eta A_{k,t}\right)^t F_{k,0}^{(0)} = \left(I - \eta A_{k,t}\right)^t\left(\theta_{k,0} - \theta'_k\right)$$

Applying Minkowski's inequality combined with Proposition 2, we derive

$$\begin{aligned}\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} F_{k,t}^{(0)}\right\|^2 &= \mathbb{E}^{1/2}\left\|\sum_{t=1}^{n}\left(I - \eta A_{k,t}\right)^t\left(\theta_{k,0} - \theta'_k\right)\right\|^2 = \mathbb{E}^{1/2}\left\|\theta_{k,0} - \theta'_k\right\|^2\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n}\left(I - \eta A_{k,t}\right)^t\right\|^2 \\ &\leq \mathbb{E}^{1/2}\left\|\theta_{k,0} - \theta'_k\right\|^2\sum_{t=1}^{n}(1-\eta a)^t \leq \frac{\mathbb{E}^{1/2}\left\|\theta_{k,0} - \theta'_k\right\|^2}{\eta a} \\ &\leq \frac{\left\|\theta_{k,0} - \theta_*\right\|}{\eta a} + \frac{\mathbb{E}^{1/2}\left\|\theta_* - \theta'_k\right\|^2}{\eta a} \\ &\leq \frac{\left\|\theta_{k,0} - \theta_*\right\|}{\eta a} + \frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)} + \left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_* - \theta_{k,0}\right\|}{\eta a\sqrt{N}}\end{aligned}$$

For subsequent bounds, we observe that Lemma 2 remains valid for $l \geq 1$ even with control

variables. Thus, for all $l \geq 1$:

$$J_{k,t}^{(l)} = -\eta \sum_{i=1}^{t} G_{k,i+1:t}^{(\eta)} (A_i - \bar{A}) J_{k,i-1}^{(l-1)},$$
$$H_{k,t}^{(l)} = -\eta \sum_{i=1}^{t} \Gamma_{k,i+1:t}^{(\eta)} (A_i - \bar{A}) J_{k,i-1}^{(l)}$$

Lemma 10 establishes the following upper bound for the third term in our decomposition

$$\mathbb{E} \left\| \sum_{t=1}^{n} J_{k,t}^{(1)} \right\|^2 \leq \frac{8\eta n V_A^2 \left\| \bar{A}^{-1} \right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N} + \frac{8\eta n V_A^3 \left\| \bar{A}^{-1} \right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}$$

Recall that for all $x, y \geq 0$, the inequality

$$\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$$

holds. Applying the square root to both sides of our bound yields

$$\begin{aligned}
\mathbb{E}^{1/2} \left\| \sum_{t=1}^{n} J_{k,t}^{(1)} \right\|^2 &\leq \sqrt{\frac{8\eta n V_A^2 \left\| \bar{A}^{-1} \right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N} + \frac{8\eta n V_A^3 \left\| \bar{A}^{-1} \right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}} \\
&\leq \sqrt{\frac{8\eta n V_A^2 \left\| \bar{A}^{-1} \right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N}} + \sqrt{\frac{8\eta n V_A^3 \left\| \bar{A}^{-1} \right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}} \\
&\leq \frac{2\sqrt{2\eta n} V_A \left\| \bar{A}^{-1} \right\| \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}\sqrt{N}} + \frac{2\sqrt{2\eta n} V_A^{3/2} \left\| \bar{A}^{-1} \right\| \|\theta_* - \theta_{k,0}\|}{a^{3/2}\sqrt{N}}
\end{aligned}$$

To analyze the final term, we apply Lemma 11, which establishes the following bound

$$\begin{aligned}
\mathbb{E}^{1/2} \left\| H_{k,t}^{(1)} \right\|^2 \leq{}& \frac{\sqrt{2}\eta^{3/2} V_A^{3/2} \left\| \bar{A}^{-1} \right\| \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}\sqrt{N}} \left(I + \frac{V_A^{1/2}}{a} \right) \\
&+ \frac{\sqrt{2}\eta^{3/2} V_A^2 \left\| \bar{A}^{-1} \right\| \|\theta_* - \theta_{k,0}\|}{a^{3/2}\sqrt{N}} \left(I + \frac{V_A^{1/2}}{a} \right)
\end{aligned}$$

The concluding step mirrors Theorem 1, requiring estimation of $\operatorname{Tr}(\Sigma_n)$. We employ the inequality (40) from its proof, which establishes

$$\sqrt{\operatorname{Tr}(\Sigma_n)} \leq \sqrt{\operatorname{Tr}(\Sigma_\infty)} + \frac{\|\Sigma_\infty\| \, dC(n,\eta)}{2n\sqrt{\operatorname{Tr}(\Sigma_\infty)}}. \tag{40}$$

To obtain the final bound, we combine all term estimates and substitute them into (39)

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_k-\theta'_k\right\|^2 \le \frac{\mathbb{E}^{1/2}\left\|\sum\limits_{t=1}^{n} J_{k,t}^{(0)}\right\|^2}{n}+\frac{\mathbb{E}^{1/2}\left\|\sum\limits_{t=1}^{n} F_{k,t}^{(0)}\right\|^2}{n}+\frac{\mathbb{E}^{1/2}\left\|\sum\limits_{t=1}^{n} J_{k,t}^{(1)}\right\|^2}{n}+\frac{\mathbb{E}^{1/2}\left\|\sum\limits_{t=1}^{n} H_{k,t}^{(1)}\right\|^2}{n}$$
$$\le \frac{\sqrt{2V_A}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_n\right)}}{\sqrt{nN}}+\frac{\sqrt{8}V_A\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{\sqrt{nN}a}$$
$$+\frac{\left\|\theta_{k,0}-\theta_*\right\|}{\eta an}+\frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}+\left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_*-\theta_{k,0}\right\|}{\eta a\sqrt{N}n}$$
$$+\frac{2\sqrt{2\eta}V_A\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{nN}}+\frac{2\sqrt{2\eta}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{a^{3/2}\sqrt{nN}}$$
$$+\frac{\sqrt{2}\eta^{3/2}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right)+\frac{\sqrt{2}\eta^{3/2}V_A^{2}\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right)$$

Finally, we obtain the following bound by combining all previous estimates with (38)

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_k-\theta_*\right\|^2$$
$$\le \frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{\sqrt{N}}+\frac{\left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_*-\theta_{k,0}\right\|}{\sqrt{N}}$$
$$+\frac{\sqrt{2V_A}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}}{\sqrt{nN}}+\frac{\sqrt{V_A}\left\|\Sigma_\infty\right\|dC(n,\eta)\left\|\bar{A}^{-1}\right\|}{n^{3/2}\sqrt{2N}\sqrt{\mathrm{Tr}\left(\Sigma_\infty\right)}}+\frac{\sqrt{8}V_A\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{\sqrt{nN}a}$$
$$+\frac{\left\|\theta_{k,0}-\theta_*\right\|}{\eta an}+\frac{\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}+\left\|\bar{A}^{-1}\right\|\sqrt{V_A}\left\|\theta_*-\theta_{k,0}\right\|}{\eta a\sqrt{N}n}$$
$$+\frac{2\sqrt{2\eta}V_A\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{nN}}+\frac{2\sqrt{2\eta}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{a^{3/2}\sqrt{nN}}$$
$$+\frac{\sqrt{2}\eta^{3/2}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right)+\frac{\sqrt{2}\eta^{3/2}V_A^{2}\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right)$$

□

At this stage, we compare our result with the single-epoch error bound from [5]. The leading terms exhibit identical asymptotics, confirming that the elimination of Assumption 2 from [5] was successful. This bound consequently serves as a foundation for extending the analysis to the complete algorithm.

Furthermore, while [5] establishes that their single-epoch bound remains valid for VRFTD, we conjecture that Theorem 2 similarly holds in this setting, thereby broadening its applicability.

Our subsequent analysis will estimate the global error of the full algorithm by employing leveraging Theorem 2, which provides an error bound for a single epoch.

