Development and Theoretical Analysis of the Algorithms for Optimal Control and Reinforcement Learning

Maksim Kaledin

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX011

Thèse de doctorat

Development and Theoretical Analysis of the Algorithms for Optimal Control and Reinforcement Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École polytechnique

École doctorale n°574 École Doctorale de Mathématique Hadamard (EDMH)

en cotutelle avec Higher School of Economics (HSE University, Moscow)

Doctoral School of Computer Science

Spécialité de doctorat : Mathématiques Appliquées

Thèse présentée et soutenue à Palaiseau, le 25.01.2023, par

MAKSIM KALEDIN

Composition du Jury :

Peter TANKOV Professeur, ENSAE, Institut Polytechnique de Paris (Center for research in economics and Statistics)	Président
Alexander GOLDENSHLUGER Professeur, Department of Statistics, University of Haifa	Rapporteur
Tony LELIÈVRE Directeur de recherche, l'Ecole des Ponts ParisTech (Centre d'Enseignement et de Recherche en MathématIques et Calcul Scientifique (CERMICS))	Rapporteur
Alexey NAUMOV Directeur de recherche, HSE University (Laboratory of Stochastic Analysis and High-Dimensional Inference)	Examinateur
Gersende FORT Directrice de recherche, Institut de Mathématiques de Toulouse	Examinateur
Eric MOULINES Professeur, École Polytechnique, Institut Polytechnique de Paris (Centre de Mathématique Appliquée(CMAP))	Directeur de thèse
Denis BELOMESTNY Professeur, University of Duisburg-Essen, HSE University	Co-directeur de thèse
Emmanuel GOBET Professeur, École Polytechnique, Institut Polytechnique de Paris (Centre de Mathématique Appliquée(CMAP))	Invité

Acknowledgements

This thesis would not be possible without mentorship of my professors. First of all, I would like to express my deep gratitude to my scientific supervisors, professors Éric Moulines and Denis Belomestny, who guided me to the final destination all these three years recommending solutions and recommending new problems to explore further. I would not achieve any of my published papers and curious results without them. Their expertise, the skills I learned, and the experience I got are of incredible importance to all my efforts. This very words would not have been written if I did not have their support during these times despite all the world challenges making the live in-person cooperation difficult. I deeply thank also my coauthors: professors John Schoenmakers, Hoi-To Wai and Vladislav Tadić, with whom we had a very intensive collaboration.

Certainly, I cannot thank enough Centre de Mathématique Appliquée and École Polytechnique. I was always given help and expert advices from the staff, the colleagues and fellow PhD students. Even in discouraging times of lockdown we still were able to have seminars, discuss the ideas concerning machine learning and statistics. Specifically, I would like to give my acknowledgement to Seminaire Palaisien, first organized in the beginning of 2020, where I learned a lot from various talks. Also I would like to thank my collaborator and friend Corentin Houpert and with whom I learned a lot about the connection between Russian and French Mathematical society through the bridge built by Vladimir Arnold. The only thing I am very sorry about is that I was not able to spend more time in Paris than I did.

Equally, I would like to thank HDI Lab of HSE University for a lot of interesting and mind-blowing projects and research we had during these three years in an incredible team of scientists of all levels. Also, I would like to express my personal gratitude to professor and the director of the lab Alexey Naumov who despite all the challenges of my PhD path encouraged me to go further and achieve the results. A lot of thanks I would like also to send to the Faculty of Computer Science at HSE.

Special thanks goes to my colleagues and acquaintance with whom we pursued or at least learned a lot together or had a very instructive discussions: professor Vladimir Ulyanov, associate professors Bruno Bauwens, Quentin Paris and Laurent Beaudou, researchers Alexander Golubev, Leonid Iosipoi, Sergey Samsonov, Daria Demidova, Daniil Tyapkin, Ilya Levin and many other colleagues and friends from HSE University.

Without any of them this text and my defense would have never been possible.

Contents

Summary

Introduction (Français)

Les problèmes de contrôle optimal stochastique sont très souvent rencontrés dans plusieurs applications pratiques: de mathématique financière [35, 77] à l'ingénierie [12]. Récemment, ils ont reçu une nouvelle attention et les nouvelles perspectives trouvé en apprentissage par renforcement (Reinforcement Learning, RL), qui est dans un certain sens se présente comme l'intersection de contrôle optimal, statistiques et l'apprentissage automatique [70].

Cette classe des problèmes peut être définie comme suit. Soit $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$ un espace de probabilité filtré avec la filtration $(\mathcal{F}_t)_{t\geq 0}$. Supposons quelque ensemble $\mathcal{U}$ de processus stochastiques mesuré $U : \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}^n$ qui s'appellent *contrôles* et l'ensemble de *processus contrôlés*

$$\mathcal{X} = \left\{X_t^U : U \in \mathcal{U}\right\}$$

où pour chaque contrôle U l'élément $(X_t^U)_{t\geq 0}$ a sa valeurs dans $\mathbb{R}^d$ et est un processus stochastique $(\mathcal{F}_t)_{t\geq 0}$-adapté. On définissons le fonctionnel $J : \mathcal{X} \to \mathbb{R}$ et le nommons *le fonctionnel de gain.*

Definition 1. *Le problème de trouver* $U_* \in \operatorname{Arg\,max}_{U\in\mathcal{U}} J(X^U)$ *est appelé le problème de contrôle optimal stochastique.*

Également, dans les applications pratiques (spécifiquement dans le domaine de l'apprentissage par renforcement [70]) nous avons besoin d'évaluation de la règle de décisions donnée. Par exemple, dans les algorithmes d'itération de politique, où l'évaluation est le composant crucial.

Definition 2. *Le problème d'évaluation de* $J(X^U)$ *avec le contrôle* U *donné est appelé le problème d'évaluation du contrôle.*

Certainement, ce n'est pas possible de faire plus avec une telle formulation très abstraite. Par exemple, nous ne pouvons pas prouver l'existence des solutions ou leur qualités. La question est plus simple quand nous considérons les formulations plus spécifiques. Dans cette thèse les deux problèmes sont considérés: le problème d'arrêt optimal pour un équation différentielle stochastique (EDS) et le problème de processus décisionnel de Markov (Markov Decision Problem, MDP).

Problem 1. (L'arrêt optimal pour EDS, [77, 35]) Soit $T > 0$ et le processus X_t est défini par l'EDS d'Ito pour $t \in [0, T)$

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{1}$$

avec $X_0^U = x_0 \in \mathbb{R}^d$, où les fonctions

$$b : [0,T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^d, \quad \sigma : [0,T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^{d \times n}$$

sont continues et de Lipschitz dans le deuxième argument est de croissance linéaire avec quelque constante K:

$$\|b(t,x,u)\|_2 + \|\sigma(t,x,u)\|_2 \leq K(1 + \|x\|_2 + \|u\|_2)$$

où $\|\cdot\|_2$ note la 2-norme euclidienne. Avec ces suppositions, nous avons maintenant la possibilité de prouver l'existence et unicité de la solution. Soit $g_t : \mathbb{R} \to \mathbb{R}$ pour $t \in [0,T]$ est quelque fonction nommé *payoff*. Considérons l'agent qui observe le processus, à $t' \in [0,T]$ il connait les valeurs de X_t pour les temps $t \leq t'$. Son objectif est choisir le temps τ pour exécuter quelque action (arrêter le processus) que lui donnera un payoff $g_\tau(X_\tau)$. Autrement dit, nous sommes intéressés à la choix de temps d'arrêt τ avec valeurs dans $[0,T]$ de l'ensemble des temps d'arrêt admissibles $\mathcal{T}$ qui maximisera l'espérance de payoff:

$$\tau_* = \arg\max_{\tau \in \mathcal{T}} \mathbb{E}\left[g_\tau(X_\tau)\right].$$

Les méthodes plus populaires en pratiques sont inventées avec les idées des algorithmes de Longstaff-Schwarz(LS)[53] et Tsitsiklis-Van Roy [80]. Ils utilisent le principe de la programmation dynamique et approchent l'espérance conditionnelle via régression linéaire et la méthode des moindres carrés avec une base des fonction fixé en chaque étape de la récursion. Longstaff et Schwarz ont démontré l'efficacité de leur approche par nombreux expériences numériques et en [22] et [89] les propriétés générales de la convergence des méthodes étaient établis.

Problem 2. (Le processus décisionnel de Markov, MDP, [70]) Considérons $\mathcal{S}, \mathcal{A}$ nommés les espaces d'*état* et d'*action* (ils doit être les espaces mesurable) et on définissons la chaîne de Markov S_t à la manière suivant. Soit Π est l'ensemble des règles de décision stochastiques (autrement dit, la *politique*) $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, i.e. chaque politique transforme l'état $s \in \mathcal{S}$ et donne la distribution de probabilité dans l'espace des actions noté comme $\pi(\cdot|s)$. Supposons la *fonction de transition* $P(\cdot|s,a)$, une distribution de probabilité dans l'espace d'états avec condition d'état et d'action en cours fixé. On défini $S_0 = s_0$ presque sûrement et après l'état se changera de S_t à S_{t+1} par l'utilisation de la formule des itérations suivant:

$$\begin{aligned} A_t &\sim \pi(\cdot|S_t), \\ S_{t+1} &\sim P(\cdot|S_t, A_t). \end{aligned}$$

Considérons la fonction des récompenses $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ déterministe et uniformément borné. L'illustration naturel de MDP est ce que nous avons l'agent dans l'environnement avec les états décrits par les éléments de $\mathcal{S}$; l'agent à chaque temps t doit exécuter certaine décision A_t par utilisation de sa politique, après il reçoit une récompense $R(S_t, A_t)$ et l'environnement se change a la manière décrite. Le problème de contrôle optimal consiste en maximisation de la somme des récompenses dévaluée

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(S_t, A_t)\right]$$

par rapport à la politique, où $\gamma \in (0, 1)$ fonctionne comme le facteur de dévaluation et l'horizon T peut être fini (le problème d'horizon fini) ou infini (le problème d'horizon infini), ou stochastique (le problème épisodique). MDP se présente comme le modèle fondamental dans l'apprentissage par renforcement(Reinforcement Learning, RL), que est maintenant fortement en développement avec les résultats prometteurs et nombreux applications en plusieurs activités de la société: à partir de l'intelligence artificielle pour les jeux d'ordinateur [82, 11, 66] à systèmes de gestion de l'énergie[49, 32], fabrication et robots [2] pour n'en nommer que quelques-uns. Naturellement, RL donne aux ingénieurs les nouvelles ensembles des outils de contrôle pour un utilisation en tout type d'automatisation [33].

L'évaluation de la politique est une parte vitale des algorithmes model-free basé sur l'itération de politique et normalement utilise les schémas d'Approximation Stochastique (Stochastic Approximation, SA), inventés par [62]. SA elle-même est maintenant devenu une belle technique [10, 47, 15], mais RL donne les problèmes et les suppositions nouvelles. Entre les autres, les schémas SA linéaires sont populaires dans l'apprentissage par renforcement car ils mènent à méthodes d'évaluation de politique avec une approximation linéaire de la fonction des valeurs, les méthodes de Temporal Difference (TD) learning [69], pour lesquels les analyses à temps fini sont rapportés en [68, 48, 13, 25], sont particulièrement importants.

Les objectifs de la recherche

L'objectif principal de notre recherche est étudier les problèmes mentionnés.

1. Concernant le problème d'arrêt optimal décrit en Section 1, nous présentons l'analyse de la complexité de la méthode de maillage stochastique (Weighted Stochastic Mesh, WSM) similaire à la méthode de [17] pour les problèmes d'arret optimal en temps discret et continu, et nous comparons WSM avec les autres méthodes populaires par introduction de la mesure de la complexité nouvelle car par rapport à la mesure de la complexité classique tous les algorithmes pour le problème d'arrêt optimal sont intraitable et il n'y a pas des possibilités de les comparer par rapport à la complexité.

2. Dans Section 2 notre objectif est obtenir l'analyse de la convergence à temps fini pour le schéma d'approximation stochastique linéaire aux deux échelles de temps sous l'hypothèse du bruit de Markov. Avec cet hypothèse c'est exactement le cas des algorithmes d'évaluation de politique pour MDP: Temporal Difference learning (TD(0) de [69]) et les algorithmes de Gradient Temporal Différence (GTD[72],GTD2 et TDC [73]). Le problème avec l'analyse que existe est que la nature de Markov de data n'était pas considéré (malgré que les praticiens travaillent avec MDP, où c'est le cas naturel) ou les suppositions sont plus restrictif.

3. Enfin, dans Section 3 nous proposons la méthode nouvelle, construit pour la réduction de la variance et basé sur la minimisation de la variance empirique présenté en [8], en cas des algorithmes de Policy-Gradient. L'objectif est, d'abord, obtenir l'algorithme que peux rendre l'amélioration de performance supplémentaire à comparaison avec l'objectif classique pour les variables de contrôle présentés en algorithme A2C (Advantage Actor-Critic) [74] et, deuxièmement, présenter quelques garanties théoriques de la réduction.

Les résultats clés

1. Dans la première direction, nous présentons pour la première fois l'analyse de la complexité d'algorithme WSM basé sur [17] et considérons aussi le cas de densité de probabilité $p(x|y)$ inconnue mais que peut être approché. Nous proposons la mesure nouvelle pour la comparaison des algorithmes d'arrêt optimal qui s'appelle *l'indice de semitraitabilité (semitractability index, ST)* et nous lui utilisons pour la comparaison des algorithmes de Longstaff et Schwartz [53] et la méthode QTM [7].

2. Nous obtient les taux de convergences améliorés pour SA linéaire aux deux échelles de temps en cas du bruit martingale et de Markov. Notre analyse nous permet l'utilisation des pas de temps plus générales, particulièrement, les pas constants, constants par morceaux, et décroissantes étudiés dans les articles précédentes [40, 24, 88, 27]. Contrairement aux articles antérieurs [51, 24, 88], nos résultats de convergence sont obtenus *sans* l'inclusion de la projection dans les itérations de SA. Enfin, avec les suppositions supplémentaires pour les pas de temps, nous avons calculé l'expansion asymptotique des erreurs quadratique attendues et montré que nos bornes sont supérieures.

3. Nous construisons des nouvelles méthodes de policy-gradient (méthodes EV) basé sur la critère de la variance empirique et nous montrons que ces méthodes fonctionnent bien dans quelque problèmes pratiques en comparaison avec la critère d'A2C. Aussi, nous proposons les bornes théoriques de la variance de l'estimation de gradient pour les méthodes d'EV par l'utilisation des techniques de [8]; c'est le première résultat que se concerne des bornes de la variance probabilistes, obtenu avec les outils de l'apprentissage statistique dans le domaine de RL. Les mesures de la variance d'estimation de gradient nous montre quelque observations. D'abord, les méthodes d'EV peux réduire la variance mieux que A2C. Deuxièmement, nous avons vu quelque confirmations de la hypothèse de [81]: la réduction de la variance a les effets mais quelques environnements ne sont pas si réactifs à la. Nous présentons les études premières de la critère d'EV pour les méthodes de policy-gradient dans les exemples classiques et nous avons présenté pour la première fois l'implémentation de ces méthodes basé sur PyTorch.

La contribution de l'Auteur. Quelques partes de l'analyse en cas de temps discret, le transfert de temps discret à temps continu, les implémentations et les expériences numériques de l'article 1 sont faites par l'Auteur. Dans article 2 l'Auteur a fait un travail conséquent de préparation de la revue de la littérature et contribué aux résultats pour le case martingale; il aussi a présenté les résultats et les illustrations numériques. Dans l'article 3 l'Auteur a fait les démonstrations des théoremes concernant les bornes probabilistes, vérification des suppositions, la revu de littérature, et participé à l'implémentation des algorithmes et contribué à la concept des expériences.

Introduction

Stochastic optimal control problems are very often encountered in various practical areas: from finance [35, 77] to engineering [12]. Recently they have got a new attention and new challenges in the light of developing Reinforcement Learning (RL), in some sense presenting itself as the intersection of optimal control, statistics and machine learning [70].

Such class of problems can be defined as follows. Let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$ be a filtered probability space with filtration $(\mathcal{F}_t)_{t\geq 0}$. Assume some set $\mathcal{U}$ of progressively measurable stochastic processes $U : \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}^n$ called *controls* and set of *controlled processes*

$$\mathcal{X} = \left\{ X_t^U : U \in \mathcal{U} \right\}$$

where for every control U each $(X_t^U)_{t\geq 0}$ is an $\mathbb{R}^d$-valued $(\mathcal{F}_t)_{t\geq 0}$-adapted stochastic process. We also set functional $J : \mathcal{X} \to \mathbb{R}$ and call it *gain functional*.

Definition 3. *The problem of searching* $U_* \in \operatorname{Arg\,max}_{U\in\mathcal{U}} J(X^U)$ *is called stochastic optimal control problem.*

Also in practice (especially in reinforcement learning, see [70]) as a technical module of some algorithms it is needed to evaluate the given decision rule and so one gets an evaluation problem.

Definition 4. *The problem of evaluating* $J(X^U)$ *given a control* U *in some form is called control evaluation problem.*

Of course, with such abstract formulation we cannot claim anything about the existence of the solutions or their qualities. The question becomes much more clear when we consider more specific formulations. In the thesis the two more specific problems are considered: optimal stopping for a stochastic differential equation(SDE) and Markov Decision Problem (MDP).

Problem 3. (Optimal stopping problem for an SDE, [77, 35]) Assume $T > 0$ and let process X_t be set with an Ito SDE for $t \in [0, T)$

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{2}$$

with initial condition $X_0^U = x_0 \in \mathbb{R}^d$, where functions

$$b : [0, T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^d, \quad \sigma : [0, T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^{d\times n}$$

are two continuous functions satisfying Lipschitz condition in the second argument and linear growth condition with constant K:

$$\|b(t, x, u)\|_2 + \|\sigma(t, x, u)\|_2 \leq K(1 + \|x\|_2 + \|u\|_2)$$

with $\|\cdot\|_2$ denoting the appropriate Euclidean 2-norm. With such assumption we may ensure that the unique strong solution exists. Let $g_t : \mathbb{R} \to \mathbb{R}$ for every $t \in [0, T]$ be some function called *payoff*. Consider an agent observing the process, at time $t' \in [0, T]$ he knows the values of X_t for all $t \leq t'$. His goal is to choose the time τ when to take one particular decision (stop the process, as it is often called) which gives him payoff $g_\tau(X_\tau)$. Formally, we are interested in choosing a stopping time τ taking values in $[0, T]$ from the

set of admissible stopping times $\mathcal{T}$ maximizing the expected discounted reward of the agent:

$$\tau_* = \arg\max_{\tau \in \mathcal{T}} \mathbb{E}\left[g_\tau(X_\tau)\right].$$

The most adopted by practitioners methods are invented with the ideas of Longstaff-Schwarz(LS)[53] and Tsitsiklis-Van Roy [80] algorithms in mind. They exploit dynamic programming principle and approximate conditional expectations using least-squares regression on a given basis of functions on each backward induction step. Longstaff and Schwarz demonstrated the efficiency of their approach through a number of numerical examples and in [22] and [89] general convergence properties of the method were established.

Problem 4. (Markov Decision Process, MDP, [70]) Assume some sets $\mathcal{S}, \mathcal{A}$ called *state* and *action* spaces (they have to be measurable spaces) and define discrete-time time-homogenuous Markov chain S_t as follows. Let there be Π, the set of stochastic decision rules (also called *policies*) $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, i.e. each policy takes the state $s \in \mathcal{S}$ and returns probability distribution over the action space denoted as $\pi(\cdot|s)$. Let us set *transition kernel* $P(\cdot|s,a)$ as a probability distribution over the state space given the current state and action. Set $S_0 = s_0$ almost surely and then iteratively update S_t to S_{t+1} using the following scheme:

$$A_t \sim \pi(\cdot|S_t),$$
$$S_{t+1} \sim P(\cdot|S_t, A_t).$$

Consider a deterministic uniformly bounded reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The natural illustration of MDP is that we have an agent in the environment with state descriptions from $\mathcal{S}$; the agent at each time t must make a decision A_t using his policy, after that he receives a reward $R(S_t, A_t)$ and the environment changes its state as shown above. The optimal control problem is to maximize with respect to policy the expected sum of discounted rewards

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(S_t, A_t)\right],$$

where $\gamma \in (0,1)$ plays the role of the discounting factor and horizon T can be finite (finite-horizon problem) or infinite (infinite-horizon problem), or even random (episodic problem). MDP is a fundamental model in Reinforcement Learning(RL) being currently a fast-developing area with promising and existing applications in numerous innovative areas of the society: starting from AI for games [82, 11, 66] and going to energy management systems [49, 32], manufacturing and robotics [2] to name a few. Naturally, RL gives the practitioners new sets of control tools for any kind of automatization [33].

Policy evaluation is a vital part of the model-free algorithms based on policy iteration and it is normally based on Stochastic Approximation(SA) schemes, invented in [62]. SA itself currently became a well-studied technique [10, 47, 15], however RL gives new challenges and new assumptions. Among others, linear SA schemes are popular in reinforcement learning (RL) as they lead to policy evaluation methods with linear function approximation, of particular importance is temporal difference (TD) learning [69] for which finite time analysis has been reported in [68, 48, 13, 25].

Aim of the Work

The aim of our research is to investigate the problems above in several ways.

1. Regarding the optimal stopping problem discussed in Section 1, we are aiming at presenting the complexity analysis of Weighted Stochastic Mesh(WSM) algorithm similar to the method of [17] for discrete- and continuous-time optimal stopping problem and compare it to other popular methods via new complexity metric since with respect to classic complexity metric all algorithms for optimal stopping are intractable and there is no way to compare them taking the complexity into account.

2. In Section 2 we aimed at obtaining finite-time convergence analysis for two-timescale linear Stochastic Approximation(SA) scheme under Markov noise assumptions. Such setting is exactly the setting of classic policy evaluation algorithms for MDP: temporal difference learning (TD(0) of [69]) and gradient temporal difference algorithms (GTD[72],GTD2 and TDC [73]). The problem with existing analysis is that it does not consider the Markov nature of the data (which is a natural thing since practitioners work in MDP setting) or the assumptions are too restrictive.

3. Finally, in Section 3 we set up to propose a new method for variance reduction based on empirical variance minimization of [8] in policy-gradient algorithms. The goal is, firstly, to obtain an algorithm able to give the improvement over the classic optimization goal for control variates in Advantage Actor-Critic(A2C) schemes [74] and, secondly, give some theoretical guarantees regarding the actual variance reduction.

Key Results

1. To address the first aim, we present for the first time the complexity analysis of WSM algorithm based on [17] and consider also the case when the transition density $p(x|y)$ is not known but can be approximated. We propose a new metric for comparison of the algorithms for optimal stopping problems called *semitractability index* and compare with it several algorithms popular in the community of practitioners: LS-algorithm [53] and QTM [7].

2. We provide improved convergence rates for the linear two-timescale SA in both martingale and Markovian noise settings. Our analysis allow for general step sizes schedules, including constant, piecewise constant, and diminishing step sizes explored in the prior works [40, 24, 88, 27]. Unlike the prior works [51, 24, 88], our convergence results are obtained *without* requiring a projection step throughout the SA iterations. Finally, with an additional assumption on the step size, we compute an exact asymptotic expansion of the expected squared error to show the tightness of our upper bounds.

3. We provide two new policy-gradient methods (EV-methods) based on EV-criterion and show that they perform well in several practical problems in comparison to A2C-criterion. Also theoretical variance bounds for EV-methods are provided using the ideas of [8], this the first result concerning the variance bounds with high probability with the help of the tools of statistical learning in the setting of RL. Measurements of the variance of the gradient estimates present several somewhat

surprising observations. Firstly, EV-methods are able to solve variance reduction problem considerably better than A2C. Secondly, we see some confirmations of the hypothesis of [81]: variance reduction has its effect but some environments are not so responsive to this. We present the first experimental investigation of EV-criterion of policy-gradient methods in classic benchmark problems and the first implementation of it in the framework of PyTorch.

Author contribution. Some part of the analysis for discrete-time case, transfer from discrete to continuous case, implementations and numerical experiments in paper 1 are done by the Author. In paper 2 the Author has done substantial work in preparing the literature review and writing the proofs for the martingale case and presented numerical results and illustrations. In the last direction the Author has done the main steps of the proof of the probabilistic bound, verification of the assumptions, literature review and has taken part in the implementation of the algorithms and experiment design.

Approbation and Publications

First-Tier Publications

1. Denis Belomestny, Maxim Kaledin, and John Schoenmakers. Semitractability of optimal stopping problems via a weighted stochastic mesh algorithm. *Mathematical Finance*, 30(4):1591–1616, 2020

2. Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2144–2203. PMLR, 09–12 Jul 2020

Other Publications

1. Maxim Kaledin, Alexander Golubev, and Denis Belomestny. Variance reduction for policy-gradient methods via empirical variance minimization. *arXiv:2206.06827v2*, 2022

Reports at Conferences and Seminars

1. Kaledin M. *Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization*, summer school "Learning and Optimization in Artificial Intelligence Models", HSE, Saint-Petersburg, June 20-26 2022.

2. Kaledin M. *Theoretical Analysis and Variance Reduction in Reinforcement Learning Algorithms*, CMAP Doctoral Student Reports, CMAP Institut Polytechnique de Paris, Palaiseau, France, May 31 2021.

3. Kaledin M. *Variance Reduction for policy-gradient methods in Reinforcement Learning*, PhD Research seminar of Doctoral School of Computer Science, HSE, Moscow,

Russia, December 21 2020 .

4. Kaledin M. *Variance Reduction for policy-gradient methods in Reinforcement Learning*, summer school "Modern methods of Information Theory, Optimization and Control" , Sirius, Sochi, Russia, August 2-23 2020.

5. Kaledin M. *Convergence of Linear Two-Timescale Stochastic Approximation*, Winter School "Math of Machine Learning" , Sirius, Sochi, Russia, February 20-23 2019.

6. Kaledin M. *Approximate Dynamic Programming for American Options*, poster session, "Data Science Summer School" (DS3), l'École Polytechnique, Paris, June 24-28th 2019.

7. Kaledin M. *Approximate Dynamic Programming with Approximation of Transition Density*, Winter School "New Frontiers in High-Dimensional Probability and Statistics 2" , HSE, Moscow, February 22-23 2019.

0.1 Contents

0.1.1 Semitractability of Optimal Stopping Problem via Weighted Stochastic Mesh Algorithm

The results of this section are published in [9].

Introduction

Optimal stopping problem consists in constructing a decision rule saying when to take one particular decision ("stop" the process). Being a classic problem in mathematical finance, it is in the core of pricing various types of options, the most popular are American and European [35]. We consider two types of problems.

1. (Continuous-time optimal stopping) Assume set of stopping opportunities $[0, T]$ and let $(X_t)_{t\in[0,T]}$ be, as set in Problem 1, an Ito diffusion process set by (35) The problem is the same as above but with g_t being a payoff function for each $t \in [0, T]$ and $\mathcal{T}$ being the set of stopping times taking values in range $[0, T]$.

2. (Discrete-time optimal stopping) Assume a time-discretized version of the problem above with some finite set of stopping opportunities $\mathcal{L} = \{0, .., L\}$ for some $L \in \mathbb{Z}_{>0}$ and let $(Z_l)_{l\in\mathcal{L}}$ be a Markov chain in $\mathbb{R}^d$ obtained after the discretization. The problem is to find stopping time τ^* giving

$$\mathbb{E}\left[g_{\tau_*}(Z_{\tau_*}) \;\middle|\; Z_0\right] = \sup_{\tau\in\mathcal{T}} \mathbb{E}\left[g_\tau(Z_\tau) \;\middle|\; Z_0\right],$$

   where g_l are payoff functions $\mathbb{R}^d \to \mathbb{R}_{\geq 0}$ at times $l \in \mathcal{L}$ and $\mathcal{T}$ is set of stopping times taking values in $\mathcal{L}$. For simplicity and without loss of generality we assume that Markov chain $(Z_l)_{l\in\mathcal{L}}$ is time-homogeneous with one-step transition density denoted by $p(y|x)$ so that

$$\mathbb{P}\left(Z_{k+1} \in dy \mid Z_k = x\right) = p(y|x)dy$$

   for all $x, y \in \mathbb{R}^d$.

Despite existing convergence results, it turns out that comparing different algorithms for optimal stopping problem based solely on their convergence rates is not possible since these algorithms may be significantly different from a computational standpoint. The core approaches to complexity analysis in numerical algorithms can be found in [58] and the references therein. The main problem studied in this literature is the computation of integrals via deterministic and stochastic algorithms. Optimal stopping problems, in fact, present computations of several nested integrals since the dynamic programming principle is used. Hence, the existing results from standard complexity theory cannot be directly transferred to the complexity analysis of optimal stopping problem. In particular, for LS algorithm [89, Cor. 3.10] results in costs

$$\mathcal{C}_L(\varepsilon, d) \sim \kappa_1 \frac{L 5^{(\kappa_2+L)(2+3d/\alpha)}}{\varepsilon^{2+3d/\alpha}}$$

with κ_1, κ_2 being certain constants. If the problem is in continuous time, then by tuning time discretization we arrive at complexity of LS algorithm possibly growing even faster than $\exp(\varepsilon^{-1/\beta})$ for some $\beta > 0$. The similar bound holds for other simulation based

regression algorithms, including the one by Tsitsiklis and Van Roy [80]. In [29] the more general regression scheme is considered with similar type of results. The main problem with these complexity estimates is that the dimensionality of the process d enters the degree of ε resulting in so-called *curse of dimensionality* still appearing even in such Monte Carlo schemes. There exists, however, work of [36] where the novel Monte-Carlo-type scheme is developed with complexity independent of d but, unfortunately, it is exponential in ε^{-1}.

Tractability is an important notion in the analysis of numerical algorithms and one of the ways to define it is as follows. A d-dimensional numerical problem, for example, computation of an integral like $\int_{[0,1]^d} f(x)dx$, is called *tractable* [58], if there is an algorithm to solve it with complexity $\mathcal{C}(\varepsilon, d)$ satisfying

$$\lim_{d+\varepsilon^{-1}\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = 0. \tag{3}$$

In the case of optimal stopping problems, however, such a definition is not very meaningful: in all regression-type algorithms already in the case of discrete-time problem one has

$$\limsup_{d+\varepsilon\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = \infty$$

due to the exponential dependence of the complexity on d (based on the convergence rates known in the literature). Thus, even for a discrete-time optimal stopping problem regression-type algorithms are intractable with respect to this definition. For example, with the results of [78] it can be shown that the error of the estimation of the value function in this case has the form

$$5^L \left(\sqrt{\frac{m^d}{N}} + e^{-\theta m} \right), \quad \theta > 0.$$

However, this observation also applies to Weighted Stochastic Mesh(WSM) algorithm of Broadie and Glasserman [17], making almost all algorithms intractable. This motivates the development of more flexible complexity metric for the comparison of the algorithms for optimal stopping problems.

In turns out that not much is known about the convergence properties of WSM method except some preliminary results in discrete case [1]. The authors, however, do not give the dependence of the errors on the underlying dimension and the number of stopping times and their analysis is based on a rather restrictive assumption of compact state space. Similar type of algorithm we present here was also analyzed in the work of Rust [63] presenting a Monte Carlo scheme which has no exponential dependence on d but just $O(1/\varepsilon^4)$. The setting of discrete-time Markov Decision Process and the techniques used, however, make the transfer to optimal stopping non-trivial. Also the paper considers very restrictive assumptions of compact state space and Lipschitz continuity of transition densities with Lipschitz constant independent on the dimension d.