Theorem 3. *In the assumptions of the Theorem 2 the expected squared deviation of the final*

iterate from the optimal parameter satisfies the following inequality:

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_K-\theta_*\right\|^2 \leq Y^K\left\|\theta_{1,0}-\theta_*\right\|+W\sum_{k=0}^{K-1}Y^k=Y^K\left\|\theta_{1,0}-\theta_*\right\|+W\frac{Y^K-1}{Y-1},$$

where

$$\begin{aligned}W=&\frac{\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}\left(\Sigma_\varepsilon\right)}}{\sqrt{N}}+\frac{\sqrt{2V_A}\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}\left(\Sigma_\infty\right)}}{\sqrt{nN}}+\frac{\sqrt{V_A}\left\|\Sigma_\infty\right\|dC(n,\eta)\left\|\bar{A}^{-1}\right\|}{n^{3/2}\sqrt{2N}\sqrt{\operatorname{Tr}\left(\Sigma_\infty\right)}}\\&+\frac{\left\|\theta_{k,0}-\theta_*\right\|}{\eta an}+\frac{\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}\left(\Sigma_\varepsilon\right)}}{\eta a\sqrt{N}n}+\frac{2\sqrt{2\eta}V_A\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{nN}}\\&+\frac{\sqrt{2}\eta^{3/2}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right)\end{aligned}$$

$$\begin{aligned}Y=&\frac{\left\|\bar{A}^{-1}\right\|\sqrt{V_A}}{\sqrt{N}}+\frac{\sqrt{8}V_A\left\|\bar{A}^{-1}\right\|\left\|\theta_*-\theta_{k,0}\right\|}{\sqrt{nN}a}+\frac{\left\|\theta_{k,0}-\theta_*\right\|}{\eta an}+\frac{\left\|\bar{A}^{-1}\right\|\sqrt{V_A}}{\eta a\sqrt{N}n}\\&+\frac{2\sqrt{2\eta}V_A^{3/2}\left\|\bar{A}^{-1}\right\|}{a^{3/2}\sqrt{nN}}+\frac{\sqrt{2}\eta^{3/2}V_A^2\left\|\bar{A}^{-1}\right\|}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right)\end{aligned}$$

Proof. Using the notation from above, Theorem 2 result can be expressed as

$$\mathbb{E}^{1/2}\left\|\hat{\theta}_k-\theta_*\right\|^2 \leq W+Y\left\|\theta_{k,0}-\theta_*\right\|$$

We now proceed to unroll this recursive bound.

$$\begin{aligned}\mathbb{E}^{1/2}\left\|\hat{\theta}_K-\theta_*\right\|^2 &\leq W+Y\left\|\theta_{K,0}-\theta_*\right\| \leq W+YW+Y^2\left\|\theta_{K-1,0}-\theta_*\right\|\\&\leq\cdots\leq Y^K\left\|\theta_{1,0}-\theta_*\right\|+W\sum_{k=0}^{K-1}Y^k=Y^K\left\|\theta_{1,0}-\theta_*\right\|+W\frac{Y^K-1}{Y-1}\end{aligned}$$

□

4.3 Experiments

In this section, we present our implementation of the VRTD algorithm, which was not published by the original authors. The source code is publicly available at:
https://github.com/horbachmp/VRTD.

To check our implementation of the VRTD algorithm we decided to compare the algorithm performance with the article [5].Similarly, to the authors we propose two-state MRPs with a

discount factor $\gamma \in (0.5, 1)$, the transition matrix P (41) and reward vector r (42).

$$P = \begin{bmatrix} \frac{2\gamma-1}{\gamma} & \frac{1-\gamma}{\gamma} \\ \frac{1-\gamma}{\gamma} & \frac{2\gamma-1}{\gamma} \end{bmatrix} \tag{41}$$

$$r = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{42}$$

The stationary distribution is $\pi = [0.5, 0.5]$ because of the symmetry of the transition matrix. For convenience, the feature matrix is chosen as

$$P = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}$$

Notice, that it forms an orthonormal basis in terms of ℓ_Π-norm.

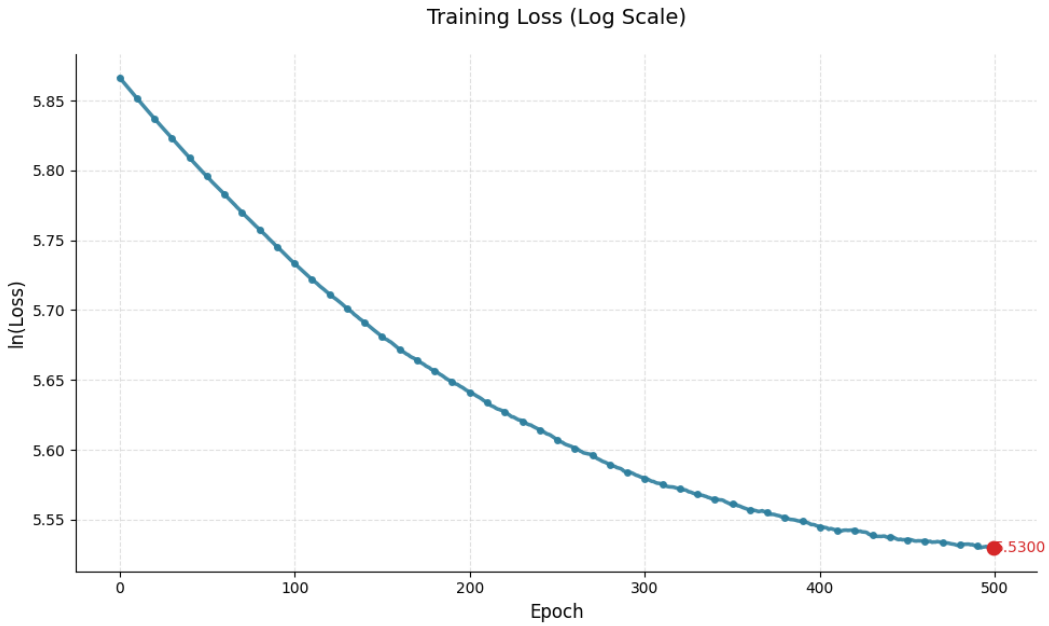

Figure 4.1: Squared error of VRTD algorithm with $\gamma = 0.999$

Our experimental reproduction differed from [5] primarily in using fixed $N_k \equiv N$ rather than variable ones. This design choice, while simplifying implementation, appears responsible for the observed higher loss values versus the original study.

The unavailability of convergence curves in [5] limits direct comparison, but our monotonic error decay (Figure 4.1 and Figure 4.2) confirms stable optimization behavior.

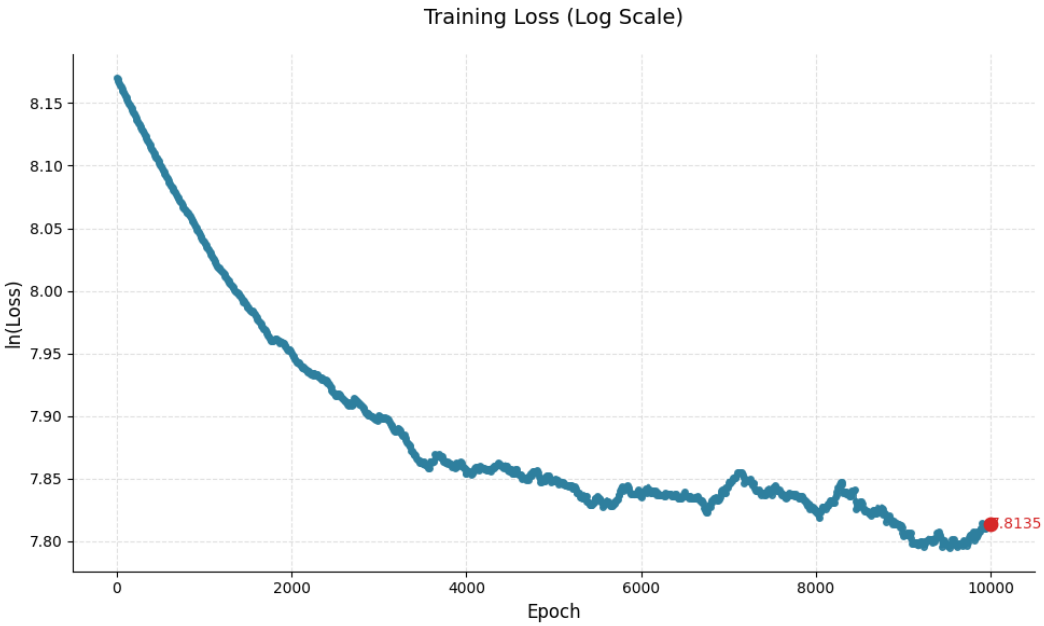

Figure 4.2: Squared error of VRTD algorithm with $\gamma = 0.9999$

5 Conclusion

This work analyzes the performance of TD and VRTD algorithms within the framework of linear stochastic approximation, establishing finite-time squared error bounds. The choice of the proof methodology is based on fundamental papers about optimality of TD learning, which were mentioned in Section 2.

Throughout the research, we investigated several approaches of error decomposition. It is then used for bounding an error per components. The main achievement of the work are new bounds for TD and VRTD algorithms. Firstly, we develop an estimation for the basic TD with Polyak-Ruppert estimation Theorem 1. After that the approach was generalized to VRTD algorithm (Theorem 2, Theorem 3). Moreover, we implemented VRTD algorithm using Python programming language, and reproduced some of the experiments analogous to [5].

TD algorithms are widely used today, which confirms the relevance of the work. By investigating modifications of the basic TD, such as VRTD that uses variance reduction, control variables and averaging, we aim to determine optimal hyperparameters and assess whether an optimal implementation is achievable in the current setting. Addressing this question would fill a literature gap and may lead to the optimization of these algorithms through the calculation of required parameters.

Based on received bound, we aim to determine optimal parameters, such as step size, which would enable optimal implementations. Future research could explore other modifications of TD algorithms and investigate their optimality, potentially leading to more efficient and robust

reinforcement learning methods. This work represents a step toward understanding the theoretical and practical limits of TD-based approaches, with the ultimate goal of improving their performance in real-world applications.

References

[1] Jalaj Bhandari, Daniel Russo, and Raghav Singal. "A finite time analysis of temporal difference learning with linear function approximation". In: *Conference on learning theory*. PMLR. 2018, pp. 1691–1692.

[2] Vivek S Borkar and Sean P Meyn. "The ODE method for convergence of stochastic approximation and reinforcement learning". In: *SIAM Journal on Control and Optimization* 38.2 (2000), pp. 447–469.

[3] Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. "Is temporal difference learning optimal? an instance-dependent analysis". In: *SIAM Journal on Mathematics of Data Science* 3.4 (2021), pp. 1013–1040.

[4] G Kotsalis, G Lan, and T Li. "Simple and optimal methods for stochastic variational inequalities". In: *II: Markovian noise and policy evaluation in reinforcement learning. arXiv, pages arXiv–2011.08434* (2020).

[5] Tianjiao Li, Guanghui Lan, and Ashwin Pananjady. "Accelerated and instance-optimal policy evaluation with linear function approximation". In: *SIAM Journal on Mathematics of Data Science* 5.1 (2023), pp. 174–200.

[6] Boris T Polyak and Anatoli B Juditsky. "Acceleration of stochastic approximation by averaging". In: *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855.

[7] Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. "Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning". In: *Advances in Neural Information Processing Systems* 37 (2025), pp. 12408–12460.

[8] Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. "Improved high-probability bounds for the temporal difference learning algorithm via exponential stability". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 4511–4547.

[9] Richard S Sutton. "Learning to predict by the methods of temporal differences". In: *Machine learning* 3 (1988), pp. 9–44.

[10] Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. "Fast gradient-descent methods for temporal-difference learning with linear function approximation". In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 993–1000.

[11] Vladislav Tadic. "On the Almost Sure Rate of Convergence of Linear Stochastic Approximation Algorithms". In: *Information Theory, IEEE Transactions on* 50 (Mar. 2004), pp. 401–409. DOI: 10.1109/TIT.2003.821971.

[12] John Tsitsiklis and Benjamin Van Roy. "Analysis of temporal-diffference learning with function approximation". In: *Advances in neural information processing systems* 9 (1996).

[13] Weichen Wu, Gen Li, Yuting Wei, and Alessandro Rinaldo. "Statistical Inference for Temporal Difference Learning with Linear Function Approximation". In: *preprint arXiv:2410.16106* (2024).

[14] Shaofeng Zou, Tengyu Xu, and Yingbin Liang. "Finite-sample analysis for sarsa with linear function approximation". In: *Advances in neural information processing systems* 32 (2019).

A Technical Proofs for J and H

A.1 TD algorithm

In this section, we present detailed proofs of the technical results used in establishing Theorem 1. Our analysis focuses on the error decomposition introduced earlier, examining each component separately to build the complete error bound.

Lemma 1. *In terms of definition (24) for any $l \in \{1, \dots, L\}$ and any $t \in \{1, \dots, n\}$ the following expression holds*

$$H_t^{(0)} = \sum_{l=1}^{L} J_t^{(l)} + H_t^{(L)},$$

Proof. Notice that the result of the lemma follows directly from the expression

$$H_t^{(l-1)} = J_t^{(l)} + H_t^{(l)}$$

We verify this formula via double induction (on both l and t).

Base case ($l = 1$, $t = 1$):

Direct computation yields

$$J_1^{(1)} + H_1^{(1)} = \left(I - \eta\bar{A}\right) J_0^{(1)} - \eta(A_1 - \bar{A})J_0^{(0)} + (I - \eta A_1)\, H_0^{(1)} - \eta(A_1 - \bar{A})J_0^{(1)} = 0$$

Simultaneously, definition (23) gives

$$H_1^{(0)} = (I - \eta A_1)\, H_0^{(0)} - \eta(A_1 - \bar{A})J_0^{(0)} = J_1^{(1)} + H_1^{(1)}$$

Hence, the base case holds.

Inductive step: Assume $J_{t-1}^{(1)} + H_{t-1}^{(1)} = H_{t-1}^{(0)}$ holds for any $t > 0$.

From definition (24) and inductive step we obtain

$$\begin{aligned}
J_t^{(1)} + H_t^{(1)} &= (I - \eta\bar{A})J_{t-1}^{(1)} - \eta(A_t - \bar{A})J_{t-1}^{(0)} + (I - \eta A_t)\, H_{t-1}^{(1)} - \eta(A_t - \bar{A})J_{t-1}^{(1)} \\
&= (I - \eta A_t)J_{t-1}^{(1)} + (I - \eta A_t)\, H_{t-1}^{(1)} - \eta(A_t - \bar{A})J_{t-1}^{(0)} \\
&= (I - \eta A_t)(J_{t-1}^{(1)} + H_{t-1}^{(1)}) - \eta(A_t - \bar{A})J_{t-1}^{(0)} \\
&= (I - \eta A_t)H_{t-1}^{(0)} - \eta(A_t - \bar{A})J_{t-1}^{(0)} = H_t^{(0)}
\end{aligned}$$

The expression above finishes the base for $l = 1$. Due to that we are able to proceed with

induction step on l. Using recurrent definitions (24) and induction proposition, imply

$$\begin{aligned} J_t^{(l)} + H_t^{(l)} &= \left(I - \eta\bar{A}\right) J_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l-1)} + (I - \eta A_t)\, H_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l)} \\ &= (I - \eta A_t)\, J_{t-1}^{(l)} + (I - \eta A_t)\, H_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l-1)} \\ &= (I - \eta A_t)\, H_{t-1}^{(l-1)} - \eta(A_t - \bar{A})J_{t-1}^{(l-1)} = H_t^{(l-1)} \end{aligned}$$

The received equation finalizes the proof. □

Corollary 1. *In terms of definition (24) for any $l \in \{1, \ldots, L\}$ and any $t \in \{1, \ldots, n\}$ the expression below follows from the proof of Lemma 1*

$$H_t^{(l-1)} = J_t^{(l)} + H_t^{(l)}$$

Lemma 2. *For any $l \in \{1, \ldots, L\}$ and any $t \in \{1, \ldots, n\}$ holds*

$$\begin{aligned} J_{t+1}^{(0)} &= -\eta \sum_{i=1}^{t+1} \left(I - \eta\bar{A}\right)^{t-i+1} \varepsilon_i. \\ J_t^{(l)} &= -\eta \sum_{i=1}^{t} G_{i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(l-1)}, \\ H_t^{(l)} &= -\eta \sum_{i=1}^{t} \Gamma_{i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(l)}, \end{aligned}$$

Proof. Observe that the first expression follows directly from definition (23), which establishes the base case for our analysis.