Complexity Metrics

It turns out that the criterion (1.3) puts too much importance on the dimension d on the one hand and on the other hand is too relaxed in dependence on ε. With such

definition the algorithm with complexity $d^2\exp\left(\varepsilon^{-1}/\ln\ln...\ln\varepsilon^{-1}\right)$ is tractable while one with complexity $2^d/\varepsilon$ is not despite that running an algorithm with the former complexity seems to be practically impossible even with $d = 1$. Therefore, we proposed another approach to tractability.

Definition 5. *For an algorithm with computational complexity $\mathcal{C}(\varepsilon, d)$ the number*

$$\Gamma_{\mathcal{C}} := \limsup_{d\to\infty} \limsup_{\varepsilon\to 0} \frac{\ln \mathcal{C}(\varepsilon, d)}{d \ln(1/\varepsilon)}.$$

is called semitractability index.

Definition 6. *The problem is called semitractable if there exists an algorithm solving it with $\Gamma_{\mathcal{C}} = 0$.*

Note that this definition nicely processes the dependencies of the complexities like $1/\varepsilon^{poly(d)}$ making possible the comparison of various Monte Carlo algorithms for solving optimal stopping and optimal control problems.

WSM Algorithm

Let us present a Weighted Stochastic Mesh (WSM) algorithm for a discrete-time optimal stopping problem. The algorithm is inspired by [17] but it differs in special choice of weights and truncation level. First, let us define the discrete Snell envelope process:

$$U_l = U_l(Z_l) := \sup_{\tau\in\mathcal{T}_{l,L}} \mathbb{E}\left[g_\tau(Z_\tau) \mid \mathcal{F}_l\right], \quad l = 0,..L,$$

where $\mathcal{T}_{l,L}$ is the set of stopping times taking values in the set $\{l,..,L\}$. Snell envelope satisfies dynamic programming principle, therefore, we can compute U_l using backward induction:

$$\begin{aligned} &U_L(Z_L) = g_L(Z_L),\\ &U_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[U_{l+1}(Z_{l+1}) \mid Z_l\right]\right\}, \quad l = 0,..,L-1. \end{aligned}$$

For technical purposes of the analysis we set truncation level $R > 0$ and define the truncated version of this backward induction:

$$\widetilde{U}_L(Z_L) = g_L(Z_L), \tag{4}$$

$$\widetilde{U}_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}) \mid Z_l\right]\right\}\cdot \mathbb{1}_{B_R}(Z_l), \quad l = 0,..,L-1, \tag{5}$$

where $\mathbb{1}_{B_R}$ is the indicator function of the 0-centered euclidean ball of radius R in $\mathbb{R}^d$. Thus, the values vanish when the process is out of B_R. We sample N independent trajectories $(Z_l^{(n)})_{l\in\mathbb{L}}$ with $Z_0^{(n)} = x_0, n = 1,..,N$ with the help of transition density $p(y|x)$. To estimate the conditional expectations, we use the following approximation:

$$\mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}) \mid Z_l = x\right] \approx \sum_{n=1}^{N} \widetilde{U}_{l+1}\left(Z_{l+1}^{(n)}\right) \frac{p\left(Z_{l+1}^{(n)} \mid x\right)}{\sum_{m=1}^{N} p\left(Z_{l+1}^{(n)} \mid Z_l^{(m)}\right)}. \tag{6}$$

To sum up, WSM algorithm is as follows:

1. Simulate N independent trajectories $(Z_l^{(1)})_{l\in\mathcal{L}}, .., (Z_l^{(N)})_{l\in\mathcal{L}}$;

2. Set $\overline{U}_L(Z_L^{(n)}) = g_L(Z_L^{(n)})$ for $n = 1, .., N$;

3. For $l = L-1, .., 1$ compute $\overline{U}_l(Z_l^{(n)})$ for all $n = 1, .., N$ using (1.6) and (1.8) for approximation of the conditional expectation;

4. Compute
$$\overline{U}_0(x_0) = \max\left\{g_0(x_0)\ ,\ \frac{1}{N}\sum_{n=1}^{N}\overline{U}_1^{(n)}\left(Z_1^{(n)}\right)\right\}.$$

One more thing to notice is that one step of backward induction with (1.6) and (1.8) takes N^2c_* with c_* being the price of multiplication. Thus, the total computational cost of the algorithms is c_*N^2L and given that $c_* \ll c_f^{(d)}$, the cost of one computation of transition density, it is bounded from above by $c_f^{(d)}N^2L$.

Main Results

Using the bounds from the literature we have computed $\Gamma_{\mathcal{C}}$ for two popular in practice methods (Longstaff-Schwarz[53] and Quantization Tree [7], see the table below) in discrete-time and continuous-time optimal stopping. For WSM algorithm we have two core results presented below.

Theorem 1. *(Proposition 2.5 in [9]) Suppose that the following conditions are satisfied:*

1. $$\max_{0\le l\le L} g_l(x) \le c_g(1+\|x\|_2), \quad x\in\mathbb{R}^d;$$

2. $$\mathbb{E}\left[\max_{l\le l'\le L}|Z_{l'}| \ \mid\ Z_l = x\right] \le c_Z(1+\|x\|_2), \quad x\in\mathbb{R}^d;$$

3. *There exist $\kappa, \alpha > 0$ such that for all $l = 1, .., L$ the l-step transition density satisfies*
$$0 < p_l(y|x) \le \frac{\kappa}{(2\pi\alpha L)^{d/2}} e^{-\frac{\|x-y\|_2^2}{2\alpha l}}.$$

Then the complexity of WSM algorithm is bounded from above by

$$\mathcal{C}(\varepsilon, d) = c_1\alpha^2 c_g^4\kappa^2 c_f^{(d)} c_2^d L^{d+7}\varepsilon^{-4} \times \ln^{d+2}\left[\frac{L\left(1+c_Z+c_Z\|x_0\|_2\right)e^{\frac{c_Z\sqrt{\alpha L}}{1+c_Z+c_Z\|x_0\|_2}}2^{3/4}\left(c_g\kappa\vee 1\right)}{\varepsilon}\right].$$

Corollary 2. *(Corollary 2.6 in [9]) Discrete-time optimal stopping under the assumptions of Theorem 1 is semitractable if the complexity of the computation of the transition density at one point $c_f^{(d)}$ is at most polynomial in d.*

One minor result we have obtained is that if the transition density itself cannot be computed but we have an approximation which is good enough, then the same result holds with slightly different constants. In particular, we get finite tractability index if approximating sequence p^n satisfies

$$\left|\frac{p^n(y|z) - p(y|z)}{p^n(y|z)}\right| \lesssim \frac{(1 + \|y - x_0\|_2^m + \|z - x_0\|_2^m)^n}{n!}, \quad y, z \in B_{R_n}$$

for some $m \in \mathbb{Z}_{>0}$ and appropriate sequence $R_n \to \infty$ as $n \to \infty$.

Considering continuous-time optimal stopping, we first build a discretization scheme based on Euler-Maruyama method with uniform time discretization having step h (for details see [9]). This essentially gives a discrete-time problem. In fact, the theorem is proven for more general approximation scheme and Euler-Maruyama scheme is just one example of the method which works.

Theorem 3. *(Proposition 3.4 in [9]) Assume the following conditions:*

1. $$\max_{0 \le t \le T} g_t(x) \le c_g(1 + \|x\|_2), \quad x \in \mathbb{R}^d;$$

2. $$\mathbb{E}\left[\max_{l \le l' \le L} \left|\overline{X}_{l'h}\right| \;\middle|\; \overline{X}_{lh} = x\right] \le c_{\overline{X}}(1 + \|x\|_2), \quad x \in \mathbb{R}^d;$$

3. *There exist $\overline{\kappa}, \overline{\alpha} > 0$ such that for all $l = 1, .., L$ the l-step transition density of $(\overline{X}_{lh})_{l \in \mathcal{L}}$ satisfies*
$$0 < \overline{p}_{lh}(y|x) \le \frac{\overline{\kappa}}{(2\pi\overline{\alpha}lh)^{d/2}} e^{-\frac{\|x-y\|_2^2}{2\overline{\alpha}lh}}.$$

Then the cost of computing the solution of obtained discrete-time optimal stopping problem is is bounded from above by

$$\mathcal{C}(\varepsilon, d) = c_1 \overline{\alpha}^2 c_g^4 \overline{\kappa}^2 c_f^{(d)} c_2^d \frac{T^{d+7}}{h^{d+5}} \varepsilon^{-4} \times \ln^{d+2}\left[\frac{(T/h)\left(1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2\right) e^{\frac{c_{\overline{X}}\sqrt{\overline{\alpha}T}}{1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2}} 2^{3/4} (c_g\overline{\kappa} \vee 1)}{\varepsilon}\right]$$

and the cost of computing the solution of continuous-time optimal stopping problem is bounded from above by

$$\mathcal{C}^\star(\varepsilon, d) = c_1 \overline{\alpha}^2 c_g^4 \overline{\kappa}^2 c_f^{(d)} c_2^d \frac{T^{d+7}}{\varepsilon^{2d+14}} \times \ln^{d+2}\left[\frac{T\left(1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2\right) e^{\frac{c_{\overline{X}}\sqrt{\alpha T}}{1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2}} 2^{3/4} (c_g\overline{\kappa} \vee 1)}{\varepsilon}\right].$$

Corollary 4. *In the setting of continuous optimal stopping problem, the WSM algorithm with time discretization satisfying the assumptions of Theorem 3 has semitractability index $\Gamma_{\mathcal{C}^\star} = 2$.*

The comparison table with semitractability indices we obtained is reported in our paper [9] and is placed below.

Setting \ Algorithm	LS	WSM	QTM
Discr. time	$3/\alpha$	0	2
Cont. time	∞	2	6

Table 1: Semitractability indices for Longstaff-Schwarz(LS), Weighted Stochastic Mesh(WSM) and Quantization Tree Method(QTM) computed in the paper.

Numerical Experiments

In the following experiments we illustrate the WSM algorithm in the case of continuous-time optimal stopping problems. A lower bound for the value function in WSM method is obtained using a suboptimal stopping rule computed on an independent set of trajectories (test set). This stopping rule can be constructed using any interpolation algorithm based on the observations from the training trajectories. The fastest and the simplest way giving good results is the nearest neighbor interpolation, in our experiments we have chosen the number of nearest neighbors to be 500.

American put option on a single asset

To illustrate the performance of the WSM algorithm in continuous time, we consider a problem of pricing American put option on a single asset driven by geometric Brownian motion

$$X_t = X_0 e^{\sigma W_t + (r-\sigma/2)t}$$

with r denoting the riskless rate of interest, assumed to be constant, and σ being the constant volatility. The payoff function is given by

$$g(x) = \max(K - x, 0).$$

The fair price of an option is defined as

$$U_0 = \sup_{\tau \in \mathcal{T}_{[0,T]}} \mathbb{E}\left[e^{-r\tau} g(X_\tau)\right]$$

for which there is no closed form solution but there exist numerical methods giving accurate approximations to U_0. We used parameters $r = 0.08, \sigma = 0.20, K = X_0 = 100, T = 3$. An accurate estimate of U_0 in this particular case is obtained and reported in [44] to be 6.9320. In Fig. 1.1 we show the lower bounds obtained by WSM, LS and VF (value function regression method of [80]) in dependence of the number of stopping opportunities L setting uniform time discretization on $[0, T]$ (the larger L the more dense is the grid). As can be seen, WSM lower bound is much more stable when L increases and LS and VF needs to use more complex regression basis to compensate for this effect.

American max-call option on five assets

The model with $d = 5$ assets is considered where each underlying asset has dividend yield δ. The dynamics is set by

$$dX_t^k = (r - \delta)X_t^k dt + \sigma X_t^k dW_t^k, \quad k = 1, .., d,$$

where W_t^k are independent one-dimensional Brownian motions. The parameters are set to be $r = 0.05, \delta = 0.1, \sigma = 0.2$. As before, the holder may exercise the option at any time $t \in [0, T]$ with $T = 3$ and receive the payoff

$$g(X_t) = max\left(\max\left(X_t^1, .., X_t^d\right) - K,\ 0\right).$$

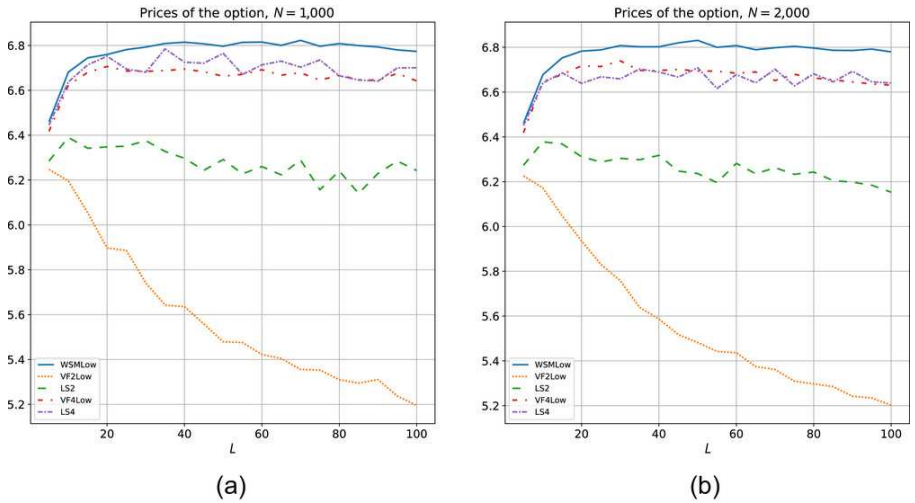


Figure 1: Lower bounds for the price of one-dimensional American put option approximated using different methods and uniform time discretization $t_k = kT/L, k = 0, .., L$ of exercise dates. The numbers of training paths are $N_{train} = 1000$(a) and $N_{train} = 2000$(b) and the number of test trajectories used for constructing the lower bounds $N_{test} = 20000$ and is the same in both cases. In LS and VF a polynomial basis of degrees 2 and 4 is used (mentioned in the legend).

We apply WSM and LS (with a basis of degree-2 polynomials) techniques to construct a lower bound. The results for different L are presented in Fig. 1.2. The option price must increase when the number of stopping opportunities increases, therefore LS-algorithm has clearly deteriorating estimate. WSM, on the other hand has increasing lower bound which shows that it performs considerably better than LS.

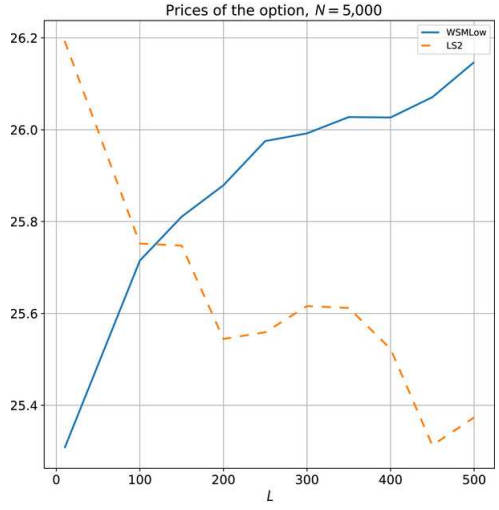


Figure 2: Lower bounds for the price of a five-dimensional American put option approximated using a uniform grid $t_k = kT/L, k = 0, .., L$ of exercise dates. The number of training paths is $N_{train} = 2000$ and the number of test trajectories is $N_{test} = 5000$.

0.1.2 Finite Time Analysis of Linear Two-Timescale Stochastic Approximation with Markovian Noise

The results of this section are published in [43].

Introduction

The TD-learning scheme based on classical (linear) SA is known to be inadequate for the off-policy learning paradigms in RL (data samples are drawn from a *behavior policy* different from the policy being evaluated [5, 79]). To circumvent this problem, [72, 73] have suggested gradient TD (GTD) method and the TD with gradient correction (TDC) method. These methods are represented as linear two-timescale SA scheme introduced by [14]:

$$\theta_{k+1} = \theta_k + \beta_k\{\widetilde{b}_1(X_{k+1}) - \widetilde{A}_{11}(X_{k+1})\theta_k - \widetilde{A}_{12}(X_{k+1})w_k\}, \tag{7}$$

$$w_{k+1} = w_k + \gamma_k\{\widetilde{b}_2(X_{k+1}) - \widetilde{A}_{21}(X_{k+1})\theta_k - \widetilde{A}_{22}(X_{k+1})w_k\}. \tag{8}$$

The above recursion involves two iterates, $\theta_k \in \mathbb{R}^{d_\theta}$, $w_k \in \mathbb{R}^{d_w}$, whose updates are coupled with each other. In the above, $\widetilde{b}_i(x)$, $\widetilde{A}_{ij}(x)$ are measurable vector/matrix valued functions on X and the random sequence $(X_k)_{k\geq 0}, X_k \in \mathrm{X}$ forms an ergodic Markov chain. The scalars $\gamma_k, \beta_k > 0$ are step sizes. The above SA scheme is said to have two timescales as the step sizes satisfy $\lim_{k\to\infty} \beta_k/\gamma_k < 1$ such that w_k is updated at a faster timescale. In fact, w_k is a 'tracking' term which seeks solution to a linear system characterized by θ_k.

Our goal is to characterize the finite-time expected error bound with improved convergence rate for the two-timescale SA (2.1),(2.2). The almost-sure convergence of two timescale SA has been established in [14, 75, 76, 15], among others and [46, 57] characterized the asymptotic convergence rates. However, finite-time risk bounds for two timescale SA have not been analyzed until recently. With martingale samples, [51] provided the first finite time analysis of GTD method, [26, 24] provided improved finite time error bounds. Unlike our analysis, they analyzed modified two timescale SA with projection and their bounds hold with high probability. With Markovian noise, [40] studied the finite time expected error bound with constant step sizes; [88] and [27] provided similar analysis for general step sizes. It is important to notice that with homogeneous martingale noise, the asymptotic rate of (2.1), (2.2) without a projection step, as shown in [46, Theorem 2.6], is in the order $\mathbb{E}\left[|\theta_k - \theta^\star|^2\right] = \mathcal{O}(\beta_k), \mathbb{E}\left[\left|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\right|^2\right] = \mathcal{O}(\gamma_k)$, where $\theta^\star$ is a stationary point of the SA scheme. However, the latter rate is not achieved in the finite-time error bounds analyzed by the above works except for [24]. It had been an open problem whether this error bound holds for the Markovian noise setting and for linear two time-scale SA scheme without projection.

Main Results

We investigate the linear two timescale SA given by the following equivalent form of (2.1), (2.2):

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{11}\theta_k - A_{12}w_k + V_{k+1}), \tag{9}$$

$$w_{k+1} = w_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}w_k + W_{k+1}), \tag{10}$$

where the mean fields are defined as $b_i := \lim_{k\to\infty} \mathbb{E}\left[\widetilde{b}_i(X_k)\right]$, $A_{ij} := \lim_{k\to\infty} \mathbb{E}\left[\widetilde{A}_{ij}(X_k)\right]$ (these limits exist as we recall that $(X_k)_{k\geq 0}$ is an ergodic Markov chain). The noise terms V_{k+1}, W_{k+1} are given by:

$$\begin{aligned} V_{k+1} &:= \widetilde{b}_1(X_{k+1}) - b_1 - (\widetilde{A}_{11}(X_{k+1}) - A_{11})\theta_k - (\widetilde{A}_{12}(X_{k+1}) - A_{12})w_k, \\ W_{k+1} &:= \widetilde{b}_2(X_{k+1}) - b_2 - (\widetilde{A}_{21}(X_{k+1}) - A_{21})\theta_k - (\widetilde{A}_{22}(X_{k+1}) - A_{22})w_k. \end{aligned} \tag{11}$$

The goal of the recursion (2.3), (2.4) is to find a stationary solution pair $(\theta^\star, w^\star)$ that solves the system of linear equations:

$$A_{11}\theta + A_{12}w = b_1, \quad A_{21}\theta + A_{22}w = b_2. \tag{12}$$

We are interested in the scenario when the solution pair $(\theta^\star, w^\star)$ is unique and is given by

$$\theta^\star = \Delta^{-1}(b_1 - A_{12}A_{22}^{-1}b_2), \quad w^\star = A_{22}^{-1}(b_2 - A_{21}\theta^\star). \tag{13}$$

where $\Delta := A_{11} - A_{12}A_{22}^{-1}A_{21}$.

To analyze the convergence of $(\theta_k, w_k)_{k\geq 0}$ in (2.3), (2.4) to $(\theta^\star, w^\star)$, we require several assumptions which are common for linear two time-scale SA, see [46].

A 1. Matrices $-A_{22}$ and $-\Delta = -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)$ are *Hurwitz*.

A 2. $(\gamma_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$ are nonincreasing sequences of positive numbers that satisfy the following.

1. There exist constant κ such that for all $k \in \mathbb{N}$, we have $\beta_k/\gamma_k \leq \kappa$.

2. For all $k \in \mathbb{N}$, it holds

$$\gamma_k/\gamma_{k+1} \leq 1+(a_{22}/8)\gamma_{k+1}, \;\; \beta_k/\beta_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}, \;\; \gamma_k/\gamma_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}. \tag{14}$$

Our conditions on step sizes are similar to [46, Assumption 2.3, 2.5]. These conditions encompass diminishing, piecewise constant and constant step sizes schedules which are common in the literature. For instance, a popular choice of diminishing step sizes satisfying A10 is

$$\beta_k = c^\beta/(k + k_0^\beta), \quad \gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3} \tag{15}$$

with some constants c^β, c^γ, k_0^γ, k_0^β, e.g., as suggested in [26, Remark 9]; or a constant step size of $\beta_k = \beta, \gamma_k = \gamma$; or a piecewise constant step size, e.g., [40].

We present new results on the convergence rate of (2.3), (2.4) depending on the types of noise with V_{k+1}, W_{k+1}. To discuss these cases, let us define the σ-field generated by the two timescale SA scheme and the initial error made by the SA scheme, respectively as:

$$\mathcal{F}_k := \sigma\big\{\theta_0, w_0, X_1, X_2, ..., X_k\big\}, \quad \mathrm{V}_0 := \mathbb{E}\left[\|\theta^0 - \theta^\star\|^2 + \|w^0 - w^\star\|^2\right]. \tag{16}$$

Our main results are presented for two sets of noise assumptions.

Martingale Noise We consider a simple setting where the random elements X_k are drawn i.i.d. from the distribution such that b_i, A_{ij} are the expected values of random variables $\widetilde{b}_i(X_k), \widetilde{A}_{ij}(X_k)$ which are assumed to have bounded second moment. This implies that the sequences $(V_{k+1})_{k\in\mathbb{N}}, (W_{k+1})_{k\in\mathbb{N}}$ are martingale difference sequences.

A 3. The noise terms are zero-mean conditioned on $\mathcal{F}_k$, i.e., $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] = \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$.

A 4. There exist constants m_W, m_V such that

$$\begin{aligned}\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\| &\le m_V(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|),\\ \|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| &\le m_W(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|)\,.\end{aligned} \tag{17}$$

Theorem 5. *Assume A9–12 and for all $k \in \mathbb{N}$, we have $\gamma_k \in [0, \gamma_\infty^{\mathsf{mtg}}]$, $\beta_k \in [0, \beta_\infty^{\mathsf{mtg}}]$ and $\kappa \in [0, \kappa_\infty]$, where $\gamma_\infty^{\mathsf{mtg}}, \beta_\infty^{\mathsf{mtg}}, \kappa_\infty$ are defined constants. Then*

$$\mathbb{E}\left[\|\theta_k - \theta^*\|^2\right] \le d_\theta\left\{\mathrm{C}_0^{\tilde{\theta},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\tilde{\theta},\mathsf{mtg}}\beta_k\right\} \tag{18}$$

$$\mathbb{E}\left[\left\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\right\|^2\right] \le d_w\Big\{\mathrm{C}_0^{\widehat{w},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\widehat{w},\mathsf{mtg}}\gamma_k\Big\} \tag{19}$$

The exact constants are provided in the paper.

Markovian Noise Consider the sequence $(X_k)_{k\ge0}$ to be samples from an exogenous Markov chain on X with the transition kernel $\mathrm{P} : \mathsf{X}\times\mathsf{X} \to \mathbb{R}_+$. For any measurable function f, we have

$$\mathbb{E}^{\mathcal{F}_k}[f(X_{k+1})] = \mathrm{P}\,f(X_k) = \int_{\mathsf{X}} f(x)\,\mathrm{P}(X_k, \mathrm{d}x)$$

B 1. The Markov kernel P has a unique invariant distribution $\mu : \mathsf{X} \to \mathbb{R}_+$. Moreover, it is irreducible and aperiodic.

Observe that

$$b_i = \int_{\mathsf{X}} \widetilde{b}_i(x)\,\mu(\mathrm{d}x), \quad A_{ij} = \int_{\mathsf{X}} \widetilde{A}_{ij}(x)\,\mu(\mathrm{d}x),\ i,j = 1,2.$$

We show that the linear two time-scale SA (2.1), (2.2) converges to a unique fixed point defined by the above mean field vectors/matrices, see (2.7). An important condition that enables our analysis is the existence of solutions to the following Poisson equations:

B 2. For any $i, j = 1, 2$, consider $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$, there exists vector/matrix valued measurable functions $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ which satisfy

$$\widetilde{b}_i(x) - b_i = \widehat{b}_i(x) - \mathrm{P}\,\widehat{b}_i(x), \quad \widetilde{A}_{ij}(x) - A_{ij} = \widehat{A}_{ij}(x) - \mathrm{P}\,\widehat{A}_{ij}(x) \tag{20}$$

for any $x \in \mathsf{X}$ and b_i, A_{ij} are the mean fields of $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$ with the stationary distribution μ.

The above assumption can be guaranteed under B5 together with some regularity conditions, see [28, Section 21.2]. Moreover,

B 3. Under B6, the vector/matrix valued functions $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ are uniformly bounded: for any $i, j = 1, 2$, $x \in \mathsf{X}$,

$$\|\widehat{b}_i(x)\| \leq \overline{\mathrm{b}}, \ \|\widehat{A}_{ij}(x)\| \leq \overline{\mathrm{A}}. \tag{21}$$

B 4. There exists constant ρ_0 such that for any $k \geq 1$, we have $\gamma_{k-1}^2 \leq \rho_0 \beta_k$.

To satisfy B7, we observe that the bounds $\overline{\mathrm{b}}, \overline{\mathrm{A}}$ depend on the mixing time of the chain $(X_k)_{k\geq 0}$ and a uniform bound on $\widetilde{b}_i(\cdot), \widetilde{A}_{ij}(\cdot)$. In the context of reinforcement learning, the latter can be satisfied when the feature vectors and reward are bounded. In fact, B7 implies A12. Meanwhile, B8 imposes further restriction on the step size. The latter can also be satisfied by (2.11). The challenges of analysis with Markovian noise lie in the biasedness of the noise term as $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] \neq 0$, $\mathbb{E}^{\mathcal{F}_k}[W_{k+1}] \neq 0$.

Theorem 6. *Assume A9–10, B5–8 hold and for all $k \in \mathbb{N}$, we have $\beta_k \in (0, \beta_\infty^{\mathsf{mark}}]$, $\gamma_k \in (0, \gamma_\infty^{\mathsf{mark}}]$, $\kappa \leq \kappa_\infty$, where $\beta_\infty^{\mathsf{mark}}$, $\gamma_\infty^{\mathsf{mark}}$, κ_∞ are defined constants. Then*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq d_\theta \left\{ \mathrm{C}_0^{\tilde{\theta},\mathsf{mark}} \prod_{\ell=0}^{k-1} \left(1 - \beta_\ell \frac{a_\Delta}{8}\right)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\tilde{\theta},\mathsf{mark}} \beta_k \right\}, \tag{22}$$

$$\mathbb{E}\left[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2\right] \leq d_w \left\{ \mathrm{C}_0^{\widehat{w},\mathsf{mark}} \prod_{\ell=0}^{k-1} \left(1 - \beta_\ell \frac{a_\Delta}{8}\right)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\widehat{w},\mathsf{mark}} \gamma_k \right\}. \tag{23}$$

The exact constants are given in the paper.

While Theorem 13 relaxes the martingale difference assumption A12 in Theorem 12, we remark that the results here do not generalize that in Theorem 12 due to the additional B7, B8. Particularly, with martingale noise, the convergence of linear two timescale SA only requires the noise to have bounded *second order moment*, yet the Markovian noise needs to be uniformly bounded.

The upper bounds in Theorem 12 and 13 consist of two terms – the first term is a 'transient' error with product such as $\prod_{i=0}^{k-1}(1 - \beta_i a_\Delta/8)$ decays to zero at the rate $o(1/k^c)$ for some $c > 1$ under an appropriate choice of step sizes such as (2.11); the second term is a 'steady-state' error. We observe that the 'steady-state' error of the iterates θ_k, w_k exhibit different behaviors. Taking the step size choices in (2.11) as an example, the steady-state error of the slow-update iterates θ_k is $\mathcal{O}(1/k)$ while the error of fast-update iterates w_k is $\mathcal{O}(1/k^{\frac{2}{3}})$. Furthermore, similar bounds hold for *both* martingale and Markovian noise.

Comparison to Related Works Our results improve the convergence rate analysis of linear two timescale SA in a number of recent works. In the martingale noise setting (Theorem 12), the closest work to ours is [24] which analyzed the linear two timescale SA with martingale samples and diminishing step sizes. The authors improved on [26] and obtained the same convergence rate (in high probability) as our Theorem 12, furthermore it is demonstrated that the obtained rates are tight. Their bounds also exhibit a sublinear dependence on the dimensions d_θ, d_w. However, their algorithm involves a sparsely executed projection step and the error bound holds only for a sufficiently large k. These restrictions are lifted in our analysis.

In the Markovian noise setting (Theorem 13), the closest works to ours are [27, 40, 88]. In particular, [40] analyzed the linear two timescale SA with constant step sizes and showed that the steady-state error for both θ_k, w_k is $\mathcal{O}(\gamma^2/\beta)$. [88] analyzed the TDC algorithm with a projection step and showed that the steady-state error for θ_k is $\mathcal{O}(1/k^{\frac{2}{3}})$ if

the step sizes in (2.11) is used. [27] analyzed the linear two timescale SA with diminishing step size and showed that the steady state error for both θ_k, w_k is $\mathcal{O}(1/k^{\frac{2}{3}})$. Interestingly, the above works do not obtain the fast rate in Theorem 13, i.e., $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = \mathcal{O}(1/k)$. One of the reasons for the sub-optimality in their rates is that their analysis are based on building a single Lyapunov function that controls both errors in θ_k and w_k. In contrast, our analysis relies on a set of coupled inequalities to obtain tight bounds for each of the iterates θ_k, w_k.

Our last result is the lower bound constructed to demonstrate the tightness of our analysis in Theorem 12, 13 writing the explicit expression for squared error $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$. We consider the following technical assumption:

A 5. There exist matrices $\Sigma^{11}, \Sigma^{12}, \Sigma^{22}$, and a constant $m_{VW}^{\mathsf{exp}} \geq 0$ such that for all $j \in \mathbb{N}$, it holds

$$\|\mathbb{E}\left[V_j V_j^\top\right] - \Sigma^{11}\| \vee \|\mathbb{E}\left[W_j W_j^\top\right] - \Sigma^{22}\| \vee \|\mathbb{E}\left[V_j W_j^\top\right] - \Sigma^{12}\| \leq m_{VW}^{\mathsf{exp}}(\|\mathbb{E}\left[\tilde{\theta}_k \tilde{\theta}_k^\top\right]\| + \|\mathbb{E}\left[\tilde{w}_k \tilde{w}_k^\top\right]\|).$$

Note that A13 implies A12 and therefore poses a stronger assumption. We have

Theorem 7. *Assume A9–11, A13 and for all $k \in \mathbb{N}$, we have $\gamma_k \in [0, \gamma_\infty^{\mathsf{mtg}}]$, $\beta_k \in [0, \beta_\infty^{\mathsf{exp}}]$ and $\kappa \in [0, \kappa_\infty^{\mathsf{exp}}]$, where $\gamma_\infty^{\mathsf{mtg}}, \beta_\infty^{\mathsf{exp}}, \kappa_\infty^{\mathsf{exp}}$ are constants defined in the paper. Then for any $k \geq k_0^{\mathsf{exp}} := \min\{\ell : \sum_{j=0}^{\ell-1} \beta_j \geq \log(2)/(2\|\Delta\|)\}$, the following expansion holds*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = I_k + J_k. \tag{24}$$

The leading term I_k is given by the following explicit formula

$$I_k := \textstyle\sum_{j=0}^k \beta_j^2 \operatorname{Tr}\left(\prod_{\ell=j+1}^k (\mathrm{I} - \beta_\ell \Delta)\, \Sigma \left\{\prod_{\ell=j+1}^k (\mathrm{I} - \beta_\ell \Delta)\right\}^\top\right),$$

where $\Sigma := \Sigma^{11} + A_{12}A_{22}^{-1}\Sigma^{22}A_{22}^{-\top}A_{12}^\top + \Sigma^{12}A_{22}^{-\top}A_{12}^\top + A_{12}A_{22}^{-1}\Sigma^{21}$. Meanwhile, the following two-sided inequality holds

$$\mathrm{C}_3^{\mathsf{exp}} \operatorname{Tr}(\Sigma) \leq \frac{I_k}{\beta_k} \leq \mathrm{C}_4^{\mathsf{exp}} \operatorname{Tr}(\Sigma), \tag{25}$$

and J_k is bounded by

$$|J_k| \leq \mathrm{C}_0^{\mathsf{exp}} \prod_{\ell=0}^{k-1} \left(1 - \frac{a_\Delta}{4}\beta_\ell\right) \mathrm{V}_0 + \mathrm{C}_1^{\mathsf{exp}} \beta_k \left(\gamma_k + \frac{\beta_k}{\gamma_k}\right), \tag{26}$$

where V_0 was defined in (2.12). All constants $\mathrm{C}_0^{\mathsf{exp}}, \mathrm{C}_1^{\mathsf{exp}}, \mathrm{C}_3^{\mathsf{exp}}, \mathrm{C}_4^{\mathsf{exp}}$ are given in the paper and they are independent of β_k, γ_k.

Observe that from (2.41), the dominant term for J_k is given by $\mathcal{O}(\beta_k\gamma_k + \frac{\beta_k^2}{\gamma_k})$. As such, using (2.40), we observe that

$$|J_k|/I_k = \mathcal{O}\left(\gamma_k + \beta_k/\gamma_k\right)$$

If $\lim_{k\to\infty} \beta_k/\gamma_k = 0$, we have $\lim_{k\to\infty} |J_k|/I_k = 0$. Combining (2.39), (2.40) shows that the expected error $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$ is lower bounded by $\Omega(\beta_k)$.

We note that the assumptions A9–11, A13 imposed by the theorem imply A9–A12 required by Theorem 12. Hence, together with (2.14) in Theorem 12, the above observations constitute a *matching* lower bound on the convergence rate of linear two timescale SA with martingale noise.

0.1.3 Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization

The results of this section are published in [42].

Introduction

In RL policy-gradient methods constitute the family of gradient algorithms directly modelling the policy and exploiting various formulas to approximate the gradient of expected reward with respect to the policy parameters [84, 74]. The straightforward way to tackle gradient estimation is Monte Carlo scheme resulting in the algorithm called REINFORCE [84]. Assume a Markov Decision Problem (MDP) $(\mathcal{S}, \mathcal{A}, R, \mathrm{P}, \Pi, \mu_0, \gamma)$ with a finite horizon T and given a class of policies $\Pi = \{\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A}) \mid \theta \in \Theta\}$ parametrized by $\theta \in \Theta \subset \mathbb{R}^D$ where $\mathcal{P}(\mathcal{A})$ is the set of probability distributions over the action set $\mathcal{A}$. We will omit the subscript in π_θ wherever possible for shorter notation, in all occurrences $\pi \in \Pi$. The optimization problem for MDP reads as

$$\text{maximize } J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t R(S_t, A_t)\right] \quad \text{w.r.t.} \quad \theta \in \Theta,$$

where we have assumed that the horizon T is fixed. Note that any sequence of states, actions, and rewards can be represented as an element X of the product space

$$(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T.$$

Let $\widetilde{\nabla} J|_{\theta'} : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T \to \mathbb{R}^D$ be an unbiased estimator of the gradient $\nabla_\theta J$ at point $\theta = \theta'$. With this notation the gradient descent algorithm for maxmization of $J(\theta)$ using the estimate $\widetilde{\nabla} J$ reads as follows:

$$\theta_{n+1} = \theta_n + \eta_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla} J|_{\theta_n}(X_n^{(k)}), \quad n = 1, 2, \ldots \tag{27}$$

with η_n being a positive sequence of step sizes. We will omit the subscript θ_n in the gradient estimate if it is clear from the context at which point the gradient is computed. REINFORCE [84] is one example of this estimator:

$$\widetilde{\nabla}^{\text{reinf}} J : \; X \mapsto \sum_{t=0}^{T-1} \gamma^t G_t(X) \nabla_\theta \log \pi(A_t|S_t)$$

with

$$G_t(X) := \sum_{t'=t}^{T-1} \gamma^{t'-t} R_t,$$

where $R_t = R(S_t, A_t)$ and

$$X = \left[(S_0, A_0, R_0), .., (S_{T-1}, A_{T-1}, R_{T-1})\right]^\top.$$

Unavoidably, there is the variance emerging from the estimation of the high-dimensional gradient [83]. This makes the problem of gradient estimation quite challenging. Variance reduction is necessarily required to construct modifications with gradient estimates of lower variance and lower computational cost than increasing the sample size.

The main developments in this direction include actor-critic by [45] and advantage actor-critic: A2C [74] and asynchronous version of it, A3C [55]. Generally, it can be considered as a modification of REINFORCE with additional use of control variate set by state-action-dependent baseline $b_\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (SA-baselines) or state-dependent baselines $b_\phi : \mathcal{S} \to \mathbb{R}$ (S-baselines) parametrized by ϕ. The estimator becomes

$$\widetilde{\nabla}_\theta^{b_\phi} J : X \mapsto \sum_{t=0}^{T-1} \gamma^t (G_t - b_\phi(S_t, A_t)) \nabla_\theta \log \pi(A_t|S_t),$$

the gradient scheme becomes two-timescale and baseline parameters are tuned so that the baseline models the state value function:

$$\theta_{n+1} = \theta_n + \alpha_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla}^{b_\phi} J(X_n^{(k)}), \tag{28}$$

$$\phi_{n+1} = \phi_n - \beta_n \nabla_\phi V_{K,n}^{A2C}(\phi)|_{\phi_n}, \tag{29}$$

where

$$V_{K,n}^{A2C}(\phi) := \frac{1}{K} \sum_{k=1}^{K} \sum_{t=0}^{T-1} (G_t(X_n^{(k)}) - b_\phi(S_t^{(k)}))^2 \tag{30}$$

is A2C goal reflecting our desire to approximate the corresponding value function from its noisy estimates $(G_t(X_n^{(k)}))$ via least squares. The motivation behind it is that if one chooses the value function as baseline, the variance will be minimized. This strategy works well in practical problems [55].

Recently a new interest in such methods has emerged due to the introduction of deep reinforcement learning [56], a very comprehensive review is done in [33]. During several decades a large number of new variance reduction methods were proposed, including subsampling methods like SVRPG [60, 86] and various control variate approaches of [64], [39], [52], [81], [85]. There are also approaches of a bit different nature: trajectory-wise control variates [19] using the control variate based on future rewards and variance reduction in input-driven environments [54]. Apart from that, in ergodic case there were both theoretic [38] and also some practical advancements [21]. The importance of the criteria for variance reduction is well-known in Monte-Carlo and MCMC [65] and recently was also addressed in RL by [30], where the Actor with Variance Estimated Critic (AVEC) was proposed.

Going to theory, it remains unclear how the procedure used in A2C is related to the variance of the gradient estimator. Moreover, the empirical studies of the variance of the gradient estimator are still very rare and available mostly for artificial problems. In the community there is still an ongoing discussion about the actual role of the variance of the gradient in the performance of the algorithms [81]. In our study we try to answer some of these questions and suggest a more direct approach inspired by the Empirical Variance(EV) Minimization recently studied by [8]. We show that the proposed EV-algorithm is not only theoretically justifiable but can also perform better than the classic A2C algorithm. It should be noted that the idea of using some kind of empirical variance functional is not new: some hints appeared, for instance, in [52]. Despite that, the implementation and theoretical studies of this approach are still missing in the literature.

Main Theoretical Results

The main object of our study is the use of empirical variance instead of A2C goal. Starting from this we could formulate two optimization goals for baseline tuning:

$$V_{n,K}^{EVv}(\theta,\phi) := \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2 - \frac{1}{K^2}\left\|\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2, \tag{31}$$

$$V_{K}^{EVm}(\theta,\phi) := \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2; \tag{32}$$

both can be shown to be an unbiased estimate of the true variance of the gradient estimator and true variance is defined for a random vector Y as

$$V(Y) := \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right].$$

The corresponding gradient algorithms can be described as

$$\theta_{n+1} = \theta_n + \alpha_n\frac{1}{K}\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi}J(X_n^{(k)}), \tag{33}$$

$$\phi_{n+1} = \phi_n - \beta_n\nabla_\phi V_K^{EV}(\phi,\theta)|_{\phi_n,\theta_n}. \tag{34}$$

We got two methods. The first one uses the full variance V_K^{EVv} and is called EVv, the second one is titled EVm and exploits V_K^{EVm}, the same variance functional but without the second term. The important fact to note is that EVv routine would work only if $K \geq 2$, otherwise we try to estimate the variance with one observation. We can note several quick facts about these methods. Firstly, it turns out that under some technical assumptions A2C goal is an upper bound (up to a constant) of EV goals (Prop.5 in [42]). Secondly, we show that if the scheme converges to a local optimum, then EVm and EVv methods are asymptotically equivalent since the second term of the variance is the squared norm of the true gradient which converges to 0.

The main theoretical result is high-probability bound for excess risk on step n of the algorithm. For this we first simplify the notation for more clarity. Let us further notate the gradient estimator as $h : \mathbb{R}^d \to \mathbb{R}^D$, fix some set of such estimators $\mathcal{H}$ and define $\mathcal{E} = \mathbb{E}[h(X)] = \nabla_\theta J$ since the estimate is assumed to be unbiased. In order to reduce the variance in the gradient estimator we would like to pick on each epoch n the best possible estimator

$$h^* = \arg\min_{h\in\mathcal{H}} V(h)$$

where variance functional V is defined for any $h \in \mathcal{H}$ via

$$V(h) := \mathbb{E}\left[\|h(X) - \mathcal{E}\|^2\right]$$

where X is random vector of concatenated states, actions and rewards described before. To solve the above optimization problem, we use empirical analogue of the variance and define

$$\hat{h} := \arg\min_{h\in\mathcal{H}} V_K(h)$$

with the empirical variance functional of the form:

$$V_K(h) := \frac{1}{K-1}\sum_{k=1}^{K}\|h(X^{(k)}) - P_K h\|^2$$

with P_K being the empirical measure, so with the given sample we could notate sample mean as

$$P_K h := \frac{1}{K}\sum_{k=1}^{K} h(X^{(k)}).$$

Let us pose several key assumptions.

A 6. Class $\mathcal{H}$ consists of bounded functions:

$$\sup_{x\in\mathcal{X}}\|h(x)\| \le b, \quad \forall h \in \mathcal{H}.$$

A 7. The solution h_* is unique and $\mathcal{H}$ is star-shaped around h_*:

$$\alpha h + (1-\alpha)h_* \in \mathcal{H}, \quad \forall h \in \mathcal{H},\ \alpha \in [0,1].$$

A 8. The class $\mathcal{H}$ has covering of polynomial size: there are $\alpha \ge 2$ and $c > 0$ such that for all $u \in (0, b]$,

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(P_K)}, u) \le \left(\frac{c}{u}\right)^{\alpha} \ a.s.$$

where

$$\|h\|_{L^2(P_K)} = \sqrt{P_K\|h\|_2^2}$$

The following result holds.

Theorem 8. *Under Assumptions 14-16 it holds with probability at least* $1 - 4e^{-t}$,

$$V(h_K) - V(h_*) \le \max_{j=1,\dots,4}\beta_j(t)$$

with

$$\beta_1 \le C_1\frac{\log K}{K},\ \beta_2 \le C_2\frac{\log K}{K},$$
$$\beta_3(t) = \frac{C_3(t+1)}{3K}, \beta_4(t) = \frac{C_4 t}{K},$$

where C_1, C_2, C_3, C_4 *are constants not depending on the dimension* D *or the sample size* K *and are defined in the paper.*

This allows to conclude that as sample size K grows, the variance reduces to that of h_*. From practical perspective, Theorem 24 firstly gives some reliability guarantee. Secondly, it also shows that if we have K large enough, we can reduce the variance even more.

Numerical Experiments

We empirically investigate the behavior of EV-algorithms on several benchmark problems:

- Gym Minigrid [20] (`Unlock-v0, GoToDoor-5x5-v0`);
- Gym Classic Control [18] (`CartPole-v1, LunarLander-v2, Acrobot-v1`).

For each of these we provide charts with mean rewards illustrating the training process, the study of gradient variance and reward variance and time complexity discussions. Here because of small amount of space we present the most important results but the reader is welcome in the Supplementary materials where more experiments and investigations are presented together with all the implementation details. The code and config-files can be found on GitHub page [37].

Overview. Below we show the discussions about several key indicators of the algorithms.

1. **Mean rewards.** They are computed at each epoch based on the rewards obtained during the training in 40 runs and characterize how good is the algorithm in interaction with the environment.
2. **Standard deviation of the rewards.** These are computed in the same way but standard deviation is computed instead of mean. This values show how stable the training goes: high values indicate that there are frequent drops or increases in rewards.
3. **Gradient variance.** It is measured every 200 epochs using (3.9) with separate set of 50 sampled trajectories with relevant policy. This is the key indicator in the discussion of variance reduction. Surprisingly, as far as we know, we are the first in the RL community presenting such results for classic benchmarks. The resulting curves are averaged over 40 runs.
4. **Variance Reduction Ratio.** Together with Gradient Variance itself we also measure reduction ratio computed as sample variance of the estimator with baseline divided by the sample variance without baseline (assuming $b_\phi = 0$) in the computations of Gradient Variance. The reduction ratio is the main value of interest in variance reduction research in Monte Carlo and MCMC.

Algorithm Performance. While observing mean rewards during the training we may notice immediately that EV-algorithms are at least as good as A2C. In `CartPole` environment (Fig. 3.1) we conducted several experiments and present here two policy configurations: one with simpler neural network (config5, see Fig. 3.1(a,b,c)) and one with more complex network (config8, see Fig. 3.1(d,e,f)). In the first case both A2C and EV have very similar performance but in the second case the agent learns considerably faster with EV-based variance reduction and we get approximately 50% improvement over A2C agent and 75% over Reinforce agent in the end and even more during the training. The phenomenon of better performance of EV in CartPole with more complex policies

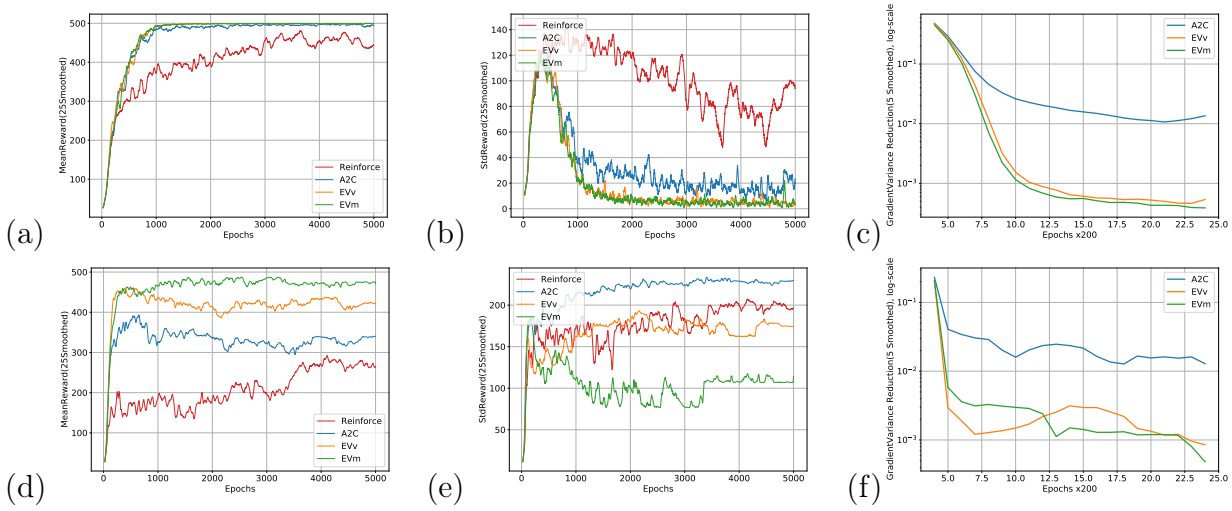


Figure 3: The charts representing the results for `CartPole` environment: (a,b,c) represent mean rewards, standard deviation of the rewards and gradient variance reduction ratio for config5 and (d,e,f) show the same information about config8.

is observed often, more detailed discussion is placed in Supplementary. As to `Acrobot` (see Fig. 3.2(a)), we see EV-algorithms giving better speed-up in the training. In the beginning EVm allows to learn faster but in the end the performance is the same as A2C. One of the reasons of such behavior can be the fact that learning rate becomes small and the agent already reaches the ceiling. `Unlock` (Fig. 3.3(a)) is the example of the environments where all algorithms work similarly: in terms of rewards we cannot see significant improvement even over Reinforce.

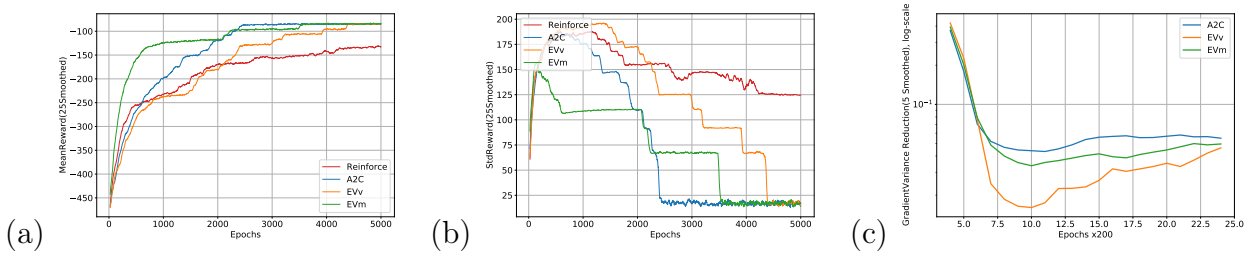


Figure 4: The charts representing the results for `Acrobot` environment: (a) depicts mean rewards, (b) shows the standard deviations of the rewards and (c) displays the gradient variance reduction ratios.

Stability of Training. When we study the charts for standard deviation of the rewards (Fig. 3.1(b,e),3.2(b),3.3(b)), we can see that EV-methods are better in terms of stability of the training, the algorithm more rarely has drops than that of A2C. This is greatly illustrated by `CartPole` in Fig. 3.1(b,e) where the standard deviation is about 2 times less than in case of A2C. This holds for both configurations. Fig. 3.2 illustrating the experiments with `Acrobot` show that until the ceiling is reached EV methods still can have lower variance. In `Unlock` presented in Fig. 3.3(b) we have not observed a significant difference in reward variance.

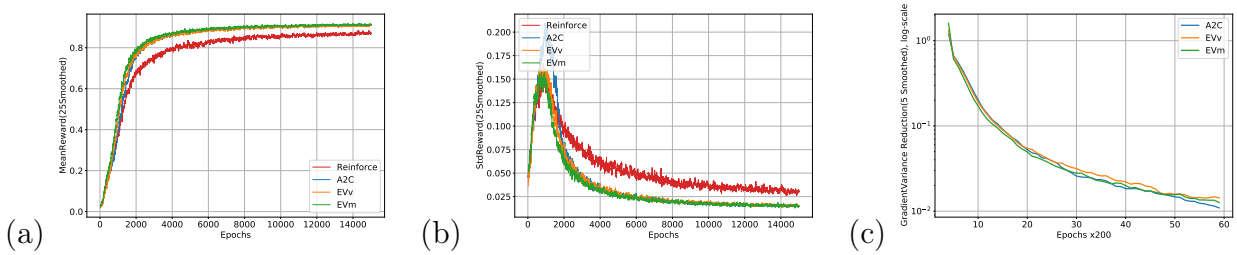


Figure 5: The charts representing the results for `Unlock` environment: (a) depicts mean rewards, (b) shows the standard deviations of the rewards and (c) displays the gradient variance reduction ratios.

Gradient Variance and its Influence. The first thing we can notice reviewing the gradient variance is that A2C and EV reduce the variance similarly in `Unlock`. `CartPole` (see Fig. 3.1(c,f)), however, gives an example of the case where EV works completely differently to A2C, it reduces the variance almost 100-1000 times in both policy configurations. Similar picture we can observe in all `CartPole` experiments. We can see that in `Unlock` shown in Fig. 3.3 the variance can also be reduced approximately 10-100 times, however, we see very little gain in rewards. It shows that in some environments training does not respond to the variance reduction; as a reason, it can be just not enough to give the improvement. The last thing we would like to note is that reward variance measured in previous sub-section is not an indicator of variance reduction since we have shown gradient variance reduction in all cases. Reward variance is decreased in relation to Reinforce, however, only in `CartPole` environement. Therefore, it cannot be used as a key metric for studying variance reduction in RL. The connection between reward variance and gradient variance seems to be an unanswered question in the literature.

Conclusion

Considering the first goal, for discrete-time optimal stopping problems we have established semitractability for the proposed WSM algorithm under weak assumption of Markov chain with transition kernel possessing a density. In the most common case of infinitely smooth continuation functions many regression based algorithms, including LS, are also semitractable for discrete-time optimal stopping problems. However, as we have shown, when going to continuous optimal stopping problem, regression method gives infinite semitractability index while WSM's index remains bounded, the experiments have clearly shown the practical consequences of it.

In the second direction we have achieved an improved finite time convergence analysis of the linear two timescale SA on both martingale and Markovian noises with relaxed conditions. Our analysis show that a tight analysis is possible through deriving and solving a sequence of recursive error bounds.

As to the third goal, we suggested to use empirical variance which in turn resulted in EV-methods. The motivation of EV-algorithms is more about actual variance reduction than in case of A2C and their performance is at least as good as A2C in terms of variance reduction and rewards. For them we also have suggested the first in the literature probabilistic bound for the variance of the gradient estimate under some mild assumptions.

EV-algorithms can be more stable in training which can allow to make sudden drops during the training less frequent. We also have for the first time presented the study of actual gradient variance reduction in classic benchmark problems. Our results have shown that variance reduction can help in the training but sometimes the environment's specific features do not allow to achieve gain in rewards. Therefore, variance reduction technique needs to be used during the training but the exact circumstances in which it helps are yet to be discovered.

Introduction

Stochastic optimal control problems are very often encountered in various practical areas: from finance [35, 77] to engineering [12]. Recently they have got a new attention and new challenges in the light of developing Reinforcement Learning (RL), in some sense presenting itself as the intersection of optimal control, statistics and machine learning [70].

Such class of problems can be defined as follows. Let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$ be a filtered probability space with filtration $(\mathcal{F}_t)_{t\geq 0}$. Assume some set $\mathcal{U}$ of progressively measurable stochastic processes $U : \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}^n$ called *controls* and set of *controlled processes*

$$\mathcal{X} = \left\{ X_t^U : U \in \mathcal{U} \right\}$$

where for every control U each $(X_t^U)_{t\geq 0}$ is an $\mathbb{R}^d$-valued $(\mathcal{F}_t)_{t\geq 0}$-adapted stochastic process. We also set functional $J : \mathcal{X} \to \mathbb{R}$ and call it *gain functional*.

Definition 1. *The problem of searching* $U_* \in \operatorname{Arg\,max}_{U\in\mathcal{U}} J(X^U)$ *is called stochastic optimal control problem.*

Also in practice (especially in reinforcement learning, see [70]) as a technical module of some algorithms it is needed to evaluate the given decision rule and so one gets an evaluation problem.

Definition 2. *The problem of evaluating* $J(X^U)$ *given a control* U *in some form is called control evaluation problem.*

Of course, with such abstract formulation we cannot claim anything about the existence of the solutions or their qualities. The question becomes much more clear when we consider more specific formulations. In the thesis the two more specific problems are considered: optimal stopping for a stochastic differential equation(SDE) and Markov Decision Problem (MDP).

Problem 1. (Optimal stopping problem for an SDE, [77, 35]) Assume $T > 0$ and let process X_t be set with an Ito SDE for $t \in [0, T)$

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{35}$$

with initial condition $X_0^U = x_0 \in \mathbb{R}^d$, where functions

$$b : [0, T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^d, \quad \sigma : [0, T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^{d\times n}$$

are two continuous functions satisfying Lipschitz condition in the second argument and linear growth condition with constant K:

$$\|b(t, x, u)\|_2 + \|\sigma(t, x, u)\|_2 \leq K(1 + \|x\|_2 + \|u\|_2)$$

with $\|\cdot\|_2$ denoting the appropriate Euclidean 2-norm. With such assumption we may ensure that the unique strong solution exists. Let $g_t : \mathbb{R} \to \mathbb{R}$ for every $t \in [0, T]$ be some function called *payoff*. Consider an agent observing the process, at time $t' \in [0, T]$ he knows the values of X_t for all $t \leq t'$. His goal is to choose the time τ when to take one particular decision (stop the process, as it is often called) which gives him payoff $g_\tau(X_\tau)$. Formally, we are interested in choosing a stopping time τ taking values in $[0, T]$ from the set of admissible stopping times $\mathcal{T}$ maximizing the expected discounted reward of the agent:

$$\tau_* = \arg\max_{\tau \in \mathcal{T}} \mathbb{E}\left[g_\tau(X_\tau)\right].$$

The most adopted by practitioners methods are invented with the ideas of Longstaff-Schwarz(LS)[53] and Tsitsiklis-Van Roy [80] algorithms in mind. They exploit dynamic programming principle and approximate conditional expectations using least-squares regression on a given basis of functions on each backward induction step. Longstaff and Schwarz demonstrated the efficiency of their approach through a number of numerical examples and in [22] and [89] general convergence properties of the method were established.

Problem 2. (Markov Decision Process, MDP, [70]) Assume some sets $\mathcal{S}, \mathcal{A}$ called *state* and *action* spaces (they have to be measurable spaces) and define discrete-time time-homogenuous Markov chain S_t as follows. Let there be Π, the set of stochastic decision rules (also called *policies*) $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, i.e. each policy takes the state $s \in \mathcal{S}$ and returns probability distribution over the action space denoted as $\pi(\cdot|s)$. Let us set *transition kernel* $P(\cdot|s, a)$ as a probability distribution over the state space given the current state and action. Set $S_0 = s_0$ almost surely and then iteratively update S_t to S_{t+1} using the following scheme:

$$A_t \sim \pi(\cdot|S_t),$$
$$S_{t+1} \sim P(\cdot|S_t, A_t).$$

Consider a deterministic uniformly bounded reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The natural illustration of MDP is that we have an agent in the environment with state descriptions from $\mathcal{S}$; the agent at each time t must make a decision A_t using his policy, after that he receives a reward $R(S_t, A_t)$ and the environment changes its state as shown above. The optimal control problem is to maximize with respect to policy the expected sum of discounted rewards

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(S_t, A_t)\right],$$

where $\gamma \in (0, 1)$ plays the role of the discounting factor and horizon T can be finite (finite-horizon problem) or infinite (infinite-horizon problem), or even random (episodic problem). MDP is a fundamental model in Reinforcement Learning(RL) being currently a fast-developing area with promising and existing applications in numerous innovative areas of the society: starting from AI for games [82, 11, 66] and going to energy management systems [49, 32], manufacturing and robotics [2] to name a few. Naturally, RL gives the practitioners new sets of control tools for any kind of automatization [33].

Policy evaluation is a vital part of the model-free algorithms based on policy iteration and it is normally based on Stochastic Approximation(SA) schemes, invented in

[62]. SA itself currently became a well-studied technique [10, 47, 15], however RL gives new challenges and new assumptions. Among others, linear SA schemes are popular in reinforcement learning (RL) as they lead to policy evaluation methods with linear function approximation, of particular importance is temporal difference (TD) learning [69] for which finite time analysis has been reported in [68, 48, 13, 25].

Aim of the Work

The aim of our research is to investigate the problems above in several ways.

1. Regarding the optimal stopping problem discussed in Section 1, we are aiming at presenting the complexity analysis of Weighted Stochastic Mesh(WSM) algorithm similar to the method of [17] for discrete- and continuous-time optimal stopping problem and compare it to other popular methods via new complexity metric since with respect to classic complexity metric all algorithms for optimal stopping are intractable and there is no way to compare them taking the complexity into account.

2. In Section 2 we aimed at obtaining finite-time convergence analysis for two-timescale linear Stochastic Approximation(SA) scheme under Markov noise assumptions. Such setting is exactly the setting of classic policy evaluation algorithms for MDP: temporal difference learning (TD(0) of [69]) and gradient temporal difference algorithms (GTD[72],GTD2 and TDC [73]). The problem with existing analysis is that it does not consider the Markov nature of the data (which is a natural thing since practitioners work in MDP setting) or the assumptions are too restrictive.

3. Finally, in Section 3 we set up to propose a new method for variance reduction based on empirical variance minimization of [8] in policy-gradient algorithms. The goal is, firstly, to obtain an algorithm able to give the improvement over the classic optimization goal for control variates in Advantage Actor-Critic(A2C) schemes [74] and, secondly, give some theoretical guarantees regarding the actual variance reduction.

Key Results

1. To address the first aim, we present for the first time the complexity analysis of WSM algorithm based on [17] and consider also the case when the transition density $p(x|y)$ is not known but can be approximated. We propose a new metric for comparison of the algorithms for optimal stopping problems called *semitractability index* and compare with it several algorithms popular in the community of practitioners: LS-algorithm [53] and QTM [7].