We analyze $J_t^{(l)}$ using definition (24), which gives $J_0^{(l)} = 0$. Unraveling the recursion yields

$$\begin{aligned} J_t^{(l)} &= \left(I - \eta\bar{A}\right) J_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l-1)} \\ &= (I - \eta\bar{A})^t J_0^{(l)} - \eta \sum_{i=0}^{t-1} (I - \eta\bar{A})^{t-1-i}(A_{i+1} - \bar{A})J_i^{l-1} = -\eta \sum_{i=1}^{t} (I - \eta\bar{A})^{t-i}(A_i - \bar{A})J_{i-1}^{l-1} = -\eta \sum_{i=1}^{t} G_{i} \end{aligned}$$

Analogously, for $H_t^{(l)}$ we obtain

$$\begin{aligned} H_t^{(l)} &= (I - \eta A_t)\, H_{t-1}^{(l)} - \eta(A_t - \bar{A})J_{t-1}^{(l)} = H_0^{(l)} \prod_{i=j}^{t} (I - \eta A_j) - \eta \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} (I - \eta A_j)(A_{i+2} - \bar{A})J_i^{(l)} \\ &= -\eta \sum_{i=1}^{t} \prod_{j=i+1}^{t} (I - \eta A_j)(A_i - \bar{A})J_{i-1}^{(l)} = -\eta \sum_{i=1}^{t} G_{i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(l)} \end{aligned}$$

□

Lemma 3. *Assume A1 and A2. Then for any natural n holds*

$$\mathsf{Cov}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} J_t^{(0)}\right) = \Sigma_n$$

Proof. We begin by substituting the expression from Lemma 2 and applying the notations introduced in (13) and (14). Rearranging the summations and using (14) yields:

$$\sum_{t=1}^{n} J_t^{(0)} = -\sum_{t=1}^{n}\eta\sum_{i=1}^{t}\left(I-\eta\bar{A}\right)^{t-i}\varepsilon_i = -\eta\sum_{t=1}^{n}\sum_{i=1}^{t} G_{i+1:t}^{(\eta)}\varepsilon_i = -\eta\sum_{i=1}^{n}\sum_{t=i}^{n} G_{i+1:t}^{(\eta)}\varepsilon_i = -\sum_{i=1}^{n} Q_i\varepsilon_i$$

This establishes the relationship between the cumulative error terms and the noise sequence $\{\varepsilon_i\}$.

Next, we compute the covariance matrix of the normalized sum. Using the independence of $\{\varepsilon_i\}$ and the definition of Σ_n from (15), we obtain

$$\mathsf{Cov}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} J_t^{(0)}\right) = \frac{1}{n}\mathsf{Cov}\left(-\sum_{t=1}^{n} Q_t\varepsilon_i\right) = \frac{1}{n}\sum_{i=t}^{n} Q_t\mathsf{Cov}(\varepsilon_i)Q_t^T = \frac{1}{n}\sum_{t=1}^{n} Q_t\Sigma_\varepsilon Q_t^T = \Sigma_n$$

□

Lemma 4. *Assume A1, A2 and A3. Then for any $t \in \{1, 2, \ldots, n\}$*

$$\begin{aligned}
\mathbb{E}\left\|J_{t+1}^{(0)}\right\|^2 &\leq \frac{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a}\eta \\
\mathbb{E}\left\|J_t^{(1)}\right\|^2 &\leq \frac{V_A\mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^2}\eta^2 \\
\mathbb{E}\left\|J_t^{(2)}\right\|^2 &\leq \frac{V_A^2\mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^3}\eta^3
\end{aligned}$$

Proof. Notice that from the definition of ε_i in (10), it follows that $\mathbb{E}J_{t+1}^{(0)} = 0$. Using the formulas from Lemma 2, the independence of $\{\varepsilon_i\}$, and standard covariance transformations, we derive:

$$\begin{aligned}
\mathbb{E}\left\|J_{t+1}^{(0)}\right\|^2 &= \mathrm{Tr}\left(\mathsf{Cov}\left(\eta\sum_{i=1}^{t+1}\left(I-\eta\bar{A}\right)^{t-i+1}\varepsilon_i\right)\right) = \eta^2\sum_{i=1}^{t+1}\mathsf{Cov}\left(\left(I-\eta\bar{A}\right)^{t-i+1}\varepsilon_i\right) \\
&= \eta^2\mathrm{Tr}\left(\Sigma_\varepsilon\right)\sum_{i=1}^{t+1}\left\|\left(I-\eta\bar{A}\right)\right\|^{2(t-i+1)} \leq \eta^2\mathrm{Tr}\left(\Sigma_\varepsilon\right)\sum_{i=1}^{t+1}\left(1-\eta a\right)^{t-i+1} \\
&\leq \eta^2\mathrm{Tr}\left(\Sigma_\varepsilon\right)\frac{1}{1-(1-\eta a)} \leq \frac{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a}\eta.
\end{aligned}$$

The inequality follows from Proposition 1 and the geometric series formula.

For $J_k^{(1)}$, we observe that it forms a sum of martingale differences by Lemma 2. This

property allows us to rewrite it as follows

$$\mathbb{E}\left\|J_t^{(1)}\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{t}\eta G_{i+1:t}^{(\eta)}(A_i-\bar{A})J_{i-1}^{(0)}\right\|^2 \leq \eta^2 V_A \sum_{i=1}^{t}\mathbb{E}\left\|G_{i+1:t}^{(\eta)}\right\|^2 \eta\frac{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a}$$
$$\leq \eta^3 V_A \frac{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a}\sum_{i=1}^{t}(1-\eta a)^{t-i} \leq \frac{V_A \mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^2}\eta^2$$

The bound for $J_t^{(2)}$ follows similarly

$$\mathbb{E}\left\|J_t^{(2)}\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{t}\eta G_{i+1:t}^{(\eta)}(A_i-\bar{A})J_{i-1}^{(1)}\right\|^2 \leq \eta^2 V_A \sum_{i=1}^{t}\left\|G_{i+1:t}^{(\eta)}\right\|^2 \frac{V_A \mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^2}\eta^2$$
$$\leq \frac{\eta^4 V_A^2 \mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^2}\sum_{i=1}^{t}(1-\eta a)^{t-i} \leq \frac{\eta^3 V_A^2 \mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^3}$$

This completes the bounds for all error components in the decomposition. □

Lemma 5. *Assume A1, A2 and A3, then any $t \in \{1,\ldots,n\}$ holds*

$$\mathbb{E}\left\|\sum_{t=1}^{n}J_t^{(1)}\right\|^2 \leq \eta\frac{4n\mathrm{Tr}\left(\Sigma_\varepsilon\right)V_A}{a^3}$$

Proof. Using the result of Lemma 2 for $l=1$, we first rewrite the cumulative sum:

$$\sum_{t=1}^{n}J_t^{(1)} = -\eta\sum_{t=1}^{n}\sum_{i=1}^{t}G_{i+1:t}^{(\eta)}(A_i-\bar{A})J_{i-1}^{(0)}$$
$$= -\eta\sum_{i=1}^{n}\left(\sum_{t=i}^{n}G_{i+1:t}^{(\eta)}\right)((A_i-\bar{A})J_{i-1}^{(0)}) = -\sum_{i=1}^{n}Q_i(A_i-\bar{A})J_{i-1}^{(0)}$$

Since $J_k^{(1)}$ forms a martingale difference sequence, we compute its second moment:

$$\mathbb{E}\left\|\sum_{t=1}^{n}J_t^{(1)}\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{n}Q_i(A_i-\bar{A})J_{i-1}^{(0)}\right\|^2 = \sum_{i=1}^{n}\mathbb{E}\left\|Q_i(A_i-\bar{A})J_{i-1}^{(0)}\right\|^2$$

The independence of $A_i-\bar{A}$ and $J_{i-1}^{(0)}$ yields:

$$\mathbb{E}\left\|\sum_{t=1}^{n}J_t^{(1)}\right\|^2 = \sum_{i=1}^{n}\mathbb{E}\left\|Q_i(A_i-\bar{A})J_{i-1}^{(0)}\right\|^2 \leq \sum_{i=1}^{n}\|Q_i\|^2\,\mathbb{E}\left\|(A_i-\bar{A})\right\|^2\mathbb{E}\left\|J_{l-1}^{(0)}\right\|^2$$

Applying Lemma 4 which gives $\mathbb{E}\left\|J_{t+1}^{(0)}\right\|^2 \leq \frac{\mathrm{Tr}(\Sigma_\varepsilon)}{a}\eta$ and the definition of Q_i from (14), we

obtain:

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_t^{(1)}\right\|^2 \leq \sum_{i=1}^{n} \|Q_i\|^2 \mathbb{E}\left\|(A_i - \bar{A})\right\|^2 \mathbb{E}\left\|J_{i-1}^{(0)}\right\|^2$$
$$= \sum_{i=1}^{n} \left\|\eta \sum_{j=i}^{n} G_{i+1:j}^{(\eta)}\right\|^2 V_A \frac{\operatorname{Tr}(\Sigma_\varepsilon)}{a}\eta = \eta^3 \frac{\operatorname{Tr}(\Sigma_\varepsilon)}{a} V_A \sum_{i=1}^{n} \left\|\sum_{j=i}^{n} G_{i+1:j}^{(\eta)}\right\|^2$$

To complete the bound, we estimate the norm $\left\|\sum_{j=i}^{n} G_{i+1:j}^{(\eta)}\right\|$ using:

- The triangle inequality
- The geometric bound from Proposition 1
- The Bernoulli inequality for the exponential terms

$$\left\|\sum_{j=i}^{n} G_{i+1:j}^{(\eta)}\right\| = \left\|\sum_{j=i}^{n} (I - \eta\bar{A})^{j-i}\right\| \leq \sum_{j=i}^{n} \left\|(I - \eta\bar{A})\right\|^{j-i} \leq \sum_{j=i}^{n} (1 - \eta a/2)^{j-i} \leq \frac{2}{\eta a} \tag{43}$$

Substituting the result we derive:

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_t^{(1)}\right\|^2 \leq \eta^3 \frac{\operatorname{Tr}(\Sigma_\varepsilon)}{a} V_A \sum_{l=1}^{n} \frac{4}{\eta^2 a^2} = 4n\eta \frac{\operatorname{Tr}(\Sigma_\varepsilon)}{a^3} V_A$$

□

Lemma 6. *Assume A1, A2 and A3. Then for any $t \in \{1, 2, \ldots, n\}$ holds*

$$\mathbb{E}^{1/2}\left\|H_t^{(2)}\right\|^2 \leq \eta^{3/2} \frac{\sqrt{\operatorname{Tr}(\Sigma_\varepsilon)} V_A^{3/2}}{a^{5/2}}$$
$$\mathbb{E}^{1/2}\left\|H_t^{(1)}\right\|^2 \leq \eta^{3/2} \frac{V_A \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}} \left(1 + \frac{V_A^{1/2}}{a}\right)$$

Proof. We begin by substituting results from Lemma 2 and applying the Minkowski inequality. Using the exponential stability from Proposition 2 for $\mathbb{E}^{1/2}\left\|\Gamma_{i+1:t}^{(\eta)}\right\|^2$, we obtain:

$$\mathbb{E}^{1/2}\left\|H_t^{(2)}\right\|^2 = \mathbb{E}^{1/2}\left\|\sum_{i=1}^{t} \eta\Gamma_{i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(2)}\right\|^2 \leq \eta\sqrt{V_A}\sum_{i=1}^{t} \mathbb{E}^{1/2}\left\|\Gamma_{i+1:t}^{(\eta)}\right\|^2 \frac{\eta^{3/2} V_A \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}}$$
$$\leq \frac{\eta^{5/2}\sqrt{\operatorname{Tr}(\Sigma_\varepsilon)} V_A^{3/2}}{a^{3/2}} \sum_{i=1}^{t} (1 - \eta a)^{(t-i)} \leq \frac{\eta^{3/2}\sqrt{\operatorname{Tr}(\Sigma_\varepsilon)} V_A^{3/2}}{a^{5/2}}$$

To establish the bound for $H_t^{(1)}$, we apply Corollary 1 and combine it with the previous result and Lemma 4:

$$\mathbb{E}^{1/2}\left\|H_t^{(1)}\right\|^2 = \mathbb{E}^{1/2}\left\|J_t^{(2)} + H_t^{(2)}\right\|^2 \leq \mathbb{E}^{1/2}\left\|J_t^{(2)}\right\|^2 + \mathbb{E}^{1/2}\left\|H_t^{(2)}\right\|^2$$
$$\leq \sqrt{\frac{\eta^3 V_A^2 \mathrm{Tr}\left(\Sigma_\varepsilon\right)}{a^3}} + \frac{\eta^{3/2}\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}V_A^{3/2}}{a^{5/2}} \leq \eta^{3/2}\frac{V_A\sqrt{\mathrm{Tr}\left(\Sigma_\varepsilon\right)}}{a^{3/2}}\left(1 + \frac{V_A^{1/2}}{a}\right)$$

□

A.2 VRTD algorithm with control variables

This section extends the error decomposition analysis from Appendix A.1 to the VRTD algorithm. We establish results for the modified error components defined in (35) and (36), accounting for the control variate structure.

We begin with the lemma that provides the expression for $J_{k,t}^{(0)}$, which will be further used in estimating higher order terms.

Lemma 7. *A1, A2 and A3, then any $k \in \{1, 2, \ldots, K\}$ and $t \in \{1, 2, \ldots, n\}$ holds*

$$J_{k,t}^{(0)} = \eta \sum_{j=1}^{t} \left(I - \eta\bar{A}\right)^{t-j} (\bar{A} - A_{k,j})(\theta_k' - \theta_{k,0})$$

Proof. Starting from the definition of $J_{k,t}^{(0)}$ in (35), we derive its explicit form through recursive expansion:

$$\begin{aligned} J_{k,t}^{(0)} &= (I - \eta\bar{A})J_{k,t-1}^{(0)} + \eta(\bar{A} - A_{k,t})(\theta_k' - \theta_{k,0}) \\ &= \left(I - \eta\bar{A}\right)^t J_{k,0}^{(0)} + \eta \sum_{j=1}^{t} \left(I - \eta\bar{A}\right)^{t-j} (\bar{A} - A_{k,j})(\theta_k' - \theta_{k,0}) \\ &= \eta \sum_{j=1}^{t} \left(I - \eta\bar{A}\right)^{t-j} (\bar{A} - A_{k,j})(\theta_k' - \theta_{k,0}) \end{aligned}$$

The final equality follows since $J_{k,0}^{(0)} = 0$ by initialization. □

Lemma 8. *Assume A1, A2 and A3, then for any $k \in \{1, 2, \ldots, K\}$ and $t \in \{1, 2, \ldots, n\}$*

$$\mathbb{E}^{1/2}\left\|\sum_{t=1}^{n} J_{k,t}^{(0)}\right\|^2 \leq \frac{\sqrt{2nV_A}\left\|\bar{A}^{-1}\right\|\sqrt{\mathrm{Tr}\left(\Sigma_n\right)}}{\sqrt{N}} + \frac{\sqrt{8n}V_A\left\|\bar{A}^{-1}\right\|\left\|\theta_* - \theta_{k,0}\right\|}{\sqrt{N}a}$$

Proof. This result generalizes Lemma 3 to the control variables setting. Although, the proof

is more complex in this case due to new structure. We begin by transforming the result from Lemma 7 and substituting the definitions (13) and (14).

$$\begin{aligned}\sum_{t=1}^{n} J_{k,t}^{(0)} &= \sum_{t=1}^{n} \eta \sum_{j=1}^{t} \left(I - \eta \bar{A}\right)^{t-j} (\bar{A} - A_{k,j})(\theta'_k - \theta_{k,0}) \\ &= \sum_{j=1}^{n} \eta \sum_{t=j}^{n} G_{j+1:t}^{(\eta)} (\bar{A} - A_{k,j})(\theta'_k - \theta_{k,0}) = \sum_{j=1}^{n} Q_j (\bar{A} - A_{k,j})(\theta'_k - \theta_{k,0})\end{aligned} \tag{44}$$

By definition, all matrices $A_{k,j}$ are mutually independent, and the approximate solution θ'_k is independent from them due to the construction. This independence structure implies that the sequence forms a martingale difference sequence with respect to the natural filtration $\mathcal{F}_{k,t} = \sigma(Z_{k,1}^k, \ldots, Z_{k,N}^k, Z_{k,1}, \ldots, Z_{k,t-1})$, where:

- $Z_{k,i}^k$ represents the samples used to construct θ'_k
- $Z_{k,t}$ denotes the samples used in the t-th update step

For such martingale difference sequences, we can bound the variance using the following decomposition:

$$\begin{aligned}\mathbb{E}\left\|\sum_{j=1}^{n} Q_j (\bar{A} - A_{k,j})(\theta'_k - \theta_{k,0})\right\|^2 &= \sum_{j=1}^{n} \mathbb{E}\left\|Q_j (\bar{A} - A_{k,j})(\theta'_k - \theta_{k,0})\right\|^2 \\ &\leq V_A \sum_{j=1}^{n} \mathbb{E}\left\|Q_j (\theta'_k - \theta_{k,0})\right\|^2\end{aligned}$$

Using the commutation of matrices Q_j and A, definition (33), and Young's inequality, by basic transformations we derive

$$\begin{aligned}&\mathbb{E}\left\|Q_j(\theta'_k - \theta_{k,0})\right\|^2 = \mathbb{E}\left\|Q_j \bar{A}^{-1}(g(\theta_{k,0}) - \widehat{g}(\theta_{k,0)})\right\|^2 \\ &\leq 2\mathbb{E}\left\|Q_j \bar{A}^{-1}\widehat{g}(\theta_*)\right\|^2 + 2\mathbb{E}\left\|Q_j \bar{A}^{-1}(g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}) - g(\theta_*) + \widehat{g}(\theta_*))\right\|^2 \\ &\leq 2\left\|\bar{A}^{-1}\right\|^2 \mathbb{E}\left\|Q_j \widehat{g}(\theta_*)\right\|^2 + \frac{2V_A \left\|Q_j\right\|^2 \left\|\bar{A}^{-1}\right\|^2 \left\|\theta_* - \theta_{k,0}\right\|^2}{N}\end{aligned}$$