2. We provide improved convergence rates for the linear two-timescale SA in both martingale and Markovian noise settings. Our analysis allow for general step sizes schedules, including constant, piecewise-constant, and diminishing step sizes explored in the prior works [40, 24, 88, 27]. Unlike the prior works [51, 24, 88], our convergence results are obtained *without* requiring a projection step throughout the SA iterations. Finally, with an additional assumption on the step size, we compute an exact asymptotic expansion of the expected squared error to show the tightness of our upper bounds.

3. We provide two new policy-gradient methods (EV-methods) based on EV-criterion and show that they perform well in several practical problems in comparison to A2C-criterion. Also theoretical variance bounds for EV-methods are provided using the ideas of [8], this the first result concerning the variance bounds with high probability with the help of the tools of statistical learning in the setting of RL. Measurements of the variance of the gradient estimates present several somewhat surprising observations. Firstly, EV-methods are able to solve variance reduction problem considerably better than A2C. Secondly, we see some confirmations of the hypothesis of [81]: variance reduction has its effect but some environments are not so responsive to this. We present the first experimental investigation of EV-criterion of policy-gradient methods in classic benchmark problems and the first implementation of it in the framework of PyTorch.

Author contribution. Some part of the analysis for discrete-time case, transfer from discrete to continuous case, implementations and numerical experiments in paper 1 are done by the Author. In paper 2 the Author has done substantial work in preparing the literature review and writing the proofs for the martingale case and presented numerical results and illustrations. In the last direction the Author has done the main steps of the proof of the probabilistic bound, verification of the assumptions, literature review and has taken part in the implementation of the algorithms and experiment design.

Aprobation and Publications

First-Tier Publications

1. Denis Belomestny, Maxim Kaledin, and John Schoenmakers. Semitractability of optimal stopping problems via a weighted stochastic mesh algorithm. *Mathematical Finance*, 30(4):1591–1616, 2020

2. Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2144–2203. PMLR, 09–12 Jul 2020

Other Publications

1. Maxim Kaledin, Alexander Golubev, and Denis Belomestny. Variance reduction for policy-gradient methods via empirical variance minimization. *arXiv:2206.06827v2*, 2022

Reports at Conferences and Seminars

1. Kaledin M. *Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization*, summer school "Learning and Optimization in Artificial Intelligence Models", HSE, Saint-Petersburg, June 20-26 2022.

2. Kaledin M. *Theoretical Analysis and Variance Reduction in Reinforcement Learning Algorithms*, CMAP Doctoral Student Reports, CMAP Institut Polytechnique de Paris, Palaiseau, France, May 31 2021.

3. Kaledin M. *Variance Reduction for policy-gradient methods in Reinforcement Learning*, PhD Research seminar of Doctoral School of Computer Science, HSE, Moscow, Russia, December 21 2020 .

4. Kaledin M. *Variance Reduction for policy-gradient methods in Reinforcement Learning*, summer school "Modern methods of Information Theory, Optimization and Control" , Sirius, Sochi, Russia, August 2-23 2020.

5. Kaledin M. *Convergence of Linear Two-Timescale Stochastic Approximation*, Winter School "Math of Machine Learning" , Sirius, Sochi, Russia, February 20-23 2019.

6. Kaledin M. *Approximate Dynamic Programming for American Options*, poster session, "Data Science Summer School" (DS3), l'École Polytechnique, Paris, June 24-28th 2019.

7. Kaledin M. *Approximate Dynamic Programming with Approximation of Transition Density*, Winter School "New Frontiers in High-Dimensional Probability and Statistics 2" , HSE, Moscow, February 22-23 2019.

Chapter 1

Semitractability of Optimal Stopping Problem via Weighted Stochastic Mesh Algorithm

The results of this chapter are published in [9].

1.1 Introduction

Optimal stopping problem consists in constructing a decision rule saying when to take one particular decision ("stop" the process). Being a classic problem in mathematical finance, it is in the core of pricing various types of options, the most popular are American and European [35]. We consider two types of problems.

1. (Continuous-time optimal stopping) Assume set of stopping opportunities $[0, T]$ and let $(X_t)_{t\in[0,T]}$ be, as set in Problem 1, an Ito diffusion process set by (35) The problem is the same as above but with g_t being a payoff function for each $t \in [0, T]$ and $\mathcal{T}$ being the set of stopping times taking values in range $[0, T]$.

2. (Discrete-time optimal stopping) Assume a time-discretized version of the problem above with some finite set of stopping opportunities $\mathcal{L} = \{0, .., L\}$ for some $L \in \mathbb{Z}_{>0}$ and let $(Z_l)_{l\in\mathcal{L}}$ be a Markov chain in $\mathbb{R}^d$ obtained after the discretization. The problem is to find stopping time τ^* giving

$$\mathbb{E}\left[g_{\tau_*}(Z_{\tau_*}) \;\middle|\; Z_0\right] = \sup_{\tau\in\mathcal{T}} \mathbb{E}\left[g_\tau(Z_\tau) \;\middle|\; Z_0\right],$$

   where g_l are payoff functions $\mathbb{R}^d \to \mathbb{R}_{\geq 0}$ at times $l \in \mathcal{L}$ and $\mathcal{T}$ is set of stopping times taking values in $\mathcal{L}$. For simplicity and without loss of generality we assume that Markov chain $(Z_l)_{l\in\mathcal{L}}$ is time-homogeneous with one-step transition density denoted by $p(y|x)$ so that

$$\mathbb{P}\left(Z_{k+1} \in dy \mid Z_k = x\right) = p(y|x)dy$$

   for all $x, y \in \mathbb{R}^d$.

Despite existing convergence results, it turns out that comparing different algorithms for optimal stopping problem based solely on their convergence rates is not possible since these algorithms may be significantly different from a computational standpoint. The

core approaches to complexity analysis in numerical algorithms can be found in [58] and the references therein. The main problem studied in this literature is the computation of integrals via deterministic and stochastic algorithms. Optimal stopping problems, in fact, present computations of several nested integrals since the dynamic programming principle is used. Hence, the existing results from standard complexity theory cannot be directly transferred to the complexity analysis of optimal stopping problem. In particular, for LS algorithm [89, Cor. 3.10] implies that for a fixed number L of stopping opportunities and a popular choice of polynomial basis functions of degree less or equal to m, the error of estimating the corresponding value function at one point is bounded by

$$\kappa\, 5^L \left(\sqrt{\frac{m^d}{N}} + \frac{1}{m^\alpha} \right), \tag{1.1}$$

therefore,

Proposition 1. *For L stopping opportunities and underlying dimension d, the computational work for achieving an accuracy ε by the LS algorithm is bounded by*

$$\mathcal{C}_L\left(\varepsilon, d\right) = \kappa_1 \frac{L\, 5^{(\kappa_2 + L)(2+3d/\alpha)}}{\varepsilon^{2+3d/\alpha}} \tag{1.2}$$

with κ_1, κ_2 being certain constants.

If the problem is in continuous time, then by tuning time discretization we arrive at complexity of LS algorithm possibly growing even faster than $\exp(\varepsilon^{-1/\beta})$ for some $\beta > 0$. The similar bound holds for other simulation based regression algorithms, including the one by Tsitsiklis and Van Roy [80]. In [29] the more general regression scheme is considered with similar type of results. It is important also to mention [36] where the novel Monte-Carlo-type scheme is developed with complexity independent of d but, unfortunately, exponential in ε^{-1}.

Tractability is an important notion in the analysis of numerical algorithms and one of the ways to define it is as follows. A d-dimensional numerical problem, for example, computation of an integral like $\int_{[0,1]^d} f(x)dx$, is called *tractable* [58], if there is an algorithm to solve it with complexity $\mathcal{C}(\varepsilon, d)$ satisfying

$$\lim_{d+\varepsilon^{-1}\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = 0. \tag{1.3}$$

In the case of optimal stopping problems, however, such a definition is not very meaningful: in all regression-type algorithms already in the case of discrete-time problem one has

$$\limsup_{d+\varepsilon\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = \infty$$

(based on the convergence rates known in the literature). Thus, even for a discrete-time optimal stopping problem regression-type algorithms are intractable with respect to this definition. For example, with the results of [78] it can be shown that the error of the estimation of the value function in this case has the form

$$5^L \left(\sqrt{\frac{m^d}{N}} + e^{-\theta m} \right), \quad \theta > 0.$$

However, this observation also applies to Weighted Stochastic Mesh(WSM) algorithm of Broadie and Glasserman [17], making almost all algorithms intractable. This motivates the development of more flexible complexity metric for the comparison of the algorithms for optimal stopping problems.

In turns out that not much is known about the convergence properties of WSM method except some preliminary results in discrete case [1]. The authors, however, do not give the dependence of the errors on the underlying dimension and the number of stopping times and their analysis is based on a rather restrictive assumption of compact state space. Similar type of algorithm we present here was also analyzed in the work of Rust [63], but there the setting of discrete-time Markov Decision Process was considered and therefore, the analysis does not directly transfer to optimal stopping. Also the paper considers very restrictive assumptions of compact state space and Lipschitz continuity of transition densities with Lipschitz constant not depending on the dimension d.

1.2 Complexity Metrics

It turns out that the criterion (1.3) puts too much importance on the dimension d on the one hand and on the other hand is too relaxed in dependence on ε. With such definition the algorithm with complexity $d^2\exp\left(\varepsilon^{-1}/\ln\ln ... \ln \varepsilon^{-1}\right)$ is tractable while one with complexity $2^d/\varepsilon$ is not despite that running an algorithm with the former complexity seems to be practically impossible even with $d = 1$. Therefore, we proposed another approach to tractability.

Definition 3. *For an algorithm with computational complexity $\mathcal{C}(\varepsilon, d)$ the number*

$$\Gamma_{\mathcal{C}} := \limsup_{d\to\infty} \limsup_{\varepsilon\to 0} \frac{\ln \mathcal{C}(\varepsilon, d)}{d \ln(1/\varepsilon)}. \tag{1.4}$$

is called semitractability index.

Definition 4. *The problem is called semitractable if there exists an algorithm solving it with $\Gamma_{\mathcal{C}} = 0$.*

1.3 WSM Algorithm

Let us present a Weighted Stochastic Mesh (WSM) algorithm for a discrete-time optimal stopping problem. The algorithm is inspired by [17] but it differs in special choice of weights and truncation level. First, let us define the discrete Snell envelope process:

$$U_l = U_l(Z_l) := \sup_{\tau\in\mathcal{T}_{l,L}} \mathbb{E}\left[g_\tau(Z_\tau) \;\middle|\; \mathcal{F}_l\right], \quad l = 0, ..L,$$

where $\mathcal{T}_{l,L}$ is the set of stopping times taking values in the set $\{l, .., L\}$. Snell envelope satisfies dynamic programming principle, therefore, we can compute U_l using backward induction:

$$\begin{aligned} &U_L(Z_L) = g_L(Z_L), \\ &U_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[U_{l+1}(Z_{l+1}) \;\middle|\; Z_l\right]\right\}, \quad l = 0, .., L-1. \end{aligned}$$

For technical purposes of the analysis we set truncation level $R > 0$ and define the truncated version of this backward induction:

$$\widetilde{U}_L(Z_L) = g_L(Z_L), \tag{1.5}$$

$$\widetilde{U}_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}) \ \mid \ Z_l\right]\right\} \cdot \mathbb{1}_{B_R}(Z_l), \quad l = 0, .., L-1, \tag{1.6}$$

where $\mathbb{1}_{B_R}$ is the indicator function of the 0-centered euclidean ball of radius R in $\mathbb{R}^d$. Thus, the values vanish when the process is out of B_R. Also by construction it holds that

$$\|\widetilde{U}_l\|_\infty \leq G_R \stackrel{\text{def}}{=} \max_{0\leq l\leq L} \sup_{z\in B_R} g_l(z), \tag{1.7}$$

We sample N independent trajectories $(Z_l^{(n)})_{l\in\mathbb{L}}$ with $Z_0^{(n)} = x_0, n = 1, .., N$ with the help of transition density $p(y|x)$. To estimate the conditional expectations, we use the following approximation:

$$\mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}) \ \mid \ Z_l = x\right] \approx \sum_{n=1}^{N} \widetilde{U}_{l+1}\left(Z_{l+1}^{(n)}\right) \frac{p\left(Z_{l+1}^{(n)} \mid x\right)}{\sum_{m=1}^{N} p\left(Z_{l+1}^{(n)} \mid Z_l^{(m)}\right)}. \tag{1.8}$$

We start by setting $\overline{U}_L(Z_L^{(n)}) = g_L(Z_L^{(n)})$ for $n = 1, .., N$. Once $\overline{U}_{l+1}$ is constructed on the grid for $0 < l + 1 \leq L$, we proceed via dynamic programming and set

$$\overline{U}_l(Z_l^{(r)}) \stackrel{\text{def}}{=} \max\left\{g_l(Z_l^{(r)}), \sum_{n=1}^{N} \overline{U}_{l+1}^{(n)}(Z_{l+1}^{(n)}) \frac{p(Z_{l+1}^{(n)}|Z_l^{(r)})}{\sum_{m=1}^{N} p(Z_{l+1}^{(n)}|Z_l^{(m)})}\right\} \mathbb{1}_{Z_l^{(r)}\in B_R}, \tag{1.9}$$

To sum up, WSM algorithm is as follows:

1. Simulate N independent trajectories $(Z_l^{(1)})_{l\in\mathcal{L}}, .., (Z_l^{(N)})_{l\in\mathcal{L}}$;
2. Set $\overline{U}_L(Z_L^{(n)}) = g_L(Z_L^{(n)})$ for $n = 1, .., N$;
3. For $l = L-1, .., 1$ compute $\overline{U}_l(Z_l^{(n)})$ for all $n = 1, .., N$ using (1.6) and (1.8) for approximation of the conditional expectation;
4. Compute
$$\overline{U}_0(x_0) = \max\left\{g_0(x_0)\ ,\ \frac{1}{N}\sum_{n=1}^{N} \overline{U}_1^{(n)}\left(Z_1^{(n)}\right)\right\}.$$

One more thing to notice is that one step of backward induction with (1.6) and (1.8) takes $N^2 c_*$ with c_* being the price of multiplication. Thus, the total computational cost of the algorithms is $c_* N^2 L$ and given that $c_* \ll c_f^{(d)}$, the cost of one computation of transition density, it is bounded from above by $c_f^{(d)} N^2 L$.

1.4 Error and complexity analysis in discrete time

In this section we analyze convergence of the WSM estimate to the solution of the discrete optimal stopping problem for $l = 0$ and a fixed $x_0 \in \mathbb{R}^d$ as $N \to \infty$. Let us first bound a distance between U_l and $\widetilde{U}_l$, $l = 0, \ldots, L$.

Proposition 2. *With*

$$\varepsilon_{l,R} \stackrel{def}{=} \int_{|x-x_0|>R} U_l(x) p_l(x|x_0)\, dx$$

$l = 0, \ldots, L$, *it holds that*

$$\int \left|U_l(x) - \widetilde{U}_l(x)\right| p_l(x|x_0)\, dx \le \sum_{j=l}^{L} \varepsilon_{j,R}. \tag{1.10}$$

Proposition 3. *Suppose that*

$$\max_{0\le l\le L} g_l(x) \le c_g(1+|x|), \quad x \in \mathbb{R}^d \tag{1.11}$$

and that

$$\mathbb{E}\left[\max_{l\le l'\le L} |Z_{l'}| \,\middle|\, Z_l = x\right] \le c_Z(1+|x|), \quad x \in \mathbb{R}^d. \tag{1.12}$$

Suppose further that for some $\varkappa$, $\alpha > 0$, *and* $l = 1, \ldots, L$,

$$0 < p_l(y|x) \le \frac{\varkappa}{(2\pi\alpha l)^{d/2}} e^{-\frac{|x-y|^2}{2\alpha l}} \tag{1.13}$$

for all $x, y \in \mathbb{R}^d$. *One then has*

$$\int \left|U_l(x) - \widetilde{U}_l(x)\right| p_l(x|x_0)\, dx \\ \le L c_g \varkappa \left(1 + c_Z + c_Z\,|x_0| + c_Z\sqrt{d\alpha L}\right) 2^{d/4} e^{-\frac{R^2}{8\alpha L}}. \tag{1.14}$$

Next we control the discrepancy between $\overline{U}_0$ and $\widetilde{U}_0$.

Proposition 4. *With*

$$F_R^2 \stackrel{def}{=} \max_{1\le l\le L} \int\int_{|y-x_0|\le R} \frac{p^2(y|x)}{p_{l+1}(y|x_0)}\, p_l(x|x_0)\, dxdy, \tag{1.15}$$

and N *such that* $(1+F_R)/\sqrt{N} < 1$, *it holds that*

$$\mathbb{E}\left[\left|\overline{U}_0 - \widetilde{U}_0\right|\right] \le \left(3+\sqrt{2}\right) L G_R \frac{1+F_R}{\sqrt{N}}.$$

Corollary 5. *Under the assumptions of Proposition 3, we have for (1.15) the estimate*

$$F_R^2 \le \frac{\varkappa}{(2\pi\alpha)^{d/2}} \mathtt{Vol}(B_R) = \frac{\varkappa R^d}{(2\alpha)^{d/2}\Gamma\left(1+d/2\right)} \le \varkappa\,(e/\alpha)^{d/2}\, R^d d^{-d/2},$$

where the last inequality follows from $\Gamma(1+a) \ge a^a e^{-a}$ *for any* $a \ge 1/2$. *Then by combining (1.14) with Proposition 4 we obtain the error estimate,*

$$\mathbb{E}\left[\left|U_0 - \overline{U}_0\right|\right] \le L c_g \varkappa \left(1 + c_Z + c_Z\,|x_0| + c_Z\sqrt{d\alpha L}\right) 2^{d/4} e^{-\frac{R^2}{8\alpha L}} \\ + \left(3+\sqrt{2}\right) L c_g (1+R) \frac{1 + \varkappa^{1/2}\,(e/\alpha)^{d/4}\, R^{d/2} d^{-d/4}}{\sqrt{N}}. \tag{1.16}$$

LS	WSM	QTM
$3/\alpha$	0	2

Table 1.1: Tractability index Γ of different algorithms for discrete time optimal stopping problems

Theorem 6. *Under the assumptions of Proposition 3 the complexity of the WSM algorithm is bounded from above by*

$$\mathcal{C}(\varepsilon, d) = c_1 \alpha^2 c_g^4 \varkappa^2 c_f^{(d)} c_2^d L^{d+7} \varepsilon^{-4} \times \log^{d+2}\left[\frac{L\left(1 + c_Z + c_Z\left|x_0\right|\right) e^{\frac{c_Z\sqrt{\alpha L}}{1+c_Z+c_Z|x_0|}} 2^{3/4}\left(c_g \varkappa \vee 1\right)}{\varepsilon}\right], \quad (1.17)$$

where $c_1 > 0$ and $c_2 > 1$ are natural constants and $c_f^{(d)}$ stands for the cost of computing the transition density $p_l(y|x)$ at one point (x, y).

Corollary 7. *For a fixed $L > 0$ the discrete-time optimal stopping problem with g and $(Z_l)_{l\geq 0}$ satisfying* (1.11), (1.12) *and* (1.13) *is semi-tractable, provided that the complexity of computing the transition density $p_l(y|x)$ at one point (x, y) is at most polynomial in d. Different approximation algorithms for discrete time optimal stopping problems can be compared using the tractability index* (1.4). *For example, it follows from* (1.2) *that the tractability index of the LS approach is equal to $3/\alpha$. If the continuation functions are analytic, then the tractability index for the LS approach becomes zero. Moreover from inspection of Theorem 2.4 in [6], we see that the Quantization Tree Method (QTM) has tractability index* 2.

1.4.1 Approximation of the transition density

A crucial condition for semi-tractability in the discrete exercise case is the availability of the transition density $p(y|x)$ of the chain $(Z_l)_{l\geq 0}$ in a closed (or cheaply computable) form. However, we can show that if the sequence of approximating densities $p^n(y|x)$, $n \in \mathbb{N}$ converging to $p(y|x)$ can be constructed in such a way that

$$\left|\frac{p^n(y|z) - p(y|z)}{p^n(y|z)}\right| \lesssim \frac{(1 + |y - x_0|^m + |z - x_0|^m)^n}{n!}, \quad y, z \in B_{R_n} \quad (1.18)$$

for some $m \in \mathbb{N}$ and a sequence $R_n \nearrow \infty$, $n \nearrow \infty$, then under proper assumptions on the growth of R_n and the cost of computing p^n (in fact it should be at most polynomial in d), one can derive a complexity bound $\mathcal{C}(\varepsilon, d)$ satisfying

$$\lim_{\varepsilon \searrow 0} \frac{\log \mathcal{C}(\varepsilon, d)}{\log \frac{1}{\varepsilon}} \text{ is finite and does not depend on } d\ .$$

The proof involves a (rather straightforward) extension of the present one based on exact transition densities. But, on the one hand, one of the main results in this paper, tractability index 2 of the continuous time stopping problem, does not rely on transition density approximation, and on the other hand, such a proof would entail a notational blow up and might detract the reader from the main lines, therefore the details are omitted.

To construct a sequence of approximations $p^n(y|z)$ satisfying the assumption (1.18), one can use various small-time expansions for transition densities of stochastic processes, see, for example, [4] and [50]. Let us exemplify this type of approximation in the case of one-dimensional diffusion processes of the form:

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t, \quad X_0 = x_0,$$

where b is a bounded function, twice continuously differentiable, with bounded derivatives and σ is a function with three continuous and bounded derivatives such that there exist two positive constants $\sigma_\circ, \sigma^\circ$ with $\sigma_\circ \le \sigma(x) \le \sigma^\circ$. Consider a Markov chain $(Z_l)_{l\ge 0}$ defined as a time discretization of $(X_t)_{t\ge 0}$, that is, $Z_l \stackrel{\text{def}}{=} X_{\Delta l}$, $l = 0, 1, 2, \ldots$ for some $\Delta > 0$. Under the above conditions the following representation for the (one-step) transition density p of the chain Z is proven in [31] (see also [23] for more general setting):

$$p(y|x) = \frac{1}{\sqrt{2\pi\Delta}}\frac{1}{\sigma(y)}\exp\left(-\frac{(s(x)-s(y))^2}{2\Delta}\right) U_\Delta(s(x), s(y)), \quad x, y \in \mathbb{R},$$

with $U_\Delta(x,y) = R_\Delta(x,y)\exp\left[\int_0^x \bar b(z)\,dz - \int_0^y \bar b(z)\,dz\right]$,

$$R_\Delta(x,y) = \mathbb{E}\left[\exp\left(-\Delta\int_0^1 \bar\rho(x + z(y-x) + \sqrt{\Delta}B_z)\,dz\right)\right], \tag{1.19}$$

where B_z is a standard Brownian bridge, $s(x) = \int_0^x \frac{dy}{\sigma(y)}$, $g = s^{-1}$ and

$$\bar\rho = (\bar b^2 + \bar b')/2 \quad \text{with} \quad \bar b = (b/\sigma)\circ g - \sigma'\circ g/2.$$

Note that the expectation in (1.19) is taken with respect to the known distribution of the Brownian bridge B_z. By expanding the exponent in (1.19) into Taylor series, we get for Δ small enough

$$p(x|y) = \frac{1}{\sqrt{2\pi\Delta}}\frac{1}{\sigma(y)}\exp\left(-\frac{(s(x)-s(y))^2}{2\Delta}\right) \times \exp\left[\int_0^x \bar b(z)\,dz - \int_0^y \bar b(z)\,dz\right]\sum_{k=0}^{\infty}\frac{\Delta^k}{k!}c_k(x,y)$$

with

$$c_k(x,y) = (-1)^k\mathbb{E}\left[\left(\int_0^1 \bar\rho(x + z(y-x) + \sqrt{\Delta}B_z)\,dz\right)^k\right].$$

If $\bar\rho$ is uniformly bounded by a constant $D > 0$, then the above series converges uniformly in x and y for all Δ small enough. Set

$$p^n(x|y) = \frac{1}{\sqrt{2\pi\Delta}}\frac{1}{\sigma(y)}\exp\left(-\frac{(s(x)-s(y))^2}{2\Delta}\right) \times \exp\left[\int_0^x \bar b(z)\,dz - \int_0^y \bar b(z)\,dz\right]\left\{\sum_{k=0}^{n}\frac{\Delta^k}{k!}c_k(x,y)\right\}.$$

It obviously holds $p^n(y|x) > 0$ for $\Delta < \Delta_0(D)$ and

$$\left| \frac{p^n(y|z) - p(y|z)}{p^n(y|z)} \right| \leq \frac{(\Delta D)^n}{(1 - \Delta D \exp(\Delta D))} \tag{1.20}$$

uniformly for all $x, y \in \mathbb{R}$. Hence the assumption (1.18) is satisfied with $m = 0$, provided that $\Delta < \Delta_0$ for some Δ_0 depending only on D. Similarly if $\bar{\rho} \leq 0$, then (1.18) holds. To sample from p^n we can use the well-known acceptance rejection method which does not require the exact knowledge of a scaling factor $\int p^n(y|x)\, dy$.

1.5 Continuous time optimal stopping for diffusions

In this section we consider diffusion processes of the form

$$dX_s^i = b_i(X_s)\, ds + \sum_{j=1}^{m} \sigma_{ij}(X_s)\, dW_s^j, \quad X_0^i = x_0^i, \quad i = 1, \ldots, d, \tag{1.21}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d \times m}$, are Lipschitz continuous and $W = (W^1, \ldots, W^m)$ is a m-dimensional standard Wiener process on a probability space $(\Omega, \mathcal{F}, P)$. As usual, the (augmented) filtration generated by $(W_s)_{s \geq 0}$ is denoted by $(\mathcal{F}_s)_{s \geq 0}$. We are interested in solving optimal stopping problems of the form:

$$U_t^\star = \operatorname*{esssup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}[e^{-r(\tau - t)} f(X_\tau) | \mathcal{F}_t], \tag{1.22}$$

where f is a given real valued function on $\mathbb{R}^d$, $r \geq 0$, and $\mathcal{T}_{t,T}$ stands for the set of stopping times τ taking values in $[t, T]$. The problem (1.22) is related to the so-called free boundary problem for the corresponding partial differential equation. Let us introduce the differential operator L_t :

$$L_t u(t,x) = \frac{1}{2} \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial^2 u}{\partial x^i \partial x^j}(t,x) + \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x^i}(t,x),$$

where

$$a_{ij}(x) = \sum_{k=1}^{d} \sigma_{ik}(x) \sigma_{jk}(x).$$

We denote by $X_s^{t,x}$ (or $X^{t,x}(s)$), $s \geq T$, the solution of (1.21) starting at moment t from x : $X_t^{t,x} = x$. Denote by $u(t,x)$ a regular solution of the following system of partial differential inequalities:

$$\begin{gathered} \frac{\partial u}{\partial t} + L_t u - ru \leq 0,\ u \geq f, \quad (t,x) \in [0,T) \times \mathbb{R}^d, \\ \left(\frac{\partial u}{\partial t} + L_t u - ru \right)(f - u) = 0, \quad (t,x) \in [0,T) \times \mathbb{R}^d, \\ u(T,x) = f(x), \quad x \in \mathbb{R}^d, \end{gathered} \tag{1.23}$$

then under some mild conditions (see, e.g. [41])

$$u(t,x) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}[e^{-r(\tau - t)} f(X_\tau^{t,x})] \quad , \ (t,x) \in [0,T] \times \mathbb{R}^d, \tag{1.24}$$

that is, $u(t,x) = U_t^\star(x)$.

With this notation established, it is worth discussing the main issue that we are going to address in this section. Our goal is to estimate $u(t,x)$ at a given point (t_0, x_0) with accuracy less than ε by an algorithm with complexity $\mathcal{C}^\star(\varepsilon, d)$ which is polynomial in $1/\varepsilon$. As already mentioned in the introduction some well known algorithms such as the regression ones fail to achieve this goal (at least according to the existing complexity bounds in the literature).

Let us introduce the Snell envelope process:

$$U_t^\star \stackrel{\text{def}}{=} \operatorname{esssup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}_{\mathcal{F}_t} \left[g(\tau, X_\tau) \right], \tag{1.25}$$

where (somewhat more general than in (1.22)) g is a given nonnegative function on $\mathbb{R}_{\geq 0} \times \mathbb{R}^d$. In the first step we perform a time discretization by introducing a finite set of stopping dates $t_l = lh$, $l = 1, \ldots, L$, with $h = T/L$ and L some natural number, and next consider the discretized Snell envelope process:

$$U_{t_l}^\circ(X_{t_l}) \stackrel{\text{def}}{=} \operatorname*{esssup}_{\tau \in \mathcal{T}_{l,L}} \mathbb{E}_{\mathcal{F}_{t_l}} \left[g(\tau, X_\tau) \right],$$

where $\mathcal{T}_{l,L}$ stands for the set of stopping times with values in the set $\{t_l, \ldots, t_L\}$. Note that the measurable functions $U_{t_l}^\circ(\cdot)$ exist due to Markovianity of the process X. The error due to the time discretization is well studied in the literature. We will rely on the following result which is implied by Thm. 2.1 in [6] for instance.

Proposition 8. *Let $g : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ be Lipschitz continuous and $p \geq 1$. Then one has that*

$$\max_{l=0,\ldots,L} \left\| U_{t_l}^\star(X_{t_l}) - U_{t_l}^\circ(X_{t_l}) \right\|_p \leq \frac{c_\circ e^{C_\circ T}(1 + |x_0|)}{L},$$

where the constants $c_\circ, C_\circ > 0$ depend on the Lipschitz constants for b, σ, and g, respectively.

In order to achieve an acceptable discretization error we choose a sufficiently large L, and then concentrate on the computation of U°.

In the next step we approximate the underlying process X using some strong discretization scheme on the time grid $t_i = iT/L$, $i = 0, \ldots, L$, yielding an approximation $\overline{X}$. It is assumed that the one step transition densities of this scheme are explicitly known. The simplest and the most popular scheme is the Euler scheme,

$$\overline{X}^i_{t_{l+1}} = \overline{X}^i_{t_l} + b_i(\overline{X}_{t_l})\, h + \sum_{j=1}^m \sigma_{ij}(\overline{X}_{t_l}) \left(W^j_{t_{l+1}} - W^j_{t_l} \right), \quad \overline{X}^i_0 = x^i_0, \tag{1.26}$$

$i = 1, \ldots, d$, which in general has strong convergence order $1/2$, and the one-step transition density of the chain $(\overline{X}_{t_{l+1}})_{l \geq 0}$ is given by

$$\bar{p}_h(y|x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi h)^d \, |\Sigma|}} \exp\left[-\frac{1}{2} h^{-1} (y - x - b(x)h)^\top \Sigma^{-1} (y - x - b(x)h) \right] \tag{1.27}$$

with $\Sigma = \sigma\sigma^\top \in \mathbb{R}^{d \times d}$ and $h = T/L$. Now we will turn to the discrete time optimal stopping problem with possible stopping times $\{t_l = lh,\ l = 0, \ldots, L\}$. To this end we introduce the discrete time Markov chain $Z_l \stackrel{\text{def}}{=} \overline{X}_{t_l}$ adapted to the filtration $(\mathcal{F}_l) \stackrel{\text{def}}{=} (\mathcal{F}_{t_l})$,

and $g_l(x) \stackrel{\text{def}}{=} g(t_l, x)$ (while abusing notation slightly) and consider the discretized Snell envelope process

$$U_{t_l}(\overline{X}_{t_l}) \stackrel{\text{def}}{=} \operatorname*{esssup}_{\tau \in \mathcal{T}_{l,L}} \mathbb{E}_{\mathcal{F}_{t_l}} \left[g(\tau, \overline{X}_\tau) \right] = \operatorname*{esssup}_{\iota \in \mathcal{I}_{l,L}} \mathbb{E}_{\mathcal{F}_l} \left[g_\iota(Z_\iota) \right] \stackrel{\text{def}}{=} U_l(Z_l), \tag{1.28}$$

where $\mathcal{I}_{l,L}$ stands for the set of stopping indices with values in $\{l, \dots, L\}$, and the measurable functions $U_{t_l}(\cdot)$ (or $U_l(\cdot)$) exist due to Markovianity of the process $\overline{X}$ (or Z). The distance between U and U° is controlled by the next proposition.