Substitute the result back to the expression (44):

$$\begin{aligned}&\mathbb{E}\left\|\sum_{t=1}^{n} J_{k,t}^{(0)}\right\|^2 \leq V_A \sum_{j=1}^{n} \mathbb{E}\left\|Q_j(\theta'_k - \theta_{k,0})\right\|^2 \\ &\leq \frac{2nV_A \left\|\bar{A}^{-1}\right\|^2 \mathrm{Tr}\left(\Sigma_n\right)}{N} + \frac{2V_A^2 \left\|\bar{A}^{-1}\right\|^2 \left\|\theta_* - \theta_{k,0}\right\|^2}{N} \sum_{j=1}^{n} \left\|Q_j\right\|^2\end{aligned}$$

From Lemma 5, we obtain the operator norm bound $\|Q_j\| \leq \frac{2}{a}$ for all j. Substituting this bound into our previous expression yields the following estimate:

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{t=1}^{n} J_{k,t}^{(0)}\right\|^2 = \mathbb{E}\left\|\sum_{j=1}^{n} Q_j(\bar{A} - A_{k,j})(\theta_k' - \theta_{k,0})\right\|^2 \\
&\leq \frac{2nV_A\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_n)}{N} + \frac{2V_A^2\left\|\bar{A}^{-1}\right\|^2\|\theta_* - \theta_{k,0}\|^2}{N}\sum_{j=1}^{n}\|Q_j\|^2 \\
&\leq \frac{2nV_A\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_n)}{N} + \frac{8nV_A^2\left\|\bar{A}^{-1}\right\|^2\|\theta_* - \theta_{k,0}\|^2}{Na^2}
\end{aligned}$$

The proof is completed by applying the Young's inequality.

□

Lemma 9. *Assume A1, A2 and A3. Then for any $k \in \{1, 2, \ldots, K\}$ and $t \in \{1, 2, \ldots, n\}$*

$$\begin{aligned}
\mathbb{E}\left\|J_{k,t}^{(0)}\right\|^2 &\leq \frac{2\eta V_A\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{aN} + \frac{2\eta V_A^2\left\|\bar{A}^{-1}\right\|^2\|\theta_* - \theta_{k,0}\|^2}{aN} \\
\mathbb{E}\left\|J_{k,t}^{(1)}\right\|^2 &\leq \frac{2\eta^2 V_A^2\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^2N} + \frac{2\eta^2 V_A^3\left\|\bar{A}^{-1}\right\|^2\|\theta_* - \theta_{k,0}\|^2}{a^2N} \\
\mathbb{E}\left\|J_{k,t}^{(2)}\right\|^2 &\leq \frac{2\eta^3 V_A^3\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3N} + \frac{2\eta^3 V_A^4\left\|\bar{A}^{-1}\right\|^2\|\theta_* - \theta_{k,0}\|^2}{a^3N}
\end{aligned}$$

Proof. The independence of $\{A_{k,1}^{(k)}, \ldots, A_{k,N}^{(k)}, A_{k,1}, \ldots, A_{k,t}\}$ implies that $J_{k,t}^{(0)}$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}_{k,t} = \sigma(Z_{k,1}^k, \ldots, Z_{k,N}^k, Z_{k,1}, \ldots, Z_{k,t-1})$.

From Lemma 7, we derive the second moment bound:

$$\begin{aligned}
&\mathbb{E}\left\|J_{k,t}^{(0)}\right\|^2 = \mathbb{E}\left\|\eta\sum_{j=1}^{t}\left(I - \eta\bar{A}\right)^{t-j}(\bar{A} - A_{k,j})(\theta_k' - \theta_{k,0})\right\|^2 \\
&\leq \eta^2\sum_{j=1}^{t}\mathbb{E}\left\|\left(I - \eta\bar{A}\right)^{t-j}(\bar{A} - A_{k,j})(\theta_k' - \theta_{k,0})\right\|^2 \leq \eta^2 V_A\|\theta_k' - \theta_{k,0}\|^2\sum_{j=1}^{t}\mathbb{E}\left\|\left(I - \eta\bar{A}\right)\right\|^{2(t-j)}
\end{aligned}$$

The term $\mathbb{E}\|\theta_k' - \theta_{k,0}\|^2$ requires careful estimation. From definition (37) and Young's inequality:

$$\begin{aligned}
&\mathbb{E}\|\theta_k' - \theta_{k,0}\|^2 = \mathbb{E}\left\|\bar{A}^{-1}(g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}))\right\|^2 \\
&\leq 2\left\|\bar{A}^{-1}\right\|^2\mathbb{E}\|\widehat{g}(\theta_*)\|^2 + 2\mathbb{E}\left\|\bar{A}^{-1}(g(\theta_{k,0}) - \widehat{g}(\theta_{k,0}) - g(\theta_*) + \widehat{g}(\theta_*))\right\|^2 \\
&\leq \frac{2\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{N} + \frac{2V_A\left\|\bar{A}^{-1}\right\|^2\|\theta_* - \theta_{k,0}\|^2}{N}
\end{aligned}$$

Combining these results and using the geometric series formula:

$$\mathbb{E}\left\|J_{k,t}^{(0)}\right\|^2 \leq \eta^2 V_A \left(\frac{2\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{N} + \frac{2V_A\left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{N}\right) \sum_{j=1}^{t} \left\|I - \eta\bar{A}\right\|^{2(t-j)}$$
$$\leq \frac{2\eta V_A\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{aN} + \frac{2\eta V_A^2\left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{aN}$$

For $J_{k,t}^{(1)}$, which remains a martingale difference sequence by Lemma 2, we obtain:

$$\mathbb{E}\left\|J_{k,t}^{(1)}\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{t} \eta G_{k,i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(0)}\right\|^2$$
$$\leq \frac{2\eta^2 V_A^2\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^2 N} + \frac{2\eta^2 V_A^3\left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^2 N}$$

The bound for $J_{k,t}^{(2)}$ follows similarly:

$$\mathbb{E}\left\|J_{k,t}^{(2)}\right\|^2 = \mathbb{E}\left\|\sum_{i=1}^{t} \eta G_{k,i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(1)}\right\|^2$$
$$\leq \left(\frac{2\eta^4 V_A^3\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^2 N} + \frac{2\eta^4 V_A^4\left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^2 N}\right) \sum_{i=1}^{t}(1-\eta a)^{t-i}$$
$$\leq \frac{2\eta^3 V_A^3\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N} + \frac{2\eta^3 V_A^4\left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}$$

This completes the proof of all required bounds. □

Lemma 10. *Assume A1, A2 and A3, then for any $k \in \{1, 2, \ldots, K\}$ and $t \in \{1, 2, \ldots, n\}$ holds*

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_{k,t}^{(1)}\right\|^2 \leq \frac{8\eta n V_A^2\left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N} + \frac{8\eta n V_A^3\left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}$$

Proof. This proof extends Lemma 5 to the setting with control variables. As established in the main text, Lemma 2 remains valid in this setting. Since $J_k^{(1)}$ forms a martingale difference sequence and $A_i - \bar{A}$ is independent of $J_{i-1}^{(0)}$, we have:

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_{k,t}^{(1)}\right\|^2 = \mathbb{E}\left\|-\eta\sum_{t=1}^{n}\sum_{i=1}^{t} G_{k,i+1:t}^{(\eta)}(A_i - \bar{A})J_{i-1}^{(0)}\right\| = \mathbb{E}\left\|-\sum_{i=1}^{n} Q_i(A_i - \bar{A})J_{i-1}^{(0)}\right\|$$
$$= \sum_{i=1}^{n} \mathbb{E}\left\|Q_i(A_i - \bar{A})J_{i-1}^{(0)}\right\|^2 \leq \sum_{i=1}^{n} \|Q_i\|^2\, \mathbb{E}\left\|(A_i - \bar{A})\right\|^2 \mathbb{E}\left\|J_{l-1}^{(0)}\right\|^2.$$

We estimate the sum using the bound from Lemma 9 for $\mathbb{E}\left\|J_{i-1}^{(0)}\right\|^2$ and the definition of

Q_i:

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_{k,t}^{(1)}\right\|^2 \leq \sum_{i=1}^{n} \|Q_i\|^2 \mathbb{E}\left\|(A_i - \bar{A})\right\|^2 \mathbb{E}\left\|J_{i-1}^{(0)}\right\|^2$$
$$= \sum_{i=1}^{n} \left\|\sum_{j=i}^{n} G_{i+1:j}^{(\eta)}\right\|^2 \left(\frac{2\eta^3 V_A^2 \left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{aN} + \frac{2\eta^3 V_A^3 \left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{aN}\right).$$