Proposition 9. *There exists a constant $C^{Euler} > 0$ depending on the Lipschitz constants of b, σ, and g, such that*

$$\max_{l=0,\dots,L} \mathbb{E} \left[\left| U^\circ_{t_l}(X_{t_l}) - U_{t_l}(\overline{X}_{t_l}) \right| \right] \leq C^{Euler} \sqrt{h}.$$

Thus, combining Proposition 8 and Proposition 9 yields:

Corollary 10. *If $\overline{X}$ is constructed by the Euler scheme with time step size $h = T/L$, where L is the number of discretization steps, then under the conditions of Proposition 8 and Proposition 9 we have that*

$$\mathbb{E} \left[|U^\star_0(x_0) - U_0(x_0)| \right] \lesssim C^{Euler} \sqrt{h} \quad \text{for } h \to 0, \tag{1.29}$$

where $\lesssim$ stands for inequality up to constant depending on $c_\circ, C_\circ$ and C^{Euler}.

Since the transition densities of the Euler scheme are explicitly known (see (1.27)), the WSM algorithm can be directly used for constructing an approximation $\overline{U}_0(x_0)$ based on the paths of the Markov chain (Z_l). To derive the complexity bounds of the resulting estimate, we shall make the following assumptions.

(AG) Suppose that $c_g > 0$ is such that

$$g(t, x) \leq c_g \left(1 + |x|\right) \quad \text{for all } 0 \leq t \leq T,\ x \in \mathbb{R}^d. \tag{1.30}$$

(AX) Assume that there exists a constant $c_{\bar{X}} > 0$ such that for all $0 \leq l \leq L$,

$$\mathbb{E}_{\mathcal{F}_{t_l}} \Big[\sup_{l \leq l' \leq L} \left| \overline{X}_{l'h} \right| \Big| \overline{X}_{lh} = x \Big] \leq c_{\bar{X}} \left(1 + |x|\right), \quad x \in \mathbb{R}^d, \tag{1.31}$$

uniformly in L (hence h). This assumption is satisfied under Lipschitz conditions on the coefficients of the SDE (1.21), and can be proved using the Burkholder-Davis-Gundy inequality and the Gronwall lemma.

(AP) Assume furthermore that $\left(\overline{X}_{lh},\ l = 0, \dots, L\right)$ is time homogeneous with transition densities $\overline{p}_{lh}(y|x)$ that satisfy the Aronson type inequality: there exist positive constants $\overline{\varkappa}$ and $\overline{\alpha}$ such that for any $x, y \in \mathbb{R}^d$ and any $l > 0$, it holds that

$$0 < \overline{p}_{lh}(y|x) \leq \frac{\overline{\varkappa}}{(2\pi\overline{\alpha} lh)^{d/2}} e^{-\frac{|x-y|^2}{2\overline{\alpha} lh}}.$$

This assumption holds if the coefficients in (1.21) are bounded and σ is uniformly elliptic.

The next proposition provides complexity bounds for the WSM algorithm in the case of continuous time optimal stopping problems.

Proposition 11. *Assume that the assumptions (AG), (AX) and (AP) hold, then*

- *the cost of computing* $U_0(x_0)$ *in* (1.28) *for a fixed* $L > 0$ *with precision* $\varepsilon > 0$ *via the WSM algorithm is bounded from above by*

$$\mathcal{C}(\varepsilon, d) = c_1 \overline{\alpha}^2 c_g^4 \varkappa^2 c_f^{(d)} c_2^d \frac{T^{d+7}}{h^{d+5}} \times \varepsilon^{-4} \log^{d+2} \left[\frac{\frac{T}{h} \left(1 + c_{\bar{X}} + c_{\bar{X}} |x_0|\right) e^{\frac{c_{\bar{X}} \sqrt{\overline{\alpha} T}}{1 + c_{\bar{X}} + c_{\bar{X}} |x_0|}} 2^{3/4} \left(c_g \varkappa \vee 1\right)}{\varepsilon} \right]. \quad (1.32)$$

- *the cost of computing* $U_0^\star(x_0)$ *with an accuracy* $\varepsilon > 0$ *via the WSM algorithm is bounded from above by*

$$\mathcal{C}^\star(\varepsilon, d) = c_1 \overline{\alpha}^2 c_g^4 \varkappa^2 c_f^{(d)} c_2^d \frac{T^{d+7}}{\varepsilon^{2d+14}} \times \log^{d+2} \left[\frac{T \left(1 + c_{\bar{X}} + c_{\bar{X}} |x_0|\right) e^{\frac{c_{\bar{X}} \sqrt{\overline{\alpha} T}}{1 + c_{\bar{X}} + c_{\bar{X}} |x_0|}} 2^{3/4} \left(c_g \varkappa \vee 1\right)}{\varepsilon} \right]. \quad (1.33)$$

The first statement follows directly from Proposition 6 by taking in (1.17), $\alpha = \overline{\alpha} h$, $c_Z = c_{\bar{X}}$, and $L = T/h$. Then by setting $h \asymp \varepsilon^2$ we obtain (1.33) (with possibly modified natural constants c_1, c_2).

Discussion. As can be seen from (1.33),

$$\Gamma_{\mathrm{WSM}} = \lim_{d \nearrow \infty} \lim_{\varepsilon \searrow 0} \frac{\log \mathcal{C}^\star(\varepsilon, d)}{d \log \varepsilon^{-1}} = 2 \quad (1.34)$$

and this shows the efficiency of the proposed algorithm as compared to the existing algorithms for continuous time optimal stopping problems at least as far as the tractability index is concerned. Indeed, the only algorithm available in the literature with a provably finite limit of type (1.34) is the quantization tree method (QTM) of Bally, Pagès, and Printems [6]. Indeed, by tending the number of stopping times and the quantization number to infinity such that the corresponding errors in Thm. 2.4-b in [6] are balanced, we derive the following complexity upper bound

$$\mathcal{C}^\star_{\mathrm{QTM}}(\varepsilon, d) = O\left(\frac{1}{\varepsilon^{6d+6}}\right) \quad (1.35)$$

Hence $\Gamma_{\mathrm{QTM}} = 6$.

Summarizing. For discrete time optimal stopping problems we have established semi-tractability for the proposed WSM algorithm with respect to rather general Markov chains governed by certain transition kernels. Note that in the most common case of infinitely smooth continuation functions, many regression algorithms including the LS and TV algorithms lead to semi-tractable in discrete time optimal stopping problems. But when passing to continuous stopping problems, the tractability index of the WSM method remains bounded (equal to two) while the tractability index of the regression methods tends to infinity.

LS	WSM	QTM
∞	2	6

Table 1.2: Tractability index Γ of different algorithms for continuous time optimal stopping problems.

1.6 Numerical Experiments

In the following experiments we illustrate the WSM algorithm in the case of continuous-time optimal stopping problems. A lower bound for the value function in WSM method is obtained using a suboptimal stopping rule computed on an independent set of trajectories (test set). This stopping rule can be constructed using any interpolation algorithm based on the observations from the training trajectories. The fastest and the simplest way giving good results is the nearest neighbor interpolation, in our experiments we have chosen the number of nearest neighbors to be 500.

American put option on a single asset

To illustrate the performance of the WSM algorithm in continuous time, we consider a problem of pricing American put option on a single asset driven by geometric Brownian motion

$$X_t = X_0 e^{\sigma W_t + (r-\sigma/2)t}$$

with r denoting the riskless rate of interest, assumed to be constant, and σ being the constant volatility. The payoff function is given by

$$g(x) = \max(K - x, 0).$$

The fair price of an option is defined as

$$U_0 = \sup_{\tau \in \mathcal{T}_{[0,T]}} \mathbb{E}\left[e^{-r\tau} g(X_\tau)\right]$$

for which there is no closed form solution but there exist numerical methods giving accurate approximations to U_0. We used parameters $r = 0.08, \sigma = 0.20, K = X_0 = 100, T = 3$. An accurate estimate of U_0 in this particular case is obtained and reported in [44] to be 6.9320. In Fig. 1.1 we show the lower bounds obtained by WSM, LS and VF (value function regression method of [80]) in dependence of the number of stopping opportunities L setting uniform time discretization on $[0, T]$ (the larger L the more dense is the grid). As can be seen, WSM lower bound is much more stable when L increases and LS and VF needs to use more complex regression basis to compensate for this effect.

American max-call option on five assets

The model with $d = 5$ assets is considered where each underlying asset has dividend yield δ. The dynamics is set by

$$dX_t^k = (r - \delta)X_t^k dt + \sigma X_t^k dW_t^k, \quad k = 1, .., d,$$

where W_t^k are independent one-dimensional Brownian motions. The parameters are set to be $r = 0.05, \delta = 0.1, \sigma = 0.2$. As before, the holder may exercise the option at any time $t \in [0, T]$ with $T = 3$ and receive the payoff

$$g(X_t) = max\left(\max\left(X_t^1, .., X_t^d\right) - K,\ 0\right).$$

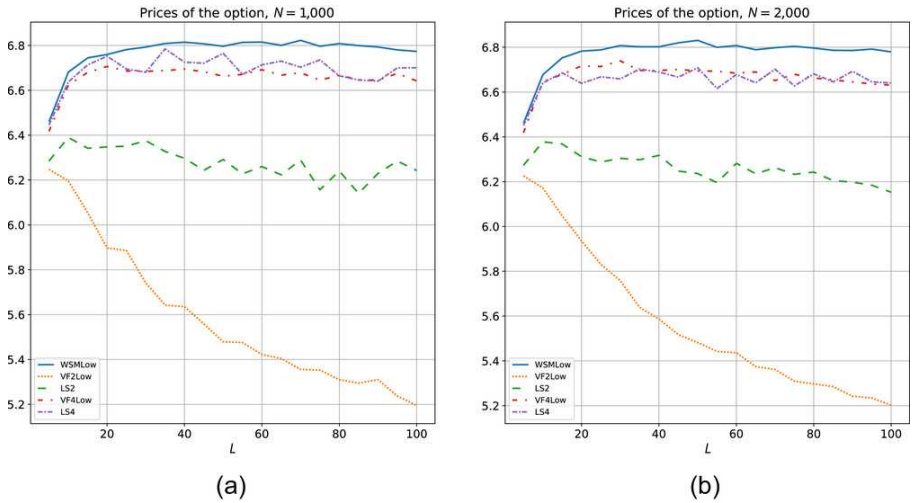


Figure 1.1: Lower bounds for the price of one-dimensional American put option approximated using different methods and uniform time discretization $t_k = kT/L, k = 0, .., L$ of exercise dates. The numbers of training paths are $N_{train} = 1000$(a) and $N_{train} = 2000$(b) and the number of test trajectories used for constructing the lower bounds $N_{test} = 20000$ and is the same in both cases. In LS and VF a polynomial basis of degrees 2 and 4 is used (mentioned in the legend).

We apply WSM and LS (with a basis of degree-2 polynomials) techniques to construct a lower bound. The results for different L are presented in Fig. 1.2. The option price must increase when the number of stopping opportunities increases, therefore LS-algorithm has clearly deteriorating estimate. WSM, on the other hand has increasing lower bound which shows that it performs considerably better than LS.

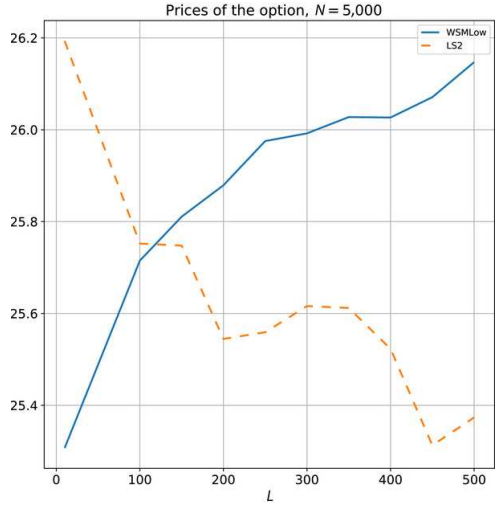


Figure 1.2: Lower bounds for the price of a five-dimensional American put option approximated using a uniform grid $t_k = kT/L, k = 0, .., L$ of exercise dates. The number of training paths is $N_{train} = 2000$ and the number of test trajectories is $N_{test} = 5000$.

1.7 Conclusions

In our work we have presented the complexity analysis of weighted stochastic mesh algorithm for solving optimal stopping problem. We also have analyzed the qualities of the algorithm not only in discrete time but also in continuous time by using sufficient degree of time discretization. Our theoretical results regarding semitractability demonstrate that the algorithm is the best in regard to this metric having semitractability index 0 in discrete-time problem and 2 in continuous one. Semitractability of continuous-time optimal stopping problem, therefore, remains to be an open question. Nevertheless, we have seen the superior performance of WSM algorithm in continuous-time problem. It turns out that the quality of the estimate in regression methods degrades very fast when making the time discretization more dense and one needs to compensate with introducing new basis functions which leads to even more computational effort. Such thing is not observed in WSM method: the estimate remains stable with all fixed parameters and varying time discretization.

Chapter 2

Finite Time Analysis of Linear Two-Timescale Stochastic Approximation with Markovian Noise

The results of this section are published in [43].

2.1 Introduction

Since its introduction close to 70 years ago, the stochastic approximation (SA) scheme [62] has been a powerful tool for root finding when only noisy samples are available. During the past two decades, considerable progresses in the practical and theoretical research of SA have been made, see [10, 47, 15] for an overview. Among others, linear SA schemes are popular in reinforcement learning (RL) as they lead to policy evaluation methods with linear function approximation, of particular importance is temporal difference (TD) learning [69] for which finite time analysis has been reported in [68, 48, 13, 25].

The TD learning scheme based on classical (linear) SA is known to be inadequate for the off-policy learning paradigms in RL, where data samples are drawn from a *behavior policy* different from the policy being evaluated [5, 79]. To circumvent this problem, [72, 73] have suggested to replace TD learning with the gradient TD (GTD) method or the TD with gradient correction (TDC) method. These methods fall within the scope of linear two-timescale SA scheme introduced by [14]:

$$\theta_{k+1} = \theta_k + \beta_k\{\widetilde{b}_1(X_{k+1}) - \widetilde{A}_{11}(X_{k+1})\theta_k - \widetilde{A}_{12}(X_{k+1})w_k\}, \tag{2.1}$$

$$w_{k+1} = w_k + \gamma_k\{\widetilde{b}_2(X_{k+1}) - \widetilde{A}_{21}(X_{k+1})\theta_k - \widetilde{A}_{22}(X_{k+1})w_k\}. \tag{2.2}$$

The above recursion involves two iterates, $\theta_k \in \mathbb{R}^{d_\theta}$, $w_k \in \mathbb{R}^{d_w}$, whose updates are coupled with each other. In the above, $\widetilde{b}_i(x)$, $\widetilde{A}_{ij}(x)$ are measurable vector/matrix valued functions on X and the random sequence $(X_k)_{k\geq 0}, X_k \in \mathsf{X}$ forms an ergodic Markov chain. The scalars $\gamma_k, \beta_k > 0$ are step sizes. The above SA scheme is said to have two timescales as the step sizes satisfy $\lim_{k\to\infty} \beta_k/\gamma_k < 1$ such that w_k is updated at a faster timescale. In fact, w_k is a 'tracking' term which seeks solution to a linear system characterized by θ_k.

The goal of this research was to characterize the finite time expected error bound with improved convergence rate for the two timescale SA (2.1),(2.2).

The almost sure convergence of two timescale SA have been established in [14, 75, 76, 15], among others; the asymptotic convergence rates have been characterized in [46,

57]. However, finite-time risk bounds for two timescale SA have not been analyzed until recently. With martingale samples, [51] provided the first finite time analysis of GTD method, [26, 24] provided improved finite time error bounds. Unlike our analysis, they analyzed modified two timescale SA with projection and their bounds hold with high probability. With Markovian noise, [40] studied the finite time expected error bound with constant step sizes; [88] and [27] provided similar analysis for general step sizes. It is important to notice that with homogeneous martingale noise, the asymptotic rate of (2.1), (2.2) without a projection step, as shown in [46, Theorem 2.6], is in the order $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = \mathcal{O}(\beta_k), \mathbb{E}\left[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2\right] = \mathcal{O}(\gamma_k)$, where $\theta^\star$ is a stationary point of the SA scheme. However, the latter rate is not achieved in the finite-time error bounds analyzed by the above works except for [24]. It remains an open problem whether this error bound holds for the Markovian noise setting and for linear two time-scale SA scheme without projection.

Contributions This chapter presents the following contributions:

- *Improved Convergence Rate* – We perform finite-time expected error bound analysis of the linear two timescale SA in both martingale and Markovian noise settings, in Theorems 12 & 13. Our analysis allow for general step sizes schedules [cf. A10, B8], including constant, piecewise constant, and diminishing step sizes explored in the prior works [40, 24, 88, 27]. We show that the error bound consists of a transient and a steady-state term, and the asymptotic rate is obtained from the latter. We show that this asymptotic rate matches those in [46, Theorem 2.6], i.e., $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = \mathcal{O}(\beta_k), \mathbb{E}\left[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2\right] = \mathcal{O}(\gamma_k)$. In particular, the fastest achievable rate for $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$ will be $\mathcal{O}(1/k)$ when we set $\beta_k = \mathcal{O}(1/k), \gamma_k = \mathcal{O}(1/k^\upsilon)$ with $\upsilon < 1$.

- *Novel Analysis without A-priori Stability Assumption* – Unlike the prior works [51, 24, 88], our convergence results are obtained *without* requiring a projection step throughout the SA iterations. In fact, [24] have pointed out that the projection step is merely included to ensure *a-priori* stability of the algorithm, and is often not used in practice. Our relaxation and the ability to achieve the optimal convergence rate are obtained through a tight analysis of the recursive inequalities of the (cross-)variances of θ_k, w_k, see Section 2.3.

- *Asymptotic Expansion* – With an additional assumption on the step size, we compute an exact asymptotic expansion of the expected error $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$, see Theorem 22. With an appropriate diminishing step sizes schedule, we show that the expected error cannot be smaller than $\Omega(\beta_k)$, which matches our upper bound results in Theorem 12 & 13.

The rest of this paper is organized as follows. In Section 2.2, we present the detailed conditions for two timescale linear SA, and the main results on finite-time performance bounds. In Section 2.3, we provide an outline of the proof, illustrating the insights behind the main steps. In Section 2.4, we show that the finite-time error bounds are tight by quantifying an exact expansion of the covariance of iterates. In Section 2.5, we illustrate the theoretical findings using numerical experiments.

Notations Let $n \in \mathbb{N}$ and Q be a symmetric definite $n \times n$ matrix. For $x \in \mathbb{R}^n$, we denote $\|x\|_Q = \{x^\top Q x\}^{1/2}$. For brevity, we set $\|x\| = \|x\|_{\mathrm{I}}$. Let $m \in \mathbb{N}$, P be a symmetric

definite $m \times m$ matrix, A be an $n \times m$ matrix. A matrix A is said to be Hurwitz if the real parts of its eigenvalues are strictly negative. We denote $\|A\|_{P,Q} = \max_{\|x\|_P=1} \|Ax\|_Q$. If A is a $n \times n$ matrix, we denote $\|A\|_Q = \|A\|_{Q,Q}$. Lastly, we give a number of auxiliary lemmas in Appendix B.4 that are instrumental to our analysis.

2.2 Linear Two Time-scale Stochastic Approximation (SA) Scheme

We investigate the linear two timescale SA given by the following equivalent form of (2.1), (2.2):

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{11}\theta_k - A_{12}w_k + V_{k+1}), \tag{2.3}$$
$$w_{k+1} = w_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}w_k + W_{k+1}), \tag{2.4}$$

where the mean fields are defined as $b_i := \lim_{k\to\infty} \mathbb{E}\left[\widetilde{b}_i(X_k)\right]$, $A_{ij} := \lim_{k\to\infty} \mathbb{E}\left[\widetilde{A}_{ij}(X_k)\right]$ (these limits exist as we recall that $(X_k)_{k\geq 0}$ is an ergodic Markov chain). The noise terms V_{k+1}, W_{k+1} are given by:

$$\begin{aligned} V_{k+1} &:= \widetilde{b}_1(X_{k+1}) - b_1 - (\widetilde{A}_{11}(X_{k+1}) - A_{11})\theta_k - (\widetilde{A}_{12}(X_{k+1}) - A_{12})w_k, \\ W_{k+1} &:= \widetilde{b}_2(X_{k+1}) - b_2 - (\widetilde{A}_{21}(X_{k+1}) - A_{21})\theta_k - (\widetilde{A}_{22}(X_{k+1}) - A_{22})w_k. \end{aligned} \tag{2.5}$$

The goal of the recursion (2.3), (2.4) is to find a stationary solution pair $(\theta^\star, w^\star)$ that solves the system of linear equations:

$$A_{11}\theta + A_{12}w = b_1, \quad A_{21}\theta + A_{22}w = b_2. \tag{2.6}$$

We are interested in the scenario when the solution pair $(\theta^\star, w^\star)$ is unique and is given by

$$\theta^\star = \Delta^{-1}(b_1 - A_{12}A_{22}^{-1}b_2), \quad w^\star = A_{22}^{-1}(b_2 - A_{21}\theta^\star). \tag{2.7}$$

where $\Delta := A_{11} - A_{12}A_{22}^{-1}A_{21}$. To analyze the convergence of $(\theta_k, w_k)_{k\geq 0}$ in (2.3), (2.4) to $(\theta^\star, w^\star)$, we require the following assumptions:

A 9. Matrices $-A_{22}$ and $-\Delta = -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)$ are *Hurwitz*.

The above assumption is common for linear two time-scale SA, see [46]. As a consequence, using the Lyapunov lemma (stated in Lemma 32 in the appendix for completeness), there exist positive definite matrices $Q_{22}^\top = Q_{22} \succ 0, Q_\Delta^\top = Q_\Delta \succ 0$ satisfying

$$A_{22}^\top Q_{22} + Q_{22}A_{22} = \mathrm{I}, \quad Q_\Delta\Delta + \Delta^\top Q_\Delta = \mathrm{I}\,. \tag{2.8}$$

This ensures the contraction (see Lemma 33 in the appendix):

$$\|\mathrm{I} - \gamma_k A_{22}\|_{Q_{22}} \leq 1 - a_{22}\gamma_k, \quad \|\mathrm{I} - \beta_k\Delta\|_{Q_\Delta} \leq 1 - a_\Delta\beta_k, \tag{2.9}$$

provided that $\gamma_k \in [0, 1/(2\|A_{22}\|_{Q_{22}}^2\|Q_{22}\|)], \beta_k \in [0, 1/(2\|A_\Delta\|_{Q_\Delta}^2\|Q_\Delta\|)]$. Moreover, we have set $a_{22} := 1/(4\|Q_{22}\|)$, $a_\Delta := 1/(4\|Q_\Delta\|)$. We consider the following conditions on the step sizes:

A 10. $(\gamma_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$ are nonincreasing sequences of positive numbers that satisfy the following.

1. There exist constants κ such that for all $k \in \mathbb{N}$, we have $\beta_k/\gamma_k \leq \kappa$.

2. For all $k \in \mathbb{N}$, it holds

$$\gamma_k/\gamma_{k+1} \leq 1+(a_{22}/8)\gamma_{k+1}, \;\; \beta_k/\beta_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}, \;\; \gamma_k/\gamma_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}. \tag{2.10}$$

As a consequence, we can define $\varsigma := 1 + \{\gamma_0 a_{22}/8 \vee \beta_0 a_\Delta/16\}$ such that $\gamma_k/\gamma_{k+1} \leq \varsigma$, $\beta_k/\beta_{k+1} \leq \varsigma$. Our conditions on step sizes are similar to [46, Assumption 2.3, 2.5]. These conditions encompass diminishing, piecewise constant and constant step sizes schedules which are common in the literature. For instance, a popular choice of diminishing step sizes satisfying A10 is

$$\beta_k = c^\beta/(k + k_0^\beta), \quad \gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3} \tag{2.11}$$

with some constants c^β, c^γ, k_0^γ, k_0^β, e.g., as suggested in [26, Remark 9]; or a constant step size of $\beta_k = \beta, \gamma_k = \gamma$; or a piecewise constant step size, e.g., [40].

We present new results on the convergence rate of (2.3), (2.4) depending on the types of noise with V_{k+1}, W_{k+1}. To discuss these cases, let us define the σ-field generated by the two timescale SA scheme and the initial error made by the SA scheme, respectively as:

$$\mathcal{F}_k := \sigma\big\{\theta_0, w_0, X_1, X_2, ..., X_k\big\}, \quad \mathrm{V}_0 := \mathbb{E}\left[\|\theta^0 - \theta^\star\|^2 + \|w^0 - w^\star\|^2\right]. \tag{2.12}$$

Our main results are presented as follows.

Martingale Noise We consider a simple setting where the random elements X_k are drawn i.i.d. from the stationary distribution such that b_i, A_{ij} are the expected values of $\widetilde{b}_i(X_k), \widetilde{A}_{ij}(X_k)$. Furthermore, the random variables $\widetilde{b}_i(X_k), \widetilde{A}_{ij}(X_k)$ have bounded second order moment. Note that this implies $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] = \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$, i.e., the sequences $(V_{k+1})_{k\in\mathbb{N}}, (W_{k+1})_{k\in\mathbb{N}}$ are martingale difference sequences. Formally, we describe this setting as the following conditions on V_{k+1}, W_{k+1}:

A 11. The noise terms are zero-mean conditioned on $\mathcal{F}_k$, i.e., $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] = \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$.

A 12. There exist constants m_W, m_V such that

$$\begin{aligned}
&\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\| \leq m_V(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|), \\
&\|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| \leq m_W(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|).
\end{aligned} \tag{2.13}$$

Theorem 12. *Assume A9–12 and for all $k \in \mathbb{N}$, we have $\gamma_k \in [0, \gamma_\infty^{\mathsf{mtg}}]$, $\beta_k \in [0, \beta_\infty^{\mathsf{mtg}}]$ and $\kappa \in [0, \kappa_\infty]$, where $\gamma_\infty^{\mathsf{mtg}}, \beta_\infty^{\mathsf{mtg}}, \kappa_\infty$ are defined constants. Then*

$$\mathbb{E}\left[\|\theta_k - \theta^*\|^2\right] \leq d_\theta\left\{\mathrm{C}_0^{\tilde{\theta},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\tilde{\theta},\mathsf{mtg}}\beta_k\right\} \tag{2.14}$$

$$\mathbb{E}\left[\left\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\right\|^2\right] \leq d_w\Big\{\mathrm{C}_0^{\widehat{w},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\widehat{w},\mathsf{mtg}}\gamma_k\Big\} \tag{2.15}$$

The exact constants are provided in the appendix, see (B.37), (B.41).

Markovian Noise Consider the sequence $(X_k)_{k\geq 0}$ to be samples from an exogenous Markov chain on X with the transition kernel $\mathrm{P} : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$. For any measurable function f, we have

$$\mathbb{E}^{\mathcal{F}_k}\left[f(X_{k+1})\right] = \mathrm{P}\, f(X_k) = \int_{\mathsf{X}} f(x)\, \mathrm{P}(X_k, \mathrm{d}x)$$

We state the following assumptions:

B 5. The Markov kernel P has a unique invariant distribution $\mu : \mathsf{X} \to \mathbb{R}_+$. Moreover, it is irreducible and aperiodic.

Observe that

$$b_i = \int_{\mathsf{X}} \widetilde{b}_i(x)\, \mu(\mathrm{d}x), \quad A_{ij} = \int_{\mathsf{X}} \widetilde{A}_{ij}(x)\, \mu(\mathrm{d}x),\ i, j = 1, 2.$$

We show that the linear two time-scale SA (2.1), (2.2) converges to a unique fixed point defined by the above mean field vectors/matrices, see (2.7). An important condition that enables our analysis is the existence of solutions to the following Poisson equations:

B 6. For any $i, j = 1, 2$, consider $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$, there exists vector/matrix valued measurable functions $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ which satisfy

$$\widetilde{b}_i(x) - b_i = \widehat{b}_i(x) - \mathrm{P}\,\widehat{b}_i(x), \quad \widetilde{A}_{ij}(x) - A_{ij} = \widehat{A}_{ij}(x) - \mathrm{P}\,\widehat{A}_{ij}(x) \tag{2.16}$$

for any $x \in \mathsf{X}$ and b_i, A_{ij} are the mean fields of $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$ with the stationary distribution μ.

The above assumption can be guaranteed under B5 together with some regularity conditions, see [28, Section 21.2]. Moreover,

B 7. Under B6, the vector/matrix valued functions $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ are uniformly bounded: for any $i, j = 1, 2$, $x \in \mathsf{X}$,

$$\|\widehat{b}_i(x)\| \leq \overline{\mathrm{b}},\ \|\widehat{A}_{ij}(x)\| \leq \overline{\mathrm{A}}. \tag{2.17}$$

B 8. There exists constant ρ_0 such that for any $k \geq 1$, we have $\gamma_{k-1}^2 \leq \rho_0 \beta_k$.

To satisfy B7, we observe that the bounds $\overline{\mathrm{b}}, \overline{\mathrm{A}}$ depend on the mixing time of the chain $(X_k)_{k\geq 0}$ and a uniform bound on $\widetilde{b}_i(\cdot), \widetilde{A}_{ij}(\cdot)$. In the context of reinforcement learning, the latter can be satisfied when the feature vectors and reward are bounded. Note that B7 implies A12, see Section 2.3.2. Meanwhile, B8 imposes further restriction on the step size. The latter can also be satisfied by (2.11).

The challenges of analysis with Markovian noise lie in the biasedness of the noise term as $\mathbb{E}^{\mathcal{F}_k}\left[V_{k+1}\right] \neq 0$, $\mathbb{E}^{\mathcal{F}_k}\left[W_{k+1}\right] \neq 0$. With a careful analysis, we obtain:

Theorem 13. *Assume A9–10, B5–8 hold and for all $k \in \mathbb{N}$, we have $\beta_k \in (0, \beta_\infty^{\mathsf{mark}}]$, $\gamma_k \in (0, \gamma_\infty^{\mathsf{mark}}]$, $\kappa \leq \kappa_\infty$, where $\beta_\infty^{\mathsf{mark}}$, $\gamma_\infty^{\mathsf{mark}}$, κ_∞ are defined in* (2.34), (2.21). *Then*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq d_\theta \left\{ \mathrm{C}_0^{\tilde{\theta},\mathsf{mark}} \prod_{\ell=0}^{k-1} \Big(1 - \beta_\ell \frac{a_\Delta}{8}\Big)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\tilde{\theta},\mathsf{mark}}\, \beta_k \right\}, \tag{2.18}$$

$$\mathbb{E}\left[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2\right] \leq d_w \left\{ \mathrm{C}_0^{\widehat{w},\mathsf{mark}} \prod_{\ell=0}^{k-1} \Big(1 - \beta_\ell \frac{a_\Delta}{8}\Big)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\widehat{w},\mathsf{mark}} \gamma_k \right\}. \tag{2.19}$$

The exact constants are given in the appendix, see (B.106), (B.109).

While Theorem 13 relaxes the martingale difference assumption A12 in Theorem 12, we remark that the results here do not generalize that in Theorem 12 due to the additional B7, B8. Particularly, with martingale noise, the convergence of linear two timescale SA only requires the noise to have bounded *second order moment*, yet the Markovian noise needs to be uniformly bounded.

Convergence Rate of Linear Two Timescale SA The upper bounds in Theorem 12 and 13 consist of two terms – the first term is a 'transient' error with product such as $\prod_{i=0}^{k-1}(1-\beta_i a_\Delta/8)$ decays to zero at the rate $o(1/k^c)$ for some $c>1$ under an appropriate choice of step sizes such as (2.11); the second term is a 'steady-state' error. We observe that the 'steady-state' error of the iterates θ_k, w_k exhibit different behaviors. Taking the step size choices in (2.11) as an example, the steady-state error of the slow-update iterates θ_k is $\mathcal{O}(1/k)$ while the error of fast-update iterates w_k is $\mathcal{O}(1/k^{\frac{2}{3}})$. Furthermore, similar bounds hold for *both* martingale and Markovian noise. In Section 2.4 we show that the obtained rates are also tight.

Comparison to Related Works Our results improve the convergence rate analysis of linear two timescale SA in a number of recent works. In the martingale noise setting (Theorem 12), the closest work to ours is [24] which analyzed the linear two timescale SA with martingale samples and diminishing step sizes. The authors improved on [26] and obtained the same convergence rate (in high probability) as our Theorem 12, furthermore it is demonstrated that the obtained rates are tight. Their bounds also exhibit a sublinear dependence on the dimensions d_θ, d_w. However, their algorithm involves a sparsely executed projection step and the error bound holds only for a sufficiently large k. These restrictions are lifted in our analysis.

In the Markovian noise setting (Theorem 13), the closest works to ours are [27, 40, 88]. In particular, [40] analyzed the linear two timescale SA with constant step sizes and showed that the steady-state error for both θ_k, w_k is $\mathcal{O}(\gamma^2/\beta)$. [88] analyzed the TDC algorithm with a projection step and showed that the steady-state error for θ_k is $\mathcal{O}(1/k^{\frac{2}{3}})$ if the step sizes in (2.11) is used. [27] analyzed the linear two timescale SA with diminishing step size and showed that the steady state error for both θ_k, w_k is $\mathcal{O}(1/k^{\frac{2}{3}})$. Interestingly, the above works do not obtain the fast rate in Theorem 13, i.e., $\mathbb{E}\left[\|\theta_k-\theta^\star\|^2\right]=\mathcal{O}(1/k)$. One of the reasons for the sub-optimality in their rates is that their analysis are based on building a single Lyapunov function that controls both errors in θ_k and w_k. In contrast, our analysis relies on a set of coupled inequalities to obtain tight bounds for each of the iterates θ_k, w_k.

2.3 Convergence Analysis

While much of the technical details and the complete constants of non-asymptotic bounds will be postponed to the appendix, this section offers insights into our main theoretical results through sketching the major steps involved in proving Theorem 12 & 13. Throughout, we shall consider the following bounds on the step sizes and step size ratio:

$$\beta_\infty^{(0)} := \frac{1}{2\|Q_\Delta\|\|\Delta\|_{Q_\Delta}^2} \wedge \frac{1}{2\|\Delta\|_{Q_\Delta}+a_\Delta}, \quad \gamma_\infty^{(0)} := \frac{1}{2\|Q_{22}\|\|A_{22}\|_{Q_{22}}^2}, \tag{2.20}$$

$$\kappa_\infty := \left(\frac{a_{22}/2}{\|A_{12}\|_{Q_{22},Q_\Delta}\|A_{22}^{-1}A_{21}\|_{Q_\Delta,Q_{22}} + \frac{a_\Delta}{2}} \Big\{ 1 \wedge \frac{a_\Delta/2}{\|\Delta\|_{Q_\Delta} + \frac{a_\Delta}{2}} \Big\} \right) \wedge \frac{a_{22}}{4a_\Delta}. \tag{2.21}$$

To begin with, let us present the reformulation of the two time-scale SA scheme (2.3), (2.4) that is borrowed from [46]. Define:

$$L_{k+1} := \big(L_k - \gamma_k A_{22}L_k + \beta_k A_{22}^{-1}A_{21}(\Delta - A_{12}L_k)\big)\big(\mathrm{I} - \beta_k(\Delta - A_{12}L_k)\big)^{-1}, \quad L_0 := 0,$$

and $\mathrm{L}_\infty := a_\Delta/(2\|A_{12}\|_{Q_{22},Q_\Delta})$. As shown in Lemma 34 of the appendix, with the step sizes $\gamma_k \le \gamma_\infty^{(0)}$, $\beta_k \le \beta_\infty^{(0)}$, $\kappa \le \kappa_\infty$, the above recursion on L_k is well defined where it holds that $\|L_k\|_{Q_\Delta,Q_{22}} \le \mathrm{L}_\infty$ for any $k \ge 0$. In addition, define the matrices:

$$B_{11}^k := \Delta - A_{12}L_k, \quad B_{22}^k := \frac{\beta_k}{\gamma_k}\big(L_{k+1} + A_{22}^{-1}A_{21}\big)A_{12} + A_{22}, \quad C_k := L_{k+1} + A_{22}^{-1}A_{21}.$$

In a similar vein as performing Gaussian elimination, we obtain a simplified two timescale SA recursions (proof in Appendix B.1):

Proposition 14. *Consider the following change-of-variables:*

$$\tilde{\theta}_k := \theta_k - \theta^\star, \quad \tilde{w}_k = w_k - w^\star + C_{k-1}\tilde{\theta}_k. \tag{2.22}$$

The two time-scale SA (2.3), (2.4) *is equivalent to the following iterations:*

$$\tilde{\theta}_{k+1} = (\mathrm{I} - \beta_k B_{11}^k)\tilde{\theta}_k - \beta_k A_{12}\tilde{w}_k - \beta_k V_{k+1}. \tag{2.23}$$

$$\tilde{w}_{k+1} = (\mathrm{I} - \gamma_k B_{22}^k)\tilde{w}_k - \beta_k C_k V_{k+1} - \gamma_k W_{k+1}. \tag{2.24}$$

Observe that $\tilde{\theta}_k = 0, \tilde{w}_k = 0$ is equivalent to having $\theta_k = \theta^\star, w_k = w^\star$, i.e., the two timescale SA solves the linear system of equations (2.6). The simplified recursion (2.23), (2.24) *decouples* the update of $\tilde{w}_k$ from $\tilde{\theta}_k$. This allows one to treat the $\tilde{w}_k$ update as a one timescale linear SA, and therefore provides a shortcut to perform a tight analysis. We focus on estimating the following operator norms of covariances:

$$\mathrm{M}_k^{\tilde{w}} := \|\mathbb{E}\left[\tilde{w}_k\tilde{w}_k^\top\right]\|, \quad \mathrm{M}_k^{\tilde{\theta}} := \|\mathbb{E}\left[\tilde{\theta}_k\tilde{\theta}_k^\top\right]\|, \quad \mathrm{M}_k^{\tilde{\theta},\tilde{w}} := \|\mathbb{E}\left[\tilde{\theta}_k\tilde{w}_k^\top\right]\|, \tag{2.25}$$

which are respectively the covariance for w_k, θ_k and the cross-variance between w_k, θ_k.

2.3.1 Proof Outline of Theorem 12

For this theorem, we assume the step sizes and their ratio are chosen such that

$$\gamma_k \le \gamma_\infty^{\mathsf{mtg}} := \gamma_\infty^{(0)} \wedge \frac{1}{\frac{a_{22}}{2} + \frac{2}{a_{22}}p_{22}(\tilde{m}_V + \kappa^2\tilde{m}_W)} \wedge \frac{a_\Delta}{4\,\mathrm{C}_2^{\tilde{\theta}}}, \quad \beta_k \le \beta_\infty^{\mathsf{mtg}} := \beta_\infty^{(0)}, \tag{2.26}$$

where $p_{22} = \lambda_{\mathsf{min}}^{-1}(Q_{22})\lambda_{\mathsf{max}}(Q_{22})$ and $\mathrm{C}_2^{\tilde{\theta}}$ is defined in (B.37) in the appendix.

While the property which the noise terms satisfy $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] = 0$, $\mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$ has greatly simplified the analysis, the challenge with our analysis lies in the coupling between slow and fast updating iterates whose convergence rates must be carefully characterized in order to obtain the desired rate in Theorem 12. To summarize, our proof consists of three steps in order: (i) we bound $\mathrm{M}_k^{\tilde{w}}$ with an inequality that is coupled with $\mathrm{M}_k^{\tilde{\theta}}$; then (ii) we bound the cross term $\mathrm{M}_k^{\tilde{\theta},\tilde{w}}$ using an inequality coupled with $\mathrm{M}_k^{\tilde{\theta}}$; lastly, (iii) these bounds are combined to bound $\mathrm{M}_k^{\tilde{\theta}}$.

Step 1: Bounding $\mathrm{M}_k^{\tilde{w}}$ Upon applying the variable transformation in Observation 14, (2.24) can be treated as a one-timescale SA which updates $\tilde{w}_k$ independently, and the contributions from $\tilde{\theta}_k$ are only found in the noise term, as seen from (B.4). This leads to:

Proposition 15. *Assume A9–12 and the step sizes satisfy* (2.26)*. For any $k \in \mathbb{N}$, it holds*

$$\mathrm{M}_{k+1}^{\tilde{w}} \leq \prod_{\ell=0}^{k} \left(1-\tfrac{\gamma_\ell a_{22}}{2}\right) \tfrac{\lambda_{\max}(Q_{22})}{\lambda_{\min}(Q_{22})}\, \mathrm{M}_0^{\tilde{w}} + \mathrm{C}_1^{\tilde{w}}\, \gamma_{k+1} + \mathrm{C}_2^{\tilde{w}} \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \left(1-\tfrac{\gamma_\ell a_{22}}{2}\right) \mathrm{M}_j^{\tilde{\theta}}, \quad (2.27)$$

where the constants $\mathrm{C}_1^{\tilde{w}}, \mathrm{C}_2^{\tilde{w}}$ can be found in (B.19) *in the appendix.*

The right hand side of (2.27) consists of three components: (i) a fast decaying term relying on the product $\prod_{\ell=0}^{k}(1-\gamma_\ell a_{22}/2)$, (ii) an $\mathcal{O}(\gamma_k)$ term, and (iii) a convolutive term between $\mathrm{M}_k^{\tilde{\theta}}$ and the fast decaying term depending on the step size sequence $(\gamma_k)_{k\geq 0}$. In the above, the second term can be viewed as a 'steady-state' term.

Step 2: Bounding $\mathrm{M}_k^{\tilde{\theta},\tilde{w}}$ Observe that $\mathrm{M}_k^{\tilde{\theta},\tilde{w}}$ refers to the cross variance between $\tilde{w}_k$ and $\tilde{\theta}_k$. We show that utilizing (2.23), (2.24), (2.27) allows us to derive:

Proposition 16. *Assume A9–12 and the step sizes satisfy* (2.26)*. For any $k \in \mathbb{N}$, it holds*

$$\mathrm{M}_{k+1}^{\tilde{\theta},\tilde{w}} \leq \mathrm{C}_0^{\tilde{\theta},\tilde{w}} \prod_{\ell=0}^{k} \left(1-\tfrac{\gamma_\ell a_{22}}{2}\right) + \mathrm{C}_1^{\tilde{\theta},\tilde{w}}\, \beta_{k+1} + \mathrm{C}_2^{\tilde{\theta},\tilde{w}} \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \left(1-\tfrac{\gamma_\ell a_{22}}{2}\right) \mathrm{M}_j^{\tilde{\theta}}, \quad (2.28)$$

where the constants $\mathrm{C}_0^{\tilde{\theta},\tilde{w}}, \mathrm{C}_1^{\tilde{\theta},\tilde{w}}, \mathrm{C}_2^{\tilde{\theta},\tilde{w}}$ can be found in (B.31) *in the appendix.*

The above bound is a crucial step in obtaining the $\mathcal{O}(\beta_k)$ rate for $\mathrm{M}_k^{\tilde{\theta}}$. To better appreciate it, note that as $\mathrm{M}_k^{\tilde{\theta},\tilde{w}} \leq (\sqrt{d_\theta d_w}/2)\{\mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}\}$ (see Lemma 39 in the appendix), one can derive a similar result to (2.28) by merely applying Proposition 15. However, doing so results in an overestimated 'steady-state' error of $\mathcal{O}(\gamma_k)$ which is worse than the $\mathcal{O}(\beta_k)$ error in (2.28). On the other hand, we take care of the two timescale nature of the algorithm to obtain (2.28) with the fast rate.

Step 3: Bounding $\mathrm{M}_k^{\tilde{\theta}}$ Having equipped ourselves with Proposition 15 and 16, we can analyze $\mathrm{M}_k^{\tilde{\theta}}$ using (2.23) and the derived bounds on $\mathrm{M}_k^{\tilde{w}}, \mathrm{M}_k^{\tilde{\theta},\tilde{w}}$, this leads to

Proposition 17. *Assume A9–12 and the step sizes satisfy* (2.26)*. For any $k \in \mathbb{N}$, it holds*

$$\mathrm{M}_{k+1}^{\tilde{\theta}} \leq \mathrm{C}_0^{\tilde{\theta}} \prod_{\ell=0}^{k} \left(1-\tfrac{\beta_\ell a_\Delta}{2}\right) + \mathrm{C}_1^{\tilde{\theta}}\, \beta_{k+1} + \mathrm{C}_2^{\tilde{\theta}} \sum_{j=0}^{k} \gamma_j \beta_j \prod_{\ell=j+1}^{k} \left(1-\tfrac{\beta_\ell a_\Delta}{2}\right) \mathrm{M}_j^{\tilde{\theta}}, \quad (2.29)$$

where the constants $\mathrm{C}_0^{\tilde{\theta}}, \mathrm{C}_1^{\tilde{\theta}}, \mathrm{C}_2^{\tilde{\theta}}$ are given in (B.37) *in the appendix.*

Besides that the middle term is now $\mathcal{O}(\beta_k)$, we also observe that the convolution term with $(\mathrm{M}_j^{\tilde{\theta}})_{j\geq 0}$ depends on the *product of step sizes* $\beta_j\gamma_j$. This bound is obtained using Proposition 16 and the fact that the cross variance $\mathrm{M}_k^{\tilde{\theta},\tilde{w}}$ has a steady-state error of $\mathcal{O}(\beta_k)$.

Eq. (2.29) is a recursive inequality as $\mathrm{M}_k^{\tilde{\theta}}$ are found on both sides. In the appendix, we show that there exists a sequence $(\mathrm{U}_k)_{k\geq 0}$ satisfying $\mathrm{M}_k^{\tilde{\theta}} \leq \mathrm{U}_k$ and

$$\mathrm{U}_{k+1} \leq (1-\beta_k a_\Delta/4)\, \mathrm{U}_k + \mathrm{C}_1^{\tilde{\theta}}(a_\Delta/2)\beta_k^2 \quad (2.30)$$

for some constant $\mathrm{C}_1^{\tilde{\theta}}$. This immediately leads to (2.14), followed by (2.15) similarly.

2.3.2 Proof Outline of Theorem 13

While our proof has largely followed the same strategy as in the martingale noise case, now that the main challenge in handling the Markovian noise case is that the noise terms V_{k+1}, W_{k+1} are no longer (conditionally) zero-mean. To circumvent this difficulty, we recall B6 and define the following using the solution of the Poisson equation: for any $i, j = 1, 2$,

$$\begin{aligned} \psi_k^{b_i} &:= \mathrm{P}\,\widehat{b}_i(X_k), \quad \Psi_k^{A_{ij}} := \mathrm{P}\,\widehat{A}_{ij}(X_k), \\ \xi_k^{b_i} &:= \widehat{b}_i(X_{k+1}) - \mathrm{P}\,\widehat{b}_i(X_k), \quad \Xi_k^{A_{ij}} := \widehat{A}_{ij}(X_{k+1}) - \mathrm{P}\,\widehat{A}_{ij}(X_k), \end{aligned} \tag{2.31}$$

where $\xi_k^{b_i}, \Xi_k^{A_{ij}}$ are zero mean when conditioned on $\mathcal{F}_k$. The noise terms (2.5) can be rewritten as

$$\begin{aligned} V_{k+1} &= \underbrace{\xi_k^{b_1} + \Xi_k^{A_{11}}\theta_k + \Xi_k^{A_{12}} w_k}_{=:V_{k+1}^{(0)}} + \underbrace{(\psi_k^{b_1} - \psi_{k+1}^{b_1}) + (\Psi_k^{A_{11}} - \Psi_{k+1}^{A_{11}})\theta_k + (\Psi_k^{A_{12}} - \Psi_{k+1}^{A_{12}}) w_k}_{=:V_{k+1}^{(1)}} \\ W_{k+1} &= \underbrace{\xi_k^{b_2} + \Xi_k^{A_{21}}\theta_k + \Xi_k^{A_{22}} w_k}_{=:W_{k+1}^{(0)}} + \underbrace{(\psi_k^{b_2} - \psi_{k+1}^{b_2}) + (\Psi_k^{A_{21}} - \Psi_{k+1}^{A_{21}})\theta_k + (\Psi_k^{A_{22}} - \Psi_{k+1}^{A_{22}}) w_k}_{=:W_{k+1}^{(1)}}. \end{aligned} \tag{2.32}$$

We observe that $\mathbb{E}^{\mathcal{F}_k}\left[V_{k+1}^{(0)}\right] = 0, \mathbb{E}^{\mathcal{F}_k}\left[W_{k+1}^{(0)}\right] = 0$ and therefore (2.32) separates the noise terms into their martingale $(V_k^{(0)}, W_k^{(0)})$ and Markovian $(V_k^{(1)}, W_k^{(1)})$ components. Under B7, the second order moment of these noise components satisfy A12. Accordingly, we define $\tilde{\theta}_0^{(0)} = \tilde{\theta}_0, \tilde{\theta}_0^{(1)} = 0$, and $\tilde{w}_0^{(0)} = \tilde{w}_0, \tilde{w}_0^{(1)} = 0$ and the recursions:

$$\begin{aligned} \tilde{\theta}_{k+1}^{(i)} &= (\mathrm{I} - \beta_k B_{11}^k)\tilde{\theta}_k^{(i)} - \beta_k A_{12}\tilde{w}_k^{(i)} - \beta_k V_{k+1}^{(i)},\ i = 0, 1, \\ \tilde{w}_{k+1}^{(i)} &= (\mathrm{I} - \gamma_k B_{22}^k)\tilde{w}_k^{(i)} - \beta_k C_k V_{k+1}^{(i)} - \gamma_k W_{k+1}^{(i)},\ i = 0, 1, \end{aligned} \tag{2.33}$$

where it holds that $\tilde{\theta}_k = \tilde{\theta}_k^{(0)} + \tilde{\theta}_k^{(1)}$, $\tilde{w}_k = \tilde{w}_k^{(0)} + \tilde{w}_k^{(1)}$ following from Observation 14. Clearly, $\tilde{\theta}_k^{(0)}, \tilde{w}_k^{(0)}$ (resp. $\tilde{\theta}_k^{(1)}, \tilde{w}_k^{(1)}$) are iterates of the two timescale SA driven by martingale (resp. Markovian) noise. The two sets of recursions are independent except the second order moments of noise are bounded by $\mathrm{M}_k^{\tilde{\theta}}, \mathrm{M}_k^{\tilde{w}}$, containing the contributions from $\tilde{\theta}_k^{(0)}, \tilde{w}_k^{(0)}$ and $\tilde{\theta}_k^{(1)}, \tilde{w}_k^{(1)}$.

In the sequel, we show the martingale noise driven terms $\|\mathbb{E}\left[\tilde{w}_k^{(0)}(\tilde{w}_k^{(0)})^\top\right]\|$, $\|\mathbb{E}\left[\tilde{w}_k^{(0)}(\tilde{\theta}_k^{(0)})^\top\right]\|$, $\|\mathbb{E}\left[\tilde{\theta}_k^{(0)}(\tilde{\theta}_k^{(0)})^\top\right]\|$ can be estimated using similar procedures as in Proposition 15–17 from the previous subsection. Meanwhile the Markovian noise driven terms $\|\mathbb{E}\left[\tilde{w}_k^{(1)}(\tilde{w}_k^{(1)})^\top\right]\|$ vanish at a faster rate than the former. Throughout this subsection, we set the step sizes to satisfy:

$$\gamma_k \le \gamma_\infty^{\mathsf{mark}} := \gamma_\infty^{(0)} \wedge \frac{1/\sqrt{d_\theta \vee d_w}}{6p_{22}\,\mathrm{E}_0^{WV}} \wedge \frac{a_{22}/4}{\widetilde{\mathrm{C}}_0 + \widetilde{\mathrm{C}}_3}, \quad \beta_k \le \beta_\infty^{\mathsf{mark}} := \beta_\infty^{(0)} \wedge \frac{1}{\sqrt{6\widetilde{\mathrm{C}}_3^{(1,1)}}} \wedge \frac{a_\Delta}{8\widetilde{\mathrm{C}}_2^{\tilde{\theta}}}, \tag{2.34}$$

where $p_{22} = \lambda_{\mathsf{min}}^{-1}(Q_{22})\lambda_{\mathsf{max}}(Q_{22})$, $\widetilde{\mathrm{C}}_0$, $\widetilde{\mathrm{C}}_3$, E_0^{WV} are defined in (B.46), (B.57), (B.51), respectively, and $\widetilde{\mathrm{C}}_3^{(1,1)}$, $\widetilde{\mathrm{C}}_2^{\tilde{\theta}}$ are defined in (B.103), (B.106), respectively, in the appendix.

Step 1: Bounding $\mathrm{M}_k^{\tilde{w}}$ We first show that the martingale and Markov noise driven iterates converge with different rates as follows:

Lemma 18. *Assume A9–10, B5–8 and the step sizes satisfy* (2.34)*. For any* $k \in \mathbb{N}$*, it holds*

$$\begin{aligned}\|\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| &\leq \textstyle\prod_{\ell=0}^k \left(1-\frac{\gamma_\ell a_{22}}{2}\right)^2 \frac{\lambda_{\max}(Q_{22})}{\lambda_{\min}(Q_{22})}\,\mathrm{M}_0^{\tilde{w}} \\ &\quad + \widetilde{\mathrm{C}}_0 \textstyle\sum_{j=0}^k \gamma_j^2 \prod_{\ell=j+1}^k \left(1-\frac{\gamma_\ell a_{22}}{2}\right)^2 (1+\mathrm{M}_j^{\tilde{w}}+\mathrm{M}_j^{\tilde{\theta}}),\end{aligned} \tag{2.35}$$

$$\begin{aligned}\|\mathbb{E}\left[\tilde{w}_{k+1}^{(1)}(\tilde{w}_{k+1}^{(1)})^\top\right]\| &\leq \widetilde{\mathrm{C}}_1 \textstyle\prod_{\ell=0}^k \left(1-\frac{\gamma_\ell a_{22}}{2}\right)^2 + \widetilde{\mathrm{C}}_2\gamma_k^2(\mathrm{M}_{k+1}^{\tilde{\theta}}+\mathrm{M}_{k+1}^{\tilde{w}}) + \widetilde{\mathrm{C}}_4\gamma_k^2 \\ &\quad + \widetilde{\mathrm{C}}_3\gamma_{k+1} \textstyle\sum_{j=0}^k \gamma_j^2 \prod_{\ell=j+1}^k \left(1-\frac{\gamma_\ell a_{22}}{2}\right)^2 (\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}}),\end{aligned} \tag{2.36}$$

where $\widetilde{\mathrm{C}}_0, \widetilde{\mathrm{C}}_1, \widetilde{\mathrm{C}}_2, \widetilde{\mathrm{C}}_3, \widetilde{\mathrm{C}}_4$ *are constants defined in* (B.46)*,* (B.57) *in the appendix.*

Let us compare the 'steady-state' error on the right hand side of both inequalities: second term of (2.35) and the second to fourth term of (2.36). We observe those in the Markovian noise driven iterates $\tilde{w}_k^{(1)}$ are $\mathcal{O}(\gamma_k)$ times smaller than the martingale noise driven counterparts, indicating a faster convergence. This is roughly due to the special structure of the Markovian noise in $V_k^{(1)}, W_k^{(1)}$, where each term can be written as successive differences of a bounded sequence, e.g., $V_k^{(1)} \approx \xi_k - \xi_{k+1}$. When the linear SA (2.33) is run over a long time horizon, the noise terms from consecutive iterations (roughly) cancels each other, leading to a significantly a smaller 'steady-state' error.

Using $\tilde{w}_k = \tilde{w}_k^{(0)} + \tilde{w}_k^{(1)}$ together with the above lemma give the following estimate for $\mathrm{M}_k^{\tilde{w}}$:

Proposition 19. *Assume A9–10, B5–8 and the step sizes satisfy* (2.34)*. For any* $k \in \mathbb{N}$*, it holds*

$$\mathrm{M}_{k+1}^{\tilde{w}} \leq \prod_{\ell=0}^k \left(1-\frac{\gamma_\ell a_{22}}{4}\right)\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\gamma_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{w}} \sum_{j=0}^k \gamma_j^2 \prod_{\ell=j+1}^k \left(1-\frac{\gamma_\ell a_{22}}{4}\right) \mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}}\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde{\theta}}, \tag{2.37}$$

where $\widetilde{\mathrm{C}}_0^{\tilde{w}}, \widetilde{\mathrm{C}}_1^{\tilde{w}}, \widetilde{\mathrm{C}}_2^{\tilde{w}}, \widetilde{\mathrm{C}}_3^{\tilde{w}}$ *are defined in* (B.66) *in the appendix.*

We note in passing that by considering a special case with $\mathrm{M}_k^{\tilde{\theta}} = 0$ for all k, the above proposition generalizes [68, Theorem 7] for linear one timescale SA with Markovian noise.

In a similar vein to the proof of Theorem 12, we bound the cross term $\|\mathbb{E}\left[\tilde{\theta}_k^{(0)}(\tilde{w}_k^{(0)})^\top\right]\|$ as:

Lemma 20. *Assume A9–10, B5–8 and the step sizes satisfy* (2.34)*. For any* $k \in \mathbb{N}$*, it holds*

$$\|\mathbb{E}\left[\tilde{\theta}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| \leq \widetilde{\mathrm{C}}_0^{\tilde{\theta},\tilde{w}} \prod_{\ell=0}^k \left(1-\frac{\gamma_\ell a_{22}}{4}\right) + \widetilde{\mathrm{C}}_1^{\tilde{\theta},\tilde{w}}\beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{\theta},\tilde{w}} \sum_{j=0}^k \gamma_j^2 \prod_{\ell=j+1}^k \left(1-\frac{\gamma_\ell a_{22}}{4}\right) \mathrm{M}_j^{\tilde{\theta}},$$

where the constants $\widetilde{\mathrm{C}}_0^{\tilde{\theta},\tilde{w}}, \widetilde{\mathrm{C}}_1^{\tilde{\theta},\tilde{w}}, \widetilde{\mathrm{C}}_2^{\tilde{\theta},\tilde{w}}$ *are defined in* (B.78) *in the appendix.*

However, we observe that it is unnecessary to derive a similar (tight) bound for $\|\mathbb{E}\left[\tilde{\theta}_k^{(1)}(\tilde{w}_k^{(1)})^\top\right]\|$ as in the above lemma. The reason is that as observed in Lemma 18, the Markovian noise driven terms are anticipated to be sufficiently small compared to the martingale noise driven terms. In particular, a crude bound suffices to obtain the desirable convergence rate of $\mathrm{M}_k^{\tilde{\theta}}$, as we observe next.

Step 2: Bounding $\mathrm{M}_k^{\tilde{\theta}}$ Again we consider the bounds on $\|\mathbb{E}\left[\tilde{\theta}_{k+1}^{(0)}(\tilde{\theta}_{k+1}^{(0)})^\top\right]\|$ and $\mathbb{E}\left[\|\tilde{\theta}_{k+1}^{(1)}\|^2\right]$ separately. As we show in the appendix, both bounds are comparable as the Markovian noise term admits a successive difference structure. Using the decomposition $\tilde{\theta}_k = \tilde{\theta}_k^{(0)} + \tilde{\theta}_k^{(1)}$, we obtain:

Proposition 21. *Assume A9–10, B5–8 and the step sizes satisfy* (2.34)*. For any* $k \in \mathbb{N}$*, it holds*

$$\mathrm{M}_{k+1}^{\tilde{\theta}} \le \widetilde{\mathrm{C}}_0^{\tilde{\theta}} \prod_{\ell=0}^{k} \left(1 - \tfrac{\beta_\ell a_\Delta}{4}\right) + \widetilde{\mathrm{C}}_1^{\tilde{\theta}} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{\theta}} \sum_{i=0}^{k} \beta_i^2 \prod_{\ell=j+1}^{k} \left(1 - \tfrac{\beta_\ell a_\Delta}{4}\right) \mathrm{M}_i^{\tilde{\theta}}, \tag{2.38}$$

where the constants $\widetilde{\mathrm{C}}_0^{\tilde{\theta}}, \widetilde{\mathrm{C}}_1^{\tilde{\theta}}, \widetilde{\mathrm{C}}_2^{\tilde{\theta}}$ *are defined in* (B.106) *in the appendix.*

Equipped with Proposition 21, we can repeat the same steps as in (2.30) to derive an upper bound for $\mathrm{M}_k^{\tilde{\theta}}$ through solving the recursive inequality (2.38). Similar steps also apply for yielding (2.19).