Next, we apply the bound $\left\|\sum_{j=i}^{n} G_{i+1:j}^{(\eta)}\right\| \leq \frac{2}{\eta a}$ received in Lemma 5:, which allows us to finish the proof:

$$\mathbb{E}\left\|\sum_{t=1}^{n} J_{k,t}^{(1)}\right\|^2 \leq \left(\frac{2\eta^3 V_A^2 \left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{aN} + \frac{2\eta^3 V_A^3 \left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{aN}\right) \sum_{l=1}^{n} \frac{4}{\eta^2 a^2}$$
$$\leq \frac{8\eta n V_A^2 \left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N} + \frac{8\eta n V_A^3 \left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}.$$

□

Lemma 11. *Assume A1, A2 and A3. Then for any $k \in \{1, 2, \ldots, K\}$ and $t \in \{1, 2, \ldots, n\}$ holds*

$$\mathbb{E}^{1/2}\left\|H_{k,t}^{(2)}\right\|^2 \leq \frac{\sqrt{2}\eta^{3/2} V_A^2 \left\|\bar{A}^{-1}\right\| \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{5/2}\sqrt{N}} + \frac{\sqrt{2}\eta^{3/2} V_A^{5/2} \left\|\bar{A}^{-1}\right\| \|\theta_* - \theta_{k,0}\|}{a^{5/2}\sqrt{N}}$$
$$\mathbb{E}^{1/2}\left\|H_{k,t}^{(1)}\right\|^2 \leq \frac{\sqrt{2}\eta^{3/2} V_A^{3/2} \left\|\bar{A}^{-1}\right\| \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}\sqrt{N}} \left(I + \frac{V_A^{1/2}}{a}\right)$$
$$+ \frac{\sqrt{2}\eta^{3/2} V_A^2 \left\|\bar{A}^{-1}\right\| \|\theta_* - \theta_{k,0}\|}{a^{3/2}\sqrt{N}} \left(I + \frac{V_A^{1/2}}{a}\right)$$

Proof. We begin by applying the Young's inequality to the expression from Lemma 9 to bound $\mathbb{E}^{1/2}\left\|J_{k,t}^{(2)}\right\|^2$:

$$\mathbb{E}^{1/2}\left\|J_{k,t}^{(2)}\right\|^2 \leq \sqrt{\frac{2\eta^3 V_A^3 \left\|\bar{A}^{-1}\right\|^2 \operatorname{Tr}(\Sigma_\varepsilon)}{a^3 N} + \frac{2\eta^3 V_A^4 \left\|\bar{A}^{-1}\right\|^2 \|\theta_* - \theta_{k,0}\|^2}{a^3 N}}$$
$$\leq \frac{\sqrt{2}\eta^{3/2} V_A^{3/2} \left\|\bar{A}^{-1}\right\| \sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}\sqrt{N}} + \frac{\sqrt{2}\eta^{3/2} V_A^2 \left\|\bar{A}^{-1}\right\| \|\theta_* - \theta_{k,0}\|}{a^{3/2}\sqrt{N}}.$$

For $H_{k,t}^{(2)}$, we invoke Lemma 2 and apply Minkowski's inequality with the exponential sta-

bility from Proposition 2:

$$\mathbb{E}^{1/2}\left\|H_{k,t}^{(2)}\right\|^2 = \mathbb{E}^{1/2}\left\|\sum_{i=1}^{t}\eta\Gamma_{k,i+1:t}^{(\eta)}(A_i-\bar{A})J_{i-1}^{(2)}\right\|^2 \leq \eta\sum_{i=1}^{t}\mathbb{E}^{1/2}\left\|\Gamma_{k,i+1:t}^{(\eta)}(A_i-\bar{A})J_{i-1}^{(2)}\right\|^2$$
$$\leq \frac{\sqrt{2}\eta^{3/2}V_A^2\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{5/2}\sqrt{N}} + \frac{\sqrt{2}\eta^{3/2}V_A^{5/2}\left\|\bar{A}^{-1}\right\|\|\theta_*-\theta_{k,0}\|}{a^{5/2}\sqrt{N}}.$$

Combining these results with Corollary 1 for $l=2$ yields the bound for $H_{k,t}^{(1)}$:

$$\mathbb{E}^{1/2}\left\|H_{k,t}^{(1)}\right\|^2 = \mathbb{E}^{1/2}\left\|J_{k,t}^{(2)}+H_{k,t}^{(2)}\right\|^2 \leq \mathbb{E}^{1/2}\left\|J_{k,t}^{(2)}\right\|^2 + \mathbb{E}^{1/2}\left\|H_{k,t}^{(2)}\right\|^2$$
$$\leq \frac{\sqrt{2}\eta^{3/2}V_A^{3/2}\left\|\bar{A}^{-1}\right\|\sqrt{\operatorname{Tr}(\Sigma_\varepsilon)}}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right) + \frac{\sqrt{2}\eta^{3/2}V_A^{2}\left\|\bar{A}^{-1}\right\|\|\theta_*-\theta_{k,0}\|}{a^{3/2}\sqrt{N}}\left(I+\frac{V_A^{1/2}}{a}\right).$$

□

B Auxiliary results

This section contains supporting technical results used in the proofs of our main theorems. We establish bounds on key quantities and derive important inequalities that facilitate the analysis in previous sections.

Lemma 12. *The following representations hold for Q_t:*

$$Q_t = \bar{A}^{-1} - \bar{A}^{-1}(I-\eta\bar{A})^{n-t+1}$$

The sum of the differences between Q_t and $\bar{A}^{-1}$ is given by

$$\sum_{t=1}^{n}(Q_t-\bar{A}^{-1}) = -\bar{A}^{-1}\sum_{t=1}^{n}(I-\eta\bar{A})^t$$

Additionally, the difference between the empirical covariance matrix Σ_n and the limiting covariance Σ_∞ can be written as

$$\Sigma_n-\Sigma_\infty = \underbrace{\frac{1}{n}\sum_{t=1}^{n}(Q_t-\bar{A}^{-1})\Sigma_\varepsilon\bar{A}^{-\top} + \frac{1}{n}\sum_{t=1}^{n}\bar{A}^{-1}\Sigma_\varepsilon(Q_t-\bar{A}^{-1})^\top}_{D_1} +$$
$$+\underbrace{\frac{1}{n}\sum_{t=1}^{n}(Q_t-\bar{A}^{-1})\Sigma_\varepsilon(Q_t-\bar{A}^{-1})^\top}_{D_2}$$

Proof. To derive a representation for Q_t, we begin by rewriting its definition (14) and applying algebraic manipulations:

$$Q_t = \eta \sum_{i=t}^{n} G_{t+1:i}^{(\eta)} = \eta(I - \eta\bar{A})^0 + \eta \sum_{i=t+1}^{n} (I - \eta\bar{A})^{i-t} = \eta I + \eta \sum_{i=t+1}^{n} (I - \eta\bar{A})^{i-t}. \tag{45}$$

Multiplying both sides by $(I - \eta\bar{A})$, we obtain:

$$(I - \eta\bar{A})Q_t = (I - \eta\bar{A})\eta \sum_{i=t}^{n} (I - \eta\bar{A})^{i-t} = \eta(I - \eta\bar{A})^{n-t+1} + \eta \sum_{i=t+1}^{n} (I - \eta\bar{A})^{i-t}. \tag{46}$$

Subtracting (46) from (45), we derive:

$$\eta\bar{A}Q_t = Q_t - (I - \eta\bar{A})Q_t = \eta I - \eta(I - \eta\bar{A})^{n-t+1},$$

which yields the desired representation:

$$Q_t = \bar{A}^{-1} - \bar{A}^{-1}(I - \eta\bar{A})^{n-t+1}.$$

Based on the derived representation, we obtain:

$$\begin{aligned} Q_t - \bar{A}^{-1} &= -\bar{A}^{-1}(I - \eta\bar{A})^{n-t+1} \\ \sum_{t=1}^{n}(Q_t - \bar{A}^{-1}) &= -\bar{A}^{-1}\sum_{t=1}^{n}(I - \eta\bar{A})^{n-t+1} = -\bar{A}^{-1}\sum_{t=1}^{n}(I - \eta\bar{A})^{t} \end{aligned}$$