2.4 Tightness of the Finite-time Error Bounds

This section examines the tightness of our finite time error bounds in Theorem 12, 13 through characterizing the squared error $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$ with expansion. We consider the assumption:

A 13. There exist matrices $\Sigma^{11}, \Sigma^{12}, \Sigma^{22}$, and a constant $m_{VW}^{\mathsf{exp}} \ge 0$ such that for all $j \in \mathbb{N}$, it holds

$$\|\mathbb{E}\left[V_j V_j^\top\right] - \Sigma^{11}\| \vee \|\mathbb{E}\left[W_j W_j^\top\right] - \Sigma^{22}\| \vee \|\mathbb{E}\left[V_j W_j^\top\right] - \Sigma^{12}\| \le m_{VW}^{\mathsf{exp}}(\|\mathbb{E}\left[\tilde{\theta}_k \tilde{\theta}_k^\top\right]\| + \|\mathbb{E}\left[\tilde{w}_k \tilde{w}_k^\top\right]\|).$$

Note that A13 implies A12 and therefore poses a stronger assumption. We have

Theorem 22. *Assume A9–11, A13 and for all* $k \in \mathbb{N}$*, we have* $\gamma_k \in [0, \gamma_\infty^{\mathsf{mtg}}]$*,* $\beta_k \in [0, \beta_\infty^{\mathsf{exp}}]$ *and* $\kappa \in [0, \kappa_\infty^{\mathsf{exp}}]$*, where* $\gamma_\infty^{\mathsf{mtg}}, \beta_\infty^{\mathsf{exp}}, \kappa_\infty^{\mathsf{exp}}$ *are constants defined in* (2.26)*,* (B.115)*,* (B.114) *in the appendix. Then for any* $k \ge k_0^{\mathsf{exp}} := \min\{\ell : \sum_{j=0}^{\ell-1} \beta_j \ge \log(2)/(2\|\Delta\|)\}$*, the following expansion holds*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = I_k + J_k. \tag{2.39}$$

The leading term I_k *is given by the following explicit formula*

$$I_k := \sum_{j=0}^{k} \beta_j^2 \operatorname{Tr}\left(\prod_{\ell=j+1}^{k} (\mathrm{I} - \beta_\ell \Delta)\, \Sigma \left\{\prod_{\ell=j+1}^{k} (\mathrm{I} - \beta_\ell \Delta)\right\}^\top\right),$$

where $\Sigma := \Sigma^{11} + A_{12}A_{22}^{-1}\Sigma^{22}A_{22}^{-\top}A_{12}^\top + \Sigma^{12}A_{22}^{-\top}A_{12}^\top + A_{12}A_{22}^{-1}\Sigma^{21}$*. Meanwhile, the following two-sided inequality holds*

$$\mathrm{C}_3^{\mathsf{exp}} \operatorname{Tr}(\Sigma) \le \frac{I_k}{\beta_k} \le \mathrm{C}_4^{\mathsf{exp}} \operatorname{Tr}(\Sigma), \tag{2.40}$$

and J_k *is bounded by*

$$|J_k| \le \mathrm{C}_0^{\mathsf{exp}} \prod_{\ell=0}^{k-1} \left(1 - \frac{a_\Delta}{4}\beta_\ell\right) \mathrm{V}_0 + \mathrm{C}_1^{\mathsf{exp}} \beta_k \left(\gamma_k + \frac{\beta_k}{\gamma_k}\right), \tag{2.41}$$

where V_0 *was defined in* (2.12)*. All constants* $\mathrm{C}_0^{\mathsf{exp}}, \mathrm{C}_1^{\mathsf{exp}}, \mathrm{C}_3^{\mathsf{exp}}, \mathrm{C}_4^{\mathsf{exp}}$ *are given in* (B.147)*,* (B.123) *and* (B.125) *in the appendix, respectively, and they are independent of* β_k, γ_k*.*

The proof is skipped in the interest of space, and it can be found in Appendix B.3. Observe that from (2.41), the dominant term for J_k is given by $\mathcal{O}(\beta_k\gamma_k + \frac{\beta_k^2}{\gamma_k})$. As such, using (2.40), we observe that

$$|J_k|/I_k = \mathcal{O}\left(\gamma_k + \beta_k/\gamma_k\right)$$

If $\lim_{k\to\infty} \beta_k/\gamma_k = 0$, we have $\lim_{k\to\infty} |J_k|/I_k = 0$. Combining (2.39), (2.40) shows that the expected error $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$ is lower bounded by $\Omega(\beta_k)$.

We note that the assumptions A9–11, A13 imposed by the theorem imply A9–A12 required by Theorem 12. Hence, together with (2.14) in Theorem 12, the above observations constitute a *matching* lower bound on the convergence rate of linear two timescale SA with martingale noise. For the Markovian noise setting, we observe that if we impose the assumption that the random elements $(X_k)_{k\geq 0}$ are i.i.d., and $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$ are bounded above for any $i, j = 1, 2$ and $x \in \mathsf{X}$, then A13, B6–B7 can be satisfied. Therefore, the lower bound on the convergence rate also holds.

2.5 Numerical Experiments, Conclusions

We present numerical experiments to support our theoretical claims. We consider (a) a toy example with a randomly generated problem parameters b_i, A_{ij} and i.i.d. samples $(X_k)_{k\in\mathbb{N}}$ such that $\mathbb{E}\left[\widetilde{b}_i(X_k)\right] = b_i$, $\mathbb{E}\left[\widetilde{A}_{ij}(X_k)\right] = A_{ij}$, (b) the Garnet problem [34] with the GTD algorithm [72] using X_k from a simulated Markov chain. For example (a), we compute the stationary point $\theta^\star, w^\star$ exactly using (2.7); for example (b), while it is known that $w^\star = 0$, the solution $\theta^\star$ is computed using Monte Carlo simulation of the matrices $\widetilde{b}_i(X_k), \widetilde{A}_{ij}(X_k)$ with $2 \cdot 10^9$ iterations. The step sizes are chosen as $\beta_k = c^\beta/(k_0^\beta + k), \gamma_k = c^\gamma/(k_0^\beta + k)^\sigma$ with $\sigma \in \{0.5, 0.67, 0.75\}$. In the toy example (a), we have $d_\theta = d_w = 10, k_0^\beta = 10^4, k_0^\gamma = 10^7, c^\beta = 140, c^\gamma = 300$; while for the Garnet problem (b), we have $k_0^\beta = 8 \cdot 10^5, k_0^\gamma = 2 \cdot 10^5, c^\beta = 2300, c^\gamma = 120$. Garnet problem is generated from family $n_S = 30, n_A = 2, b = 2, p = 8$, see [34]. Further details about both experiments are described in Appendix B.5.

We illustrate the convergence rates of the linear two timescale SA on the two problems in Figure 2.1. Note that the plots show the (normalized) steady state errors are $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = \mathcal{O}(\beta_k)$, $\mathbb{E}\left[\|w_k - w^\star\|^2\right] = \mathcal{O}(\gamma_k)$, which hold for both examples on martingale and Markovian noise. In addition, they are independent of the choice of σ. These observations agree with our main results.

2.6 Conclusions

We have provided an improved finite time convergence analysis of the linear two timescale SA on both martingale and Markovian noises with relaxed conditions. Our analysis show that a tight analysis is possible through deriving and solving a sequence of recursive error bounds. Future works include the finite time analysis of nonlinear two timescale SA.

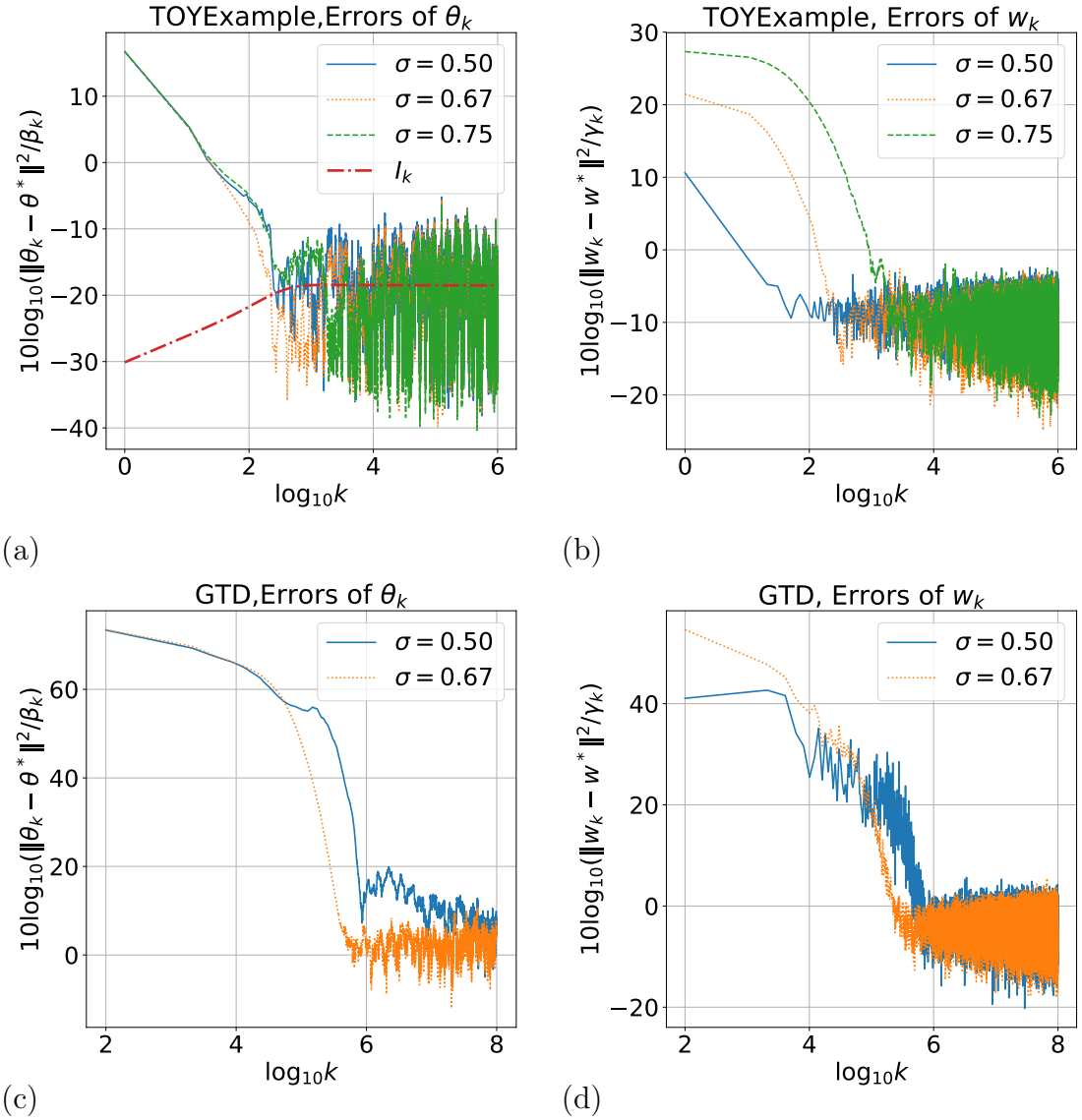


Figure 2.1: Deviations from stationary point $(\theta^\star, w^\star)$ *normalized* by step sizes β_k, γ_k: (a,b) the toy example, note we also show I_k using the exact formula in Theorem 22 (unnormalized plot also available in the Appendix); (c,d) the Garnet problem.

Chapter 3

Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization

The results of this section are published in [42].

3.1 Introduction

Reinforcement learning (RL) is a framework of stochastic control problems where one needs to derive a decision rule (policy) for an agent performing in some stochastic environment giving him rewards or penalties for taking actions. The decision rule is naturally desired to be optimal with respect to some criterion; commonly, expected sum of discounted rewards is used as such criterion[71]. Reinforcement learning is currently a fast-developing area with promising and existing applications in numerous innovative areas of the society: starting from AI for games [82, 11, 66] and going to energy management systems [49, 32], manufacturing and robotics [2] to name a few. Naturally, RL gives the practitioners new sets of control tools for any kind of automatization [33].

Policy-gradient methods constitute the family of gradient algorithms which directly model the policy and exploit various formulas to approximate the gradient of expected reward with respect to the policy parameters [84, 74]. One of the main drawbacks of these approaches is the variance emerging from the estimation of the gradient [83], which typically is high-dimensional. Apart from that, the total sum of rewards is itself a random variable with high variance. Both facts imply that the problem of gradient estimation might be quite challenging. The straightforward way to tackle gradient estimation is Monte Carlo scheme resulting in the algorithm called REINFORCE [84]. In REINFORCE increasing the number of trajectories for gradient estimation naturally reduces the variance but costs a lot of time spent on simulation. Therefore, variance reduction is necessarily required to construct procedures with gradient estimates of lower variance and lower computational cost than increasing the sample size.

The main developments in this direction include actor-critic by [45] and advantage actor-critic: A2C [74] and asynchronous version of it, A3C [55]. Recently a new interest in such methods has emerged due to the introduction of deep reinforcement learning [56] and the frameworks for training nonlinear models like a neural network in RL setting, a very comprehensive review of this area is done by [33]. During several decades a large number of new variance reduction methods were proposed, including sub-sampling meth-

ods like SVRPG [60, 86] and various control variate approaches of [64], [39], [52], [81], [85]. There are also approaches of a bit different nature: trajectory-wise control variates [19] using the control variate based on future rewards and variance reduction in input-driven environments [54]. Apart from that, in ergodic case there were both theoretic [38] and also some practical advancements [21]. The importance of the criteria for variance reduction is well-known in Monte-Carlo and MCMC [65] and recently was also addressed in RL by [30], where the Actor with Variance Estimated Critic (AVEC) was proposed.

Being successful in practice, A2C method is difficult to analyze theoretically. In particular, it remains unclear how the goal functional used in A2C is related to the variance of the gradient estimator. Moreover, the empirical studies of the variance of the gradient estimator are still very rare and available mostly for artificial problems. In the community there is still an ongoing discussion, whether the variance of the gradient really plays main role in the performance of the developed algorithms, according to [81]. In our paper we try to answer some of these questions and suggest a more direct approach inspired by the Empirical Variance(EV) Minimization recently studied by [8]. We show that the proposed EV-algorithm is not only theoretically justifiable but can also perform better than the classic A2C algorithm. It should be noted that the idea of using some kind of empirical variance functional is not new: some hints appeared, for instance, in [52]. Despite that, the implementation and theoretical studies of this approach are still missing in the literature.

3.1.1 Main Contributions

- We provide two new policy-gradient methods (EV-methods) based on EV-criterion and show that they perform well in several practical problems in comparison to A2C-criterion. We have deliberately chosen A2C and Reinforce as baseline algorithms to be less design-specific and have fair comparison of the two criteria. We show that in terms of training and mean rewards EV-methods perform at least as good as A2C but are considerably better in cases with complex policies.

- Theoretical variance bounds are proven for EV-methods. Also we show that EV-criterion addresses the stability of the gradient scheme directly while A2C-criterion is in fact an upper bound for EV. As far as we know, we are the first in the setting of RL who formulates the variance bounds with high probability with the help of the tools of statistical learning.

- We also provide the measurements of the variance of the gradient estimates which present several somewhat surprising observations. Firstly, EV-methods are able to solve this task much better allowing for reduction ratios of 10^3 times . Secondly, in general we see another confirmation the hypothesis of [81]: variance reduction has its effect but some environments are not so responsive to this.

- To our knowledge, we are the first who provide an experimental investigation of EV-criterion of policy-gradient methods in classic benchmark problems and the first implementation of it in the framework of PyTorch. Despite the idea is not new (it is mentioned, for example, by [52]), so far EV-criterion was out of sight mainly because of A2C-criterion is computationally cheaper and is simpler to implement in the current deep learning frameworks since it does not need any complex operations with the gradient.

3.2 EV-Algorithms

3.2.1 Preliminaries

Let us assume a Markov Decision Problem (MDP) $(\mathcal{S}, \mathcal{A}, R, \mathrm{P}, \Pi, \mu_0, \gamma)$ with a finite horizon T, an arbitrary state space $\mathcal{S}$, an action space $\mathcal{A}$, a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, Markov transition kernel P. We are also given a class of policies $\Pi = \{\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A}) \mid \theta \in \Theta\}$ parametrized by $\theta \in \Theta \subset \mathbb{R}^D$ where $\mathcal{P}(\mathcal{A})$ is the set of probability distributions over the action set $\mathcal{A}$. We will omit the subscript in π_θ wherever possible for shorter notation, in all occurrences $\pi \in \Pi$. Additionally we are provided with an initial distribution μ_0, so that $S_0 \sim \mu_0$, and a discounting factor $\gamma \in (0, 1)$. The optimization problem reads as

$$\text{maximize } J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t R(S_t, A_t)\right] \quad \text{w.r.t.} \quad \theta \in \Theta,$$

where we have assumed that the horizon T is fixed. Let us note that any sequence of states, actions, and rewards can be represented as an element X of the product space

$$(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T.$$

A generalization to the cases of infinite horizon and episodes is straightforward: we need to consider the space of sequences

$$(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^\infty$$

for infinite horizon or

$$\bigcup_{L=1}^{\infty} (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^L$$

for the episodes, where the union is the set of all finite sequences. It turns out that the gradient scheme described below still works for these two cases, so we will focus on the finite horizon case only to simplify the exposition.

3.2.2 General Policy Gradient Scheme and REINFORCE

Let $\widetilde{\nabla} J|_{\theta'} : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T \to \mathbb{R}^D$ be an unbiased estimator of the gradient $\nabla_\theta J$ at point $\theta = \theta'$. With this notation the gradient descent algorithm for minimization of $J(\theta)$ using the estimate $\widetilde{\nabla} J$ reads as follows:

$$\theta_{n+1} = \theta_n + \eta_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla} J|_{\theta_n}(X_n^{(k)}), \quad n = 1, 2, \ldots \tag{3.1}$$

with η_n being a positive sequence of step sizes. We will omit the subscript θ_n in the gradient estimate if it is clear from the context at which point the gradient is computed. A simple example of the estimator $\widetilde{\nabla} J$ is the one called REINFORCE [84]:

$$\widetilde{\nabla}^{\text{reinf}} J : \; X \mapsto \sum_{t=0}^{T-1} \gamma^t G_t(X) \nabla_\theta \log \pi(A_t|S_t)$$

with

$$G_t(X) := \sum_{t'=t}^{T-1} \gamma^{t'-t} R_t,$$

where $R_t = R(S_t, A_t)$ and

$$X = \left[(S_0, A_0, R_0), .., (S_{T-1}, A_{T-1}, R_{T-1})\right]^\top.$$

This form is obtained with the help of the following policy gradient theorem.

Proposition 23. *(Policy gradient theorem [84]) If $X \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^\infty$ is sampled from the MDP, then under mild regularity conditions on π, P,*

$$\nabla_\theta J = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t G_t(X) \nabla_\theta \log \pi(A_t|S_t)\right].$$

Note that the above proposition is formulated for the infinite horizon case, but similar statement also holds for the finite-horizon and episodic cases. To see that, one can rewrite the problem as the one with infinite horizon giving zero reward after the end of the trajectory and almost sure transition from the end state to itself.

The baseline approach modifies the above Monte Carlo estimate by incorporating a family of state-action-dependent baselines $b_\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (SA-baselines) or state-dependent baselines $b_\phi : \mathcal{S} \to \mathbb{R}$ (S-baselines) parametrized by ϕ. The resulting gradient estimate reads as

$$\widetilde{\nabla}^{b_\phi} J : X \mapsto \sum_{t=0}^{T-1} \gamma^t (G_t(X) - b_\phi(S_t, A_t)) \nabla_\theta \log \pi(A_t|S_t). \tag{3.2}$$

In order to keep this estimate unbiased, we need to additionally require that for all $\theta \in \Theta$,

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t b_\phi(S_t, A_t) \nabla_\theta \log \pi(A_t|S_t)\right] = 0.$$

It is known that every S-baseline $b_\phi : \mathcal{S} \to \mathbb{R}$ will satisfy this requirement (we have placed the proof in Supplementary Materials, Prop. 50). Such baselines, in particular, are often used in A2C algorithms [33]. When action dependence is presented, special care is required. In fact, the SA-baselines keeping the gradient estimate unbiased are known only in the case of continuous action spaces, see [39, 52, 81, 85]. The main drawback of these methods is that they often are problem-specific. For example, QProp and SteinCV algorithms require the actions to be from continuous set so that one could differentiate the policy with respect to them. QProp additionally needs a notion of mean action. In the case of factorized baselines we need to require the policy to be factorized in coordinates and to construct a vector representation for each action which is not trivial in practice. In this paper we experiment with S-baselines since these allow us fair comparison of the variance reduction procedures with the same models for baseline and policy but the algorithm is applicable generally: it could be used in the gradient routines in place of A2C least-squares criterion.

3.2.3 Two-Timescale Gradient Algorithm with Variance Reduction

If we consider A2C algorithm, we might notice that it can be written as a two-timescale scheme with two step sizes α_n, β_n

$$\theta_{n+1} = \theta_n + \alpha_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla}^{b_\phi} J(X_n^{(k)}), \tag{3.3}$$

$$\phi_{n+1} = \phi_n - \beta_n \nabla_\phi V_{K,n}^{A2C}(\phi)|_{\phi_n} \tag{3.4}$$

where

$$V_{K,n}^{A2C}(\phi) := \frac{1}{K}\sum_{k=1}^{K}\sum_{t=0}^{T-1}(G_t(X_n^{(k)}) - b_\phi(S_t^{(k)}))^2 \tag{3.5}$$

is A2C goal reflecting our desire to approximate the corresponding value function from its noisy estimates $(G_t(X_n^{(k)}))$ via least squares. The motivation behind it is that if one chooses the value function as baseline, the variance will be minimized. This strategy works well in practical problems [55].

If one would like to improve the baseline method there are two ways. One can either construct better baseline families (in which much effort was already invested) or change variance functional in the second timescale. In this work we address the variance of the gradient estimate directly via empirical variance (EV). Since a gradient estimate at iteration n, $\widetilde{\nabla}^{b_\phi} J(X_n)|_{\theta_n}$, is a random vector, we could define its variance as

$$V_n(\theta,\phi) := \mathrm{Tr}\left(\mathbb{E}\left[\left(\widetilde{\nabla}^{b_\phi} J(X_n)|_\theta - \mathbb{E}\left[\widetilde{\nabla}^{b_\phi} J(X_n)|_\theta\right]\right)\cdot\left(\widetilde{\nabla}^{b_\phi} J(X_n)|_\theta - \mathbb{E}\left[\widetilde{\nabla}^{b_\phi} J(X_n)|_\theta\right]\right)^*\right]\right), \tag{3.6}$$

or, what is the same, as

$$V_n(\theta,\phi) = \mathbb{E}\left[\left\|\widetilde{\nabla}^{b_\phi} J(X_n)|_\theta\right\|_2^2\right] - \left\|\mathbb{E}\left[\widetilde{\nabla}^{b_\phi} J(X_n)|_\theta\right]\right\|_2^2. \tag{3.7}$$

Therefore, its empirical analogue is

$$V_{n,K}^{EVv}(\theta,\phi) := \tag{3.8}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi} J(X_n^{(k)})|_\theta\right\|_2^2 - \frac{1}{K^2}\left\|\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi} J(X_n^{(k)})|_\theta\right\|_2^2. \tag{3.9}$$

It can be noticed also, that the second term in the variance (3.7) does not depend on ϕ if the baseline does not add any bias. In this case we could safely discard it before going to sample estimates and use instead

$$V_K^{EVm}(\theta,\phi) := \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi} J(X_n^{(k)})|_\theta\right\|_2^2.$$

The corresponding gradient descent algorithms can be described as

$$\theta_{n+1} = \theta_n + \alpha_n \frac{1}{K}\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi} J(X_n^{(k)}), \tag{3.10}$$

$$\phi_{n+1} = \phi_n - \beta_n \nabla_\phi V_K^{EV}(\phi,\theta)|_{\phi_n,\theta_n}. \tag{3.11}$$

So we have constructed two methods. The first one uses the full variance V_K^{EVv} and is called EVv, the second one is titled EVm and exploits V_K^{EVm}, the same variance functional but without the second term. The important fact to note is that EVv routine would work only if $K \geq 2$, otherwise we try to estimate the variance with one observation.

As was pointed out in [52], the methods addressing the minimization of empirical variance would be computationally very demanding. This though strongly depends on the implementation. EV-methods are indeed more time-consuming than A2C, partially because of PyTorch which is not made for parallel computing of the gradients: the larger K we want, the more time is needed. We are inclined to think that our implementation can be significantly optimized. The main complexity discussion with charts is placed in Supplementary.

3.3 Theoretical Guarantees

The main advantage of using empirical variance is that we have the machinery of statistical learning to prove the upper bounds for the variance of the gradient estimator. Our main theoretical result is concerned about one step of the update of θ_n with the best possible baseline chosen from the class of control variates. In this section we give some background and problem formulation and in the end discuss how the results are applied to our initial problem and give theoretical guarantees for EV-methods.

3.3.1 Variance Reduction

Classic problem for variance reduction is formulated for Monte Carlo estimation of some expectation. Let X be random variable and $X_1, .., X_K$ be a sample from the same distribution. Given a function $h : \mathbb{R} \to \mathbb{R}$ we want to evaluate $\mathbb{E}[h(X)] = \mathcal{E}$ using $\frac{1}{K}\sum_{k=1}^{K} h(X_k)$. This estimate, however, may possess large variance $V(h) := \mathbb{E}\left[(h(X) - \mathcal{E})^2\right]$, one could avoid that using other function h' such that $\mathbb{E}[h'(X)] = \mathcal{E}$ but $V(h') < V(h)$. Such estimate would be more reliable since there is less uncertainty.

This leads us to the following formulation. Given a class $\mathcal{H}$ of functions $h : \mathbb{R}^d \to \mathbb{R}$ find function $h_* \in \mathcal{H}$ with the least possible variance $V(h_*) \leq V(h) \quad \forall h \in \mathcal{H}$. Such problem is well-investigated in the literature and many methods have been suggested for variance reduction both for Monte Carlo and MCMC settings, for examples see [59, 67].

3.3.2 Variance Reduction in Multivariate Case

Let us consider scheme (3.1), the variance of the gradient estimate affects the convergence properties of the scheme so one is interested in reducing it, as can be seen in [90, 87] . We also provide a discussion about it in Supplementary. Note that now we are in setting different from the one above: it is needed to construct a vector estimate.

Let $X_1, ..., X_K \sim P$ be a sample of random vectors taking values in $\mathcal{X} \subset \mathbb{R}^d$ and let $\mathcal{H}$ be a class of functions $h : \mathbb{R}^d \to \mathbb{R}^D$ such that $\mathbb{E}[h(X)] = \mathcal{E}$. Later we will also need the corresponding empirical measure P_K based on $X_1, ..., X_K$. Define the variance

$$V(h) := \mathbb{E}[\|h(X) - \mathcal{E}\|^2]$$

with $\|\cdot\|$ being Euclidean 2-norm. Our goal is to find a function $h_* \in \mathcal{H}$ such that $V(h_*) \leq V(h)$ for all $h \in \mathcal{H}$. Then we have a variance reduced Monte Carlo estimate $\frac{1}{K}\sum_{k=1}^{K} h_*(X_k)$.

3.3.3 Variance Representation in Terms of Excess Risk

It is obvious that the exact solution h_* cannot be computed meaning that we are left always with some suboptimal solution $\hat{h} \in \mathcal{H}$ given by a particular method of ours. The quantity $V(\hat{h}) - V(h_*)$ where h_* is defined with

$$V(h_*) := \inf_{h \in \mathcal{H}} V(h)$$

is usually called *excess risk* in statistics and represents optimality gap, i.e. it shows how far the current solution h is from the optimal one. We can always write the variance of $\hat{h}$

as

$$V(\hat{h}) = \left[V(\hat{h}) - \inf_{h\in\mathcal{H}} V(h)\right] + \inf_{h\in\mathcal{H}} V(h) \tag{3.12}$$

from which we can clearly see the excess risk (the first term) and the second term representing approximation richness of the class $\mathcal{H}$: generally speaking, the better this class is, the lower the infimum can be. As more concrete example, consider variance reduction using the method of control variates. In this setting the goal is to estimate $\mathbb{E}[f(X)]$ for a fixed $f : \mathbb{R}^d \to \mathbb{R}^D$. To reduce the variance of the Monte Carlo sample mean one adds some *control variate* $g \in \mathcal{G}$ with zero expectation giving us the class of unbiased estimates

$$\mathcal{H} = \{f - g : \quad g \in \mathcal{G}, \quad \mathbb{E}[g(X)] = 0\}.$$

The excess risk is now of the form

$$\left[V(f - \hat{g}) - \inf_{g\in\mathcal{G}} V(f - g)\right],$$

sometimes called *stochastic error* and the second term

$$\inf_{g\in\mathcal{H}} V(f - g)$$

is known as *approximation error*. So, the way to analyze variance reduction is to estimate the excess risk and the approximation error.

In our analysis we consider a class of estimators with control variates implemented as baselines. Specifically, the class of estimators is

$$\mathcal{H} := \left\{\widetilde{\nabla}^{b_\phi} J \mid b_\phi \in \mathcal{B}_\Phi\right\}, \tag{3.13}$$

where

$$\widetilde{\nabla}^{b_\phi}_\theta J : X \mapsto \sum_{t=0}^{T-1} \gamma^t (G_t - b_\phi(S_t, A_t)) \nabla_\theta \log \pi(A_t|S_t)$$

and $b_\phi \in \mathcal{B}_\Phi$ is a map $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. The set of baselines $\mathcal{B}_\Phi$ is a parametric class parametrized by $\phi \in \Phi$. We require that for each $b_\phi \in \mathcal{B}_\Phi$ and for all policies $\pi \in \Pi$

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t b_\phi(S_t, A_t) \nabla_\theta \log \pi(A_t|S_t)\right] = 0,$$

requiring therefore that the estimator $\widetilde{\nabla}^{b_\phi} J$ is unbiased for all $b_\phi \in \mathcal{B}_\Phi$. For example, any set of maps $b_\phi : \mathcal{S} \to \mathbb{R}$ will satisfy the above condition leading to S-baselines.

We start with one-step analysis, showing how well the variance behaves when variance reduction with EV is applied at nth iteration. Let us further notate the estimator as $h : \mathbb{R}^d \to \mathbb{R}^D$ and note that $\mathbb{E}[h(X)] = \mathcal{E}$ with constant $\mathcal{E} = \nabla_\theta J$ since the estimate is assumed to be unbiased. In order to reduce the variance in the gradient estimator we would like to pick on each epoch n the best possible estimator

$$h^* = \arg\min_{h\in\mathcal{H}} V(h)$$

where variance functional V is defined for any $h \in \mathcal{H}$ via

$$V(h) := \mathbb{E}\left[\|h(X) - \mathcal{E}\|^2\right]$$

where X is random vector of concatenated states, actions and rewards described before. To solve the above optimization problem, we use empirical analogue of the variance and define

$$\hat{h} := \arg\min_{h \in \mathcal{H}} V_K(h)$$

with the empirical variance functional of the form:

$$V_K(h) := \frac{1}{K-1} \sum_{k=1}^{K} \|h(X^{(k)}) - P_K h\|^2$$

with P_K being the empirical measure, so with the given sample we could notate sample mean as

$$P_K h := \frac{1}{K} \sum_{k=1}^{K} h(X^{(k)}).$$

Let us pose several key assumptions.

A 14. Class $\mathcal{H}$ consists of bounded functions:

$$\sup_{x \in \mathcal{X}} \|h(x)\| \leq b, \quad \forall h \in \mathcal{H}.$$

A 15. The solution h_* is unique and $\mathcal{H}$ is star-shaped around h_*:

$$\alpha h + (1-\alpha) h_* \in \mathcal{H}, \quad \forall h \in \mathcal{H},\ \alpha \in [0,1].$$

A 16. The class $\mathcal{H}$ has covering of polynomial size: there are $\alpha \geq 2$ and $c > 0$ such that for all $u \in (0, b]$,

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(P_K)}, u) \leq \left(\frac{c}{u}\right)^{\alpha} \ a.s.$$

where

$$\|h\|_{L^2(P_K)} = \sqrt{P_K \|h\|_2^2}$$

The following result holds.