Recall that the limiting covariance is given by:

$$\Sigma_\infty = \bar{A}^{-1}\Sigma_\varepsilon\bar{A}^{-T}$$

Using the definition (15), the difference between the empirical and limiting covariances can be written as:

$$\begin{aligned} \Sigma_n - \Sigma_\infty &= \frac{1}{n}\sum_{t=1}^{n}\left(Q_t\Sigma_\varepsilon Q_t^T - \bar{A}^{-1}\Sigma_\varepsilon\bar{A}^{-T}\right) \\ &= \frac{1}{n}\sum_{t=1}^{n}\left(Q_t - \bar{A}^{-1}\right)\Sigma_\varepsilon\bar{A}^{-T} + \frac{1}{n}\sum_{t=1}^{n}\bar{A}^{-1}\Sigma_\varepsilon\left(Q_t - \bar{A}^{-1}\right)^T + \frac{1}{n}\sum_{t=1}^{n}\left(Q_t - \bar{A}^{-1}\right)\Sigma_\varepsilon\left(Q_t - \bar{A}^{-1}\right)^T \end{aligned}$$

For convenience, we define the following terms, which will be analysed separately:

$$D_1 = \frac{1}{n}\sum_{t=1}^{n}\left(Q_t - \bar{A}^{-1}\right)\Sigma_\varepsilon \bar{A}^{-T} + \frac{1}{n}\sum_{t=1}^{n}\bar{A}^{-1}\Sigma_\varepsilon\left(Q_t - \bar{A}^{-1}\right)^T$$
$$D_2 = \frac{1}{n}\sum_{t=1}^{n}\left(Q_t - \bar{A}^{-1}\right)\Sigma_\varepsilon\left(Q_t - \bar{A}^{-1}\right)^T$$

□

Proposition 3. *Suppose the conditions of Proposition 1 hold for η. Therefore*

$$\left\|(\mathrm{I} - \eta\bar{A})^n\right\| \le \cdot\left(1 - \frac{a}{2}\cdot\eta\right)^n \le \exp\left(-\frac{na\eta}{2}\right)$$

Proof. Applying Proposition 1 and combining inequalities $\sqrt{1-x} \le 1 - \frac{x}{2}$ and $1 - x \le \exp(-x)$, we get

$$\left\|(\mathrm{I} - \eta\bar{A})\right\|^n \le \left(\sqrt{\left\|\mathrm{I} - \eta\bar{A}\right\|^2}\right)^n \le \left(\sqrt{1-a\eta}\right)^n \le \left(1 - \frac{a\eta}{2}\right)^n \le \exp\left(-\frac{a\eta}{2}\right)^n = \exp\left(-\frac{a}{2}n\eta\right)$$

□

Proposition 4. *Under the assumptions of Proposition 1 it holds*

$$\left\|\sum_{t=1}^{n} G_{1:t}^{(\eta)}\right\| = \left\|\sum_{t=1}^{n}(I - \eta\bar{A})^t\right\| \le \frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}}$$

Proof. The triangle inequality and Proposition 3 immediately imply that

$$\left\|\sum_{t=1}^{n} G_{1:t}^{(\eta)}\right\| \le \sum_{t=1}^{n}\left\|G_{1:t}^{(\eta)}\right\| = \sum_{t=1}^{n}\left\|(I - \eta\bar{A})^t\right\| \le \sum_{t=1}^{n}\exp\left(-\frac{t\eta a}{2}\right) = \frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}}$$

□

Lemma 13. *Under assumptions A1, A2, A3 the following bound holds for any natural n:*

$$\left\|\Sigma_n - \Sigma_\infty\right\| \le \frac{2\left\|\Sigma_\infty\right\|}{n}\cdot\frac{e^{-\frac{a\eta}{2}}(1 - e^{-\frac{an\eta}{2}})}{1 - e^{-\frac{a\eta}{2}}} + \frac{\left\|\Sigma_\infty\right\|}{n}\frac{e^{-an\eta}(1 - e^{-an\eta})}{1 - e^{-a\eta}}$$

Proof. We use the representation obtained in Lemma 12

$$\Sigma_n - \Sigma_\infty = \underbrace{\frac{1}{n}\sum_{t=1}^{n}(Q_t - \bar{A}^{-1})\Sigma_\varepsilon\bar{A}^{-\top} + \frac{1}{n}\sum_{t=1}^{n}\bar{A}^{-1}\Sigma_\varepsilon(Q_t - \bar{A}^{-1})^\top}_{D_1} + \underbrace{\frac{1}{n}\sum_{t=1}^{n}(Q_t - \bar{A}^{-1})\Sigma_\varepsilon(Q_t - \bar{A}^{-1})^\top}_{D_2}$$

First, we will bound D_1. The operator norms of both terms are equal because one is a transposed version of another, so it is sufficient to bound only one of them. Note that $G^{(\eta)}_{n:m}$, Q_t, $\bar{A}$, $\bar{A}^{-1}$, commute as polynomials in $\bar{A}$. Remember another representation from Lemma 12 and obtain

$$\left\|\frac{1}{n}\sum_{t=1}^{n}(Q_t-\bar{A}^{-1})\Sigma_\varepsilon\bar{A}^{-\top}\right\| = \left\|-\frac{1}{n}\bar{A}^{-1}\sum_{j=1}^{n}(I-\eta\bar{A})^t\Sigma_\varepsilon\bar{A}^{-\top}\right\|$$
$$= \left\|n^{-1}\Sigma_\infty\sum_{j=1}^{n}(I-\eta\bar{A})^t\right\| \le n^{-1}\left\|\Sigma_\infty\right\|\cdot\left\|\sum_{j=1}^{n}(I-\eta\bar{A})^t\right\|$$

Proposition 4 directly imply the bound for D_1:

$$\left\|\frac{1}{n}\sum_{t=1}^{n}(Q_t-\bar{A}^{-1})\Sigma_\varepsilon\bar{A}^{-\top}\right\| \le n^{-1}\left\|\Sigma_\infty\right\|\cdot\left\|\sum_{j=1}^{n}G_{1:j}\right\| \le n^{-1}\left\|\Sigma_\infty\right\|\cdot\frac{e^{-\frac{a\eta}{2}}(1-e^{-\frac{an\eta}{2}})}{1-e^{-\frac{a\eta}{2}}}$$

Hence

$$\|D_1\| \le \frac{2}{n}\left\|\Sigma_\infty\right\|\cdot\frac{e^{-\frac{a\eta}{2}}(1-e^{-\frac{an\eta}{2}})}{1-e^{-\frac{a\eta}{2}}}$$

We analyze the term D_2 using the representation from Lemma 12:

$$D_2 = n^{-1}\sum_{t=1}^{n}(Q_t-\bar{A}^{-1})\Sigma_\varepsilon(Q_t-\bar{A}^{-1})^\top = n^{-1}\sum_{t=1}^{n}\left(-\bar{A}^{-1}(I-\eta\bar{A})^{n-t+1}\right)\Sigma_\varepsilon\left(-\bar{A}^{-1}(I-\eta\bar{A})^{n-t+1}\right)^\top$$
$$= n^{-1}\sum_{t=1}^{n}\bar{A}^{-1}(I-\eta\bar{A})^{n-t+1}\Sigma_\varepsilon\left((I-\eta\bar{A})^{n-t+1}\right)^\top\bar{A}^{-\top}$$

Applying Proposition 3, we obtain the following bound for the operator norm $\|D_2\|$:

$$\|D_2\| = \left\|n^{-1}\sum_{t=1}^{n}\bar{A}^{-1}(I-\eta\bar{A})^{n-t+1}\Sigma_\varepsilon\left((I-\eta\bar{A})^{n-t+1}\right)^\top\bar{A}^{-\top}\right\|$$
$$\le \frac{1}{n}\sum_{t=1}^{n}\left\|\bar{A}^{-1}(I-\eta\bar{A})^{n-t+1}\Sigma_\varepsilon\left((I-\eta\bar{A})^{n-t+1}\right)^\top\right\|\bar{A}^{-\top}$$
$$= \frac{1}{n}\sum_{t=1}^{n}\left\|(I-\eta\bar{A})^{n-t+1}\Sigma_\infty\left((I-\eta\bar{A})^{n-t+1}\right)^\top\right\|$$
$$\le \frac{\|\Sigma_\infty\|}{n}\sum_{t=1}^{n}\left\|(I-\eta\bar{A})^{n-t+1}\right\|^2 \le \frac{\|\Sigma_\infty\|}{n}\sum_{t=1}^{n}\exp\left(-\frac{at\eta}{2}\right)^2 = \frac{\|\Sigma_\infty\|}{n}\frac{e^{-a\eta}(1-e^{-an\eta})}{1-e^{-a\eta}}$$

□