Theorem 24. *Under Assumptions 14-16 it holds with probability at least* $1 - 4e^{-t}$,

$$V(h_K) - V(h_*) \leq \max_{j=1,\ldots,4} \beta_j(t)$$

with

$$\beta_1 \leq C_1 \frac{\log K}{K},\ \beta_2 \leq C_2 \frac{\log K}{K},$$
$$\beta^3(t) = \frac{8(40b^2 t + 72b^2)}{3K}, \beta^4(t) = \frac{9216 b^2 t}{K},$$

where C_1, C_2 *are constants not depending on the dimension* D *or the sample size* K.

This allows to conclude from the variance decomposition (3.12) that as sample size K grows, the variance reduces to that of h_*. From practical perspective, Theorem 24 firstly gives some reliability guarantee. Secondly, it also shows that if we have K large enough, we can reduce the variance even more.

3.3.4 Verifying the Assumptions in Policy-Gradient Setting

Let us now discuss how we can satisfy the assumptions in our policy-gradient scheme (3.11).

As to Assumption 14, we can prove

Proposition 25. *(see Supplementary) If there exist constants $C_L > 0$ and $C_R > 0$ such that*

$$\forall \theta \in \Theta,\ a \in \mathcal{A}, s \in \mathcal{S} \quad \begin{aligned} &\|\nabla_\theta \log \pi(a|s)\| \leq C_L, \\ &|R(s,a)| \leq C_R, \end{aligned}$$

then Assumption 1 is satisfied.

In order to satisfy Assumption 15 in the context of policy gradient estimators $\mathcal{H}$ defined in (3.13), one might notice that

$$V(h_K) - \arg\min_{h \in \mathcal{H}} V(h) \leq V(h_K) - \arg\min_{h \in \mathrm{conv}(\mathcal{H})} V(h).$$

Indeed, Assumption 2 is a weaker notion than convexity.

Proposition 26. *(see Supplementary) If Assumption 3 holds for $\mathcal{B}_\phi$, under the conditions of Proposition 25 Assumption 3 holds also for $\mathcal{H}$ with other constants c, α.*

Let us also note that we could use the more realistic As. 16 stating the same for $\log \mathcal{N}$ (therefore considering more complex classes of baselines) and get weaker bounds with weaker rates, see [8].

3.3.5 Asymptotic Equivalence of EVv and EVm

Let us have a closer look on the variance functional with fixed baseline b_ϕ,

$$V(\tilde{\nabla}^{b_\phi} J(X_n)) = \mathbb{E}[\|\tilde{\nabla}^{b_\phi} J(X_n)\|^2] - \|\mathbb{E}[\tilde{\nabla}^{b_\phi} J(X_n)]\|^2.$$

Note that the right term equals $\|\nabla_\theta J\|^2$ since the estimate is unbiased. Therefore, if the gradient scheme converges to local optimum, i.e. $\theta_n \to \theta_*$ with $\nabla_{\theta_*} J = 0$ and the baseline parameters $\phi_n \to \phi_*$ as $n \to \infty$ a.s., then we can define the limiting variance as

$$V_\infty(\tilde{\nabla}^{b_{\phi_*}} J) = \mathbb{E}[\|\tilde{\nabla}^{b_{\phi_*}} J|_{\theta=\theta_*}\|^2]$$

which will strongly depend on the baseline we have chosen. This fact, firstly, implies that EVm and EVv algorithms are asymptotically equivalent because they differ in the second term converging to 0 and the first term is dominating by Jensen's inequality. Indeed, in our experiments we see that EVm and EVv behave similarly, so one would accept EVm as computationally cheaper version which works with $K \geq 1$. Secondly, EV-methods give additional stability guarantees for large n because they are directly related to the asymptotic gradient variance. It is an open question though to characterize the convergence of the presented two-timescale scheme to (θ_*, ϕ_*) more precisely.

3.3.6 Relation to A2C

Proposition 27. *(see Supplementary) If the conditions of Proposition 25 are satisfied, then for all $K \geq 2$ A2C goal function $V_K^{A2C}(\phi)$ is an upper bound (up to a constant) for EV goal functions:*

$$V_K^{EVm}(\phi) \leq 2C_L^2 V_K^{A2C}(\phi), \quad V_K^{EVv}(\phi) \leq 2C_L^2 V_K^{A2C}(\phi).$$

So, A2C is more computationally friendly method which exploits the upper bound on empirical variance for baseline training. This, in a sense, explains the success of A2C and different performance of A2C and EV-methods.

3.4 Experimental Results

We empirically investigate the behavior of EV-algorithms on several benchmark problems:

- Gym Minigrid [20] (`Unlock-v0, GoToDoor-5x5-v0`);
- Gym Classic Control [18] (`CartPole-v1, LunarLander-v2, Acrobot-v1`).

For each of these we provide charts with mean rewards illustrating the training process, the study of gradient variance and reward variance and time complexity discussions. Here because of small amount of space we present the most important results but the reader is welcome in the Supplementary materials where more experiments and investigations are presented together with all the implementation details. The code and config-files can be found on GitHub page [37].

3.4.1 Overview

Below we show the discussions about several key indicators of the algorithms.

1. **Mean rewards.** They are computed at each epoch based on the rewards obtained during the training in 40 runs and characterize how good is the algorithm in interaction with the environment.
2. **Standard deviation of the rewards.** These are computed in the same way but standard deviation is computed instead of mean. This values show how stable the training goes: high values indicate that there are frequent drops or increases in rewards.
3. **Gradient variance.** It is measured every 200 epochs using (3.9) with separate set of 50 sampled trajectories with relevant policy. This is the key indicator in the discussion of variance reduction. Surprisingly, as far as we know, we are the first in the RL community presenting such results for classic benchmarks. The resulting curves are averaged over 40 runs.

4. **Variance Reduction Ratio.** Together with Gradient Variance itself we also measure reduction ratio computed as sample variance of the estimator with baseline divided by the sample variance without baseline (assuming $b_\phi = 0$) in the computations of Gradient Variance. The reduction ratio is the main value of interest in variance reduction research in Monte Carlo and MCMC.

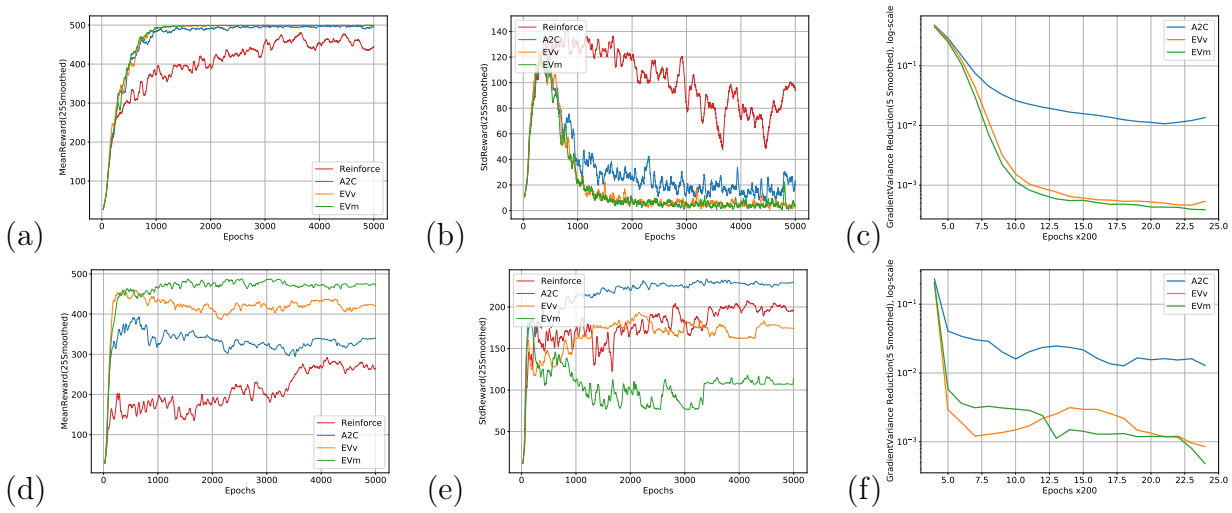


Figure 3.1: The charts representing the results for `CartPole` environment: (a,b,c) represent mean rewards, standard deviation of the rewards and gradient variance reduction ratio for config5 and (d,e,f) show the same information about config8.

3.4.2 Algorithm Performance

While observing mean rewards during the training we may notice immediately that EV-algorithms are at least as good as A2C.

In `CartPole` environment (Fig. 3.1) we conducted several experiments and present here two policy configurations: one with simpler neural network (config5, see Fig. 3.1(a,b,c)) and one with more complex network (config8, see Fig. 3.1(d,e,f)). In the first case both A2C and EV have very similar performance but in the second case the agent learns considerably faster with EV-based variance reduction and we get approximately 50% improvement over A2C agent and 75% over Reinforce agent in the end and even more during the training. The phenomenon of better performance of EV in CartPole with more complex policies is observed often, more detailed discussion is placed in Supplementary.

Experiments in `Acrobot` (see Fig. 3.2(a)) show that EV-algorithms can give better speed-up in the training. In the beginning EVm allows to learn faster but in the end the performance is the same as A2C. One of the reasons of such behavior can be the fact that learning rate becomes small and the agent already reaches the ceiling.

`Unlock` (Fig. 3.3(a)) is the example of the environments where all algorithms work similarly: in terms of rewards we cannot see significant improvement even over Reinforce. In `Unlock`, however, there is a difference presented but very small.

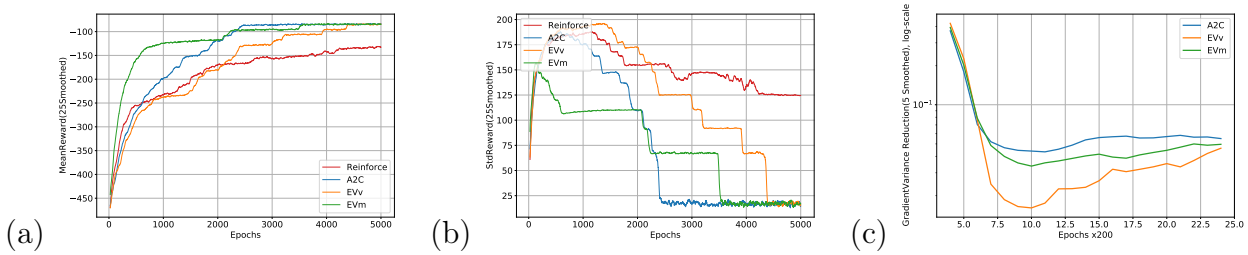


Figure 3.2: The charts representing the results for `Acrobot` environment: (a) depicts mean rewards, (b) shows the standard deviations of the rewards and (c) displays the gradient variance reduction ratios.

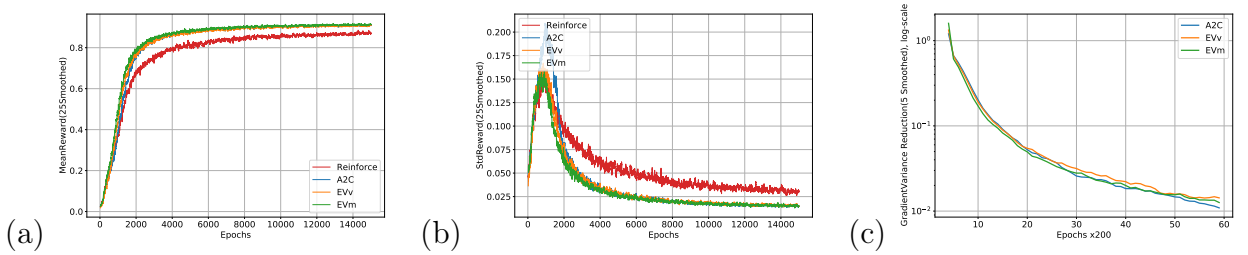


Figure 3.3: The charts representing the results for `Unlock` environment: (a) depicts mean rewards, (b) shows the standard deviations of the rewards and (c) displays the gradient variance reduction ratios.

3.4.3 Stability of Training

When we study the charts for standard deviation of the rewards (Fig. 3.1(b,e),3.2(b),3.3(b)), we can see that EV-methods are better in terms of stability of the training, the algorithm more rarely has drops than that of A2C. This is greatly illustrated by `CartPole` in Fig. 3.1(b,e) where the standard deviation is about 2 times less than in case of A2C. This holds for both configurations.

Fig. 3.2 illustrating the experiments with `Acrobot` show that until the ceiling is reached EV methods still can have lower variance. In `Unlock` presented in Fig. 3.3(b) we have not observed a significant difference in reward variance.

3.4.4 Gradient Variance and its Influence

The first thing we can notice reviewing the gradient variance is that A2C and EV reduce the variance similarly in `Unlock`. `CartPole` (see Fig. 3.1(c,f)), however, gives an example of the case where EV works completely differently to A2C, it reduces the variance almost 100-1000 times in both policy configurations. Similar picture we can observe in all `CartPole` experiments.

We can see that in `Unlock` showed in Fig. 3.3 the variance can also be reduced approximately 10-100 times, however, we see very little gain in rewards. It shows that in some environments training does not respond to the variance reduction; as a reason, it can be just not enough to give the improvement.

As answer to the discussion [81] about whether variance reduction helps in training we

would note the following phenomenon. In all cases we have confirmed variance reduction but not everywhere we have seen different performance of A2C, Reinforce and EV. Note that we designed our experiment in such a way that the only thing differentiating the agents is the goal function for baseline training. Some environments due to their specific setting and structure just do not respond to this variance reduction. In some cases (like in `Acrobot`) we can see that there are moments in training where variance reduction helps and where it does not change anything or even make training slower. It is natural to suppose that all these specific features should be addressed by some training algorithm which would combine in a clever way several variance reduction techniques or would decide that variance reduction is not needed at all. The last thing can be vital for exploration properties. Hence, we would conclude that variance reduction is the technique for improvement but how and when to apply it during the training is an interesting open question.

The last thing we would like to note is that reward variance measured in previous subsection is not an indicator of variance reduction since we have shown gradient variance reduction in all cases. Reward variance is decreased in relation to Reinforce, however, only in `CartPole` environement. Therefore, it cannot be used as a key metric for studying variance reduction in RL. The connection between reward variance and gradient variance seems to be an unanswered question in the literature.

3.5 Conclusions

In conclusion, we would like to state that sometimes the desired effect from variance reduction cannot be reached due to the specific nature of the environment. However, as we have seen above, it has the potential to influence the training process in a good way. As a new method for constructing variance reduction goals we suggested to use empirical variance which in turn resulted in EV-methods. Their motivation is more about actual variance reduction than in case of A2C and their performance is at least as good as A2C in terms of variance reduction and rewards. For them we also have suggested a probabilistic bound for the variance of the gradient estimate under some mild assumptions. Finally, EV-algorithms can be more stable in training which can allow to make sudden drops during the training less frequent. We also have for the first time presented the study of actual gradient variance reduction in classic benchmark problems. Our results have shown that variance reduction can help in the training but sometimes the environment's specific features do not allow to achieve gain in rewards. Therefore, variance reduction technique needs to be used during the training but the exact circumstances in which it helps are yet to be discovered.

Conclusion

In our work we have addressed two settings: optimal stopping for SDE and reinforcement learning in MDP setting.

Regarding the first direction we have presented the complexity analysis of WSM-algorithm and have suggested a new methodology for comparison of the algorithms for optimal stopping problem which includes the computational complexity. Our results have demonstrated superior qualities of the WSM algorithm and its robustness when trying to approximate the solution of continuous-time problem with a discrete-time one.

Regarding the second direction, we have contributed in several areas. Firstly, we have proven finite-time convergence analysis of linear stochastic approximation scheme which serves numerous policy evaluation algorithms. The analysis is shown to be tight by constructing an exact expansion of the error giving a lower bound. Secondly, in the last chapter we have designed a new method for variance reduction in policy-gradient algorithms based on empirical variance. The algorithm shows an inprovement over A2C least-squares criterion and can be used in various modifications of A2C schemes encorporating a variance reduction component.

Overall, our contribution demonstrated itself to be not in the sole direction but rather in the several areas: mathematical finance and reinforcement learning. Obtained ideas can be incorporated in many possible future research directions including stochastic algorithms and their analysis in stochastic optimal control and reinforcement learning.

References

[1] Ankush Agarwal and Sandeep Juneja. Comparing optimal convergence rate of stochastic mesh and least squares method for bermudan option pricing. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, WSC '13, page 701–712. IEEE Press, 2013.

[2] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[3] T. W. Archibald, K. I. M. McKinnon, and L. C. Thomas. On the generation of Markov decision processes. *The Journal of the Operational Research Society*, 46(3):354–361, 1995.

[4] Robert Azencott. Densité des diffusions en temps petit: développements asymptotiques. I. In *Seminar on probability, XVIII*, volume 1059 of *Lecture Notes in Math.*, pages 402–498. Springer, Berlin, 1984.

[5] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 30–37, 1995.

[6] Vlad Bally, Gilles Pagès, and Jacques Printems. A quantization tree method for pricing and hedging multidimensional American options. *Math. Finance*, 15(1):119–168, 2005.

[7] Vlad Bally, Gilles Pagès, and Jacques Printems. A quantization tree method for pricing and hedging multidimensional american options. *Mathematical Finance*, 15(1):119–168, 2005.

[8] Denis Belomestny, Leonid Iosipoi, Quentin Paris, and Nikita Zhivotovskiy. Empirical variance minimization with applications in variance reduction and optimal control. *Bernoulli*, 28(2):1382 – 1407, 2022.

[9] Denis Belomestny, Maxim Kaledin, and John Schoenmakers. Semitractability of optimal stopping problems via a weighted stochastic mesh algorithm. *Mathematical Finance*, 30(4):1591–1616, 2020.

[10] Michel Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, 33:1–68, 1999.

[11] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov,

Henrique Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. 12 2019.

[12] D. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific optimization and computation series. Athena Scientific, 2019.

[13] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

[14] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[15] Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[16] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[17] Mark Broadie and Paul Glasserman. A stochastic mesh method for pricing high-dimensional american options. *Journal of Computational Finance*, 7:35–72, 2004.

[18] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[19] Ching-An Cheng, Xinyan Yan, and Byron Boots. Trajectory-wise control variates for variance reduction in policy gradient methods. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1379–1394. PMLR, 30 Oct–01 Nov 2020.

[20] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[21] Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 21(52):1–51, 2020.

[22] E. Clément, D. Lamberton, and Philip Protter. An analysis of a least squares regression algorithm for american option pricing. *Finance and Stochastics*, 17, 01 2002.

[23] D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4):263–284, 1986.

[24] Gal Dalal, Balazs Szorenyi, and Gugan Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. *arXiv preprint arXiv:1911.09157*, 2019.

[25] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233, 2018.

[27] Thinh T Doan. Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation. *arXiv preprint arXiv:1912.10583*, 2019.

[28] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

[29] Daniel Egloff, Michael Kohler, and Nebojsa Todorovic. A dynamic look-ahead monte carlo algorithm for pricing bermudan options. *The Annals of Applied Probability*, 17(4):1138–1171, 2007.

[30] Yannis Flet-Berliac, reda ouhamma, odalric-ambrym maillard, and Philippe Preux. Learning value functions in deep policy gradients using residual variance. In *International Conference on Learning Representations*, 2021.

[31] Daniéle Florens-Zmirou. On estimating the diffusion coefficient from discrete observations. *J. Appl. Probab.*, 30(4):790–804, 1993.

[32] Vincent Francois, David Taralla, Damien Ernst, and Raphael Fonteneau. Deep reinforcement learning solutions for energy microgrids management. 12 2016.

[33] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354, 2018.

[34] Matthieu Geist and Bruno Scherrer. Off-policy learning with eligibility traces: A survey. *Journal of Machine Learning Research*, 15:289–333, 2014.

[35] Paul Glasserman. *Monte Carlo methods in financial engineering*. Springer, New York, 2004.

[36] David A. Goldberg and Yilun Chen. Beating the curse of dimensionality in options pricing and optimal stopping, 2018.

[37] Alexander Golubev and Maksim Kaledin. EVRLLib, a library implementing policy gradient algorithms in Reinforcement Learning. `https://github.com/DJAlexJ/EVRLLib`, June 2022.

[38] Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *J. Mach. Learn. Res.*, 5:1471–1530, December 2004.

[39] Shixiang Gu, Timothy P. Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[40] Harsh Gupta, R Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4706–4715, 2019.

[41] Patrick Jaillet, Damien Lamberton, and Bernard Lapeyre. Variational inequalities and the pricing of american options. *Acta Applicandae Mathematica*, 21(3):263–289, 1990.

[42] Maxim Kaledin, Alexander Golubev, and Denis Belomestny. Variance reduction for policy-gradient methods via empirical variance minimization. *arXiv:2206.06827v2*, 2022.

[43] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2144–2203. PMLR, 09–12 Jul 2020.

[44] Beom Kim, Yong-Ki Ma, and Hi Choe. A simple numerical method for pricing an american put option. *Journal of Applied Mathematics*, 2013, 01 2013.

[45] Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1008–1014. MIT Press, 2000.

[46] Vijay R. Konda and John N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *Ann. Appl. Probab.*, 14(2):796–819, 05 2004.

[47] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[48] 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, 2018.

[49] Tanguy Levent, Philippe Preux, Erwan Le Pennec, Jordi Badosa, Gonzague Henri, and Yvan Bonnassieux. Energy Management for Microgrids: a Reinforcement Learning Approach. In *ISGT-Europe 2019 - IEEE PES Innovative Smart Grid Technologies Europe*, pages 1–5, Bucharest, France, September 2019. IEEE.

[50] Chenxu Li. Maximum-likelihood estimation for diffusion processes via closed-form density expansions. *Ann. Statist.*, 41(3):1350–1380, 2013.

[51] Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *UAI*, pages 504–513, 2015.

[52] Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-dependent control variates for policy optimization via stein identity. In *International Conference on Learning Representations*, 2018.

[53] Francis Longstaff and Eduardo Schwartz. Valuing american options by simulation: A simple least-squares approach. *Review of Financial Studies*, 14:113–47, 02 2001.

[54] Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. In *International Conference on Learning Representations*, 2019.

[55] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.

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

[57] Abdelkader Mokkadem, Mariane Pelletier, et al. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671–1702, 2006.

[58] Erich Novak and Henryk Wo'zniakowski. *Tractability of multivariate problems. Vol. 1: Linear information*. 01 2008.

[59] Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

[60] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4026–4035, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[61] Alexander S. Poznyak. *Advanced Mathematical Tools for Automatic Control Engineers: Deterministic Techniques*. Elsevier, Oxford, 2008.

[62] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[63] John Rust. Using randomization to break the curse of dimensionality. *Econometrica*, 65(3):487–516, 1997.

[64] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

[65] Shijing Si, Chris. J. Oates, Andrew B. Duncan, Lawrence Carin, and François-Xavier Briol. Scalable control variates for monte carlo methods via stochastic optimization, 2021.

[66] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 10 2017.

[67] Leah F. South, Chris J. Oates, Antonietta Mira, and Christopher Drovandi. Regularised zero-variance control variates for high-dimensional variance reduction, 2020.

[68] R. Srikant and Lei Ying. Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning. In *Conference on Learning Theory*, 2019.

[69] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[70] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

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

[72] Richard S Sutton, Hamid Maei, and Csaba Szepesvári. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[73] 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*, ICML '09, page 993–1000, New York, NY, USA, 2009. Association for Computing Machinery.

[74] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[75] Vladislav Tadic. Almost sure convergence of two time-scale stochastic approximation algorithms. *Proceedings of the 2004 American Control Conference*, 4:3802–3807 vol.4, 2004.

[76] Vladislav Tadic. Asymptotic analysis of temporal-difference learning algorithms with constant step-sizes. *Machine Learning*, 63:107–133, 05 2006.

[77] Nizar Touzi. Optimal stochastic control, stochastic target problems, and backward sde. *Fields Institute Monographs*, 29, 01 2013.

[78] Lloyd Trefethen. Multivariate polynomial approximation in the hypercube. *Proceedings of the American Mathematical Society*, 145, 08 2016.

[79] 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.

[80] John Tsitsiklis and Benjamin Roy. Regression methods for pricing complex american style options. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 12:694–703, 02 2001.

[81] George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5015–5024. PMLR, 10–15 Jul 2018.

[82] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang,

Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, Nov 2019.

[83] L. Weaver and N. Tao. The optimal reward baseline for gradient·based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2001.

[84] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[85] Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations*, 2018.

[86] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 541–551, Tel Aviv, Israel, 22–25 Jul 2020. PMLR.

[87] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 541–551. PMLR, 22–25 Jul 2020.

[88] Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643, 2019.

[89] Daniel Zanger. Quantitative error estimates for a least-squares monte carlo algorithm for american option pricing. *Finance and Stochastics*, 17, 07 2013.

[90] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

Appendices

Appendix A

A.1 Proofs

A.1.1 Proof of Proposition 1

For achieving a target accuracy of order ε it is reasonable to divide the error equally over the variance and the bias part of (1.1). One thus chooses m such that $\kappa 5^L/m^\alpha \approx \varepsilon/2$, that is, $m \approx (2\kappa 5^L/\varepsilon)^{1/\alpha}$, and then takes N such that $\kappa 5^L m^{d/2}/N^{1/2} \approx \varepsilon/2$, i.e. $N \approx (2\kappa)^2 5^{2L} m^d/\varepsilon^2$, yielding a computational work load $\mathcal{C}_L(\varepsilon, d) = \kappa_1 N m^{2d} L$ as stated.

A.1.2 Proof of Proposition 2

For $l = L$ the statement reads

$$\int \left| U_L(x) - \widetilde{U}_L(x) \right| p_L(x|x_0)dx = \int 1_{|x-x_0|>R}\, g(x) p_L(x|x_0)dx = \varepsilon_{L,R},$$

so then it is true. Suppose (1.10) is true for $0 < l+1 \leq L$. Then, by using $|\max(a,b) - \max(a,c)| \leq |b-c|$ and the fact that $\widetilde{U}_l(x)$ vanishes for $|x - x_0| > R$,

$$\begin{aligned}
\left| U_l(x) - \widetilde{U}_l(x) \right| &\leq 1_{|x-x_0|\leq R} \left| \max\left[g(x), \mathbb{E}\left[U_{l+1}(X_{l+1}) \right| X_l = x \right] \right] \right. \\
&\quad \left. - \max\left[g(x), \mathbb{E}\left[\widetilde{U}_{l+1}(X_{l+1}) \right| X_l = x \right] \right] \right| + 1_{|x-x_0|>R} U_l(x) \\
&\leq 1_{|x-x_0|\leq R} \mathbb{E}\left[\left| U_{l+1}(X_{l+1}) - \widetilde{U}_{l+1}(X_{l+1}) \right| \right| X_l = x \right] + 1_{|x-x_0|>R} U_l(x).
\end{aligned}$$

Hence we have by induction,

$$\begin{aligned}
&\int \left| U_l(x) - \widetilde{U}_l(x) \right| p_l(x|x_0)dx \\
&\leq \int 1_{|x-x_0|>R} \mathbb{E}\left[\left| U_{l+1}(X_{l+1}) - \widetilde{U}_{l+1}(X_{l+1}) \right| \right| X_l = x \right] p_l(x|x_0)dx + \varepsilon_{l,R} \\
&\leq \int \left| U_{l+1}(y) - \widetilde{U}_{l+1}(y) \right| p_{l+1}(y|x_0)dy + \varepsilon_{l,R} \\
&= \sum_{j=l+1}^{L} \varepsilon_{j,R} + \varepsilon_{l,R} = \sum_{j=l}^{L} \varepsilon_{j,R}.
\end{aligned}$$

A.1.3 Proof of Proposition 3

Combining the assumptions (1.11) and (1.12) yields

$$\begin{aligned} U_l(x) &= \operatorname*{esssup}_{\tau\in\mathcal{T}_{l,L}} \mathbb{E}\left[g_\tau(Z_\tau)\middle|\, Z_l = x\right] \\ &\le c_g \mathbb{E}\left[1 + \max_{l\le l'\le L} |Z_{l'}|\middle|\, Z_l = x\right] \\ &\le c_g\left(1 + c_Z\right) + c_g c_Z \left|x\right|. \end{aligned}$$

By the estimate

$$\int_{|x-x_0|>R} e^{-\frac{|x-x_0|^2}{2\alpha l}} dx \le e^{-\frac{R^2}{8\alpha l}} \left(4/3\right)^{d/2} \left(2\pi\alpha l\right)^{d/2},$$

and (using Cauchy-Schwarz) the estimate

$$\begin{aligned} \int_{|x-x_0|>R} |x-x_0|\, e^{-\frac{|x-x_0|^2}{2\alpha l}} dx &\le \sqrt{\int_{|x-x_0|>R} e^{-\frac{|x-x_0|^2}{2\alpha l}} dx}\sqrt{\int |x-x_0|^2\, e^{-\frac{|x-x_0|^2}{2\alpha l}} dx} \\ &\le e^{-\frac{R^2}{8\alpha l}} 2^{d/4} (2\pi\alpha l)^{d/2}\sqrt{d\alpha l}, \end{aligned}$$

we get (note that $(4/3)^{1/2} < 2^{1/4}$)

$$\begin{aligned} \varepsilon_{l,R} &\le \frac{\varkappa}{(2\pi\alpha l)^{d/2}} \int_{|x-x_0|>R} \left(c_g\left(1 + c_Z\right) + c_g c_Z \left|x\right|\right) e^{-\frac{|x-x_0|^2}{2\alpha l}} dx \\ &\le \frac{\varkappa c_g\left(1 + c_Z + c_Z\left|x_0\right|\right)}{(2\pi\alpha l)^{d/2}} \int_{|x-x_0|>R} e^{-\frac{|x-x_0|^2}{2\alpha l}} dx \\ &+ \frac{\varkappa c_g c_Z}{(2\pi\alpha l)^{d/2}} \int_{|x-x_0|>R} |x-x_0|\, e^{-\frac{|x-x_0|^2}{2\alpha l}} dx \\ &\le \varkappa c_g \left(1 + c_Z + c_Z\left|x_0\right| + c_Z\sqrt{d\alpha}\sqrt{l}\right) 2^{d/4} e^{-\frac{R^2}{8\alpha l}} \\ &\equiv \left(A + B\sqrt{l}\right) c_g \varkappa e^{-\frac{R^2}{8\alpha l}}, \end{aligned}$$

for $l \ge 1$ ($\varepsilon_{0,R} = 0$ for $R > 0$). Now by (1.10), i.e. Proposition 2, we get

$$\int \left|U_l(x) - \widetilde{U}_l(x)\right| p_l(x|x_0)\, dx \le L\left(A + B\sqrt{L}\right) c_g \varkappa e^{-\frac{R^2}{8\alpha L}},$$

whence the estimate (1.14).

A.1.4 Proof of Proposition 4

Let us write the sample based backward dynamic program (1.9) for step $l < L$ in the form

$$\overline{U}_l\left(Z_l^{(i)}\right) = \mathbb{1}_{\left|Z_l^{(i)} - x_0\right| \le R} \max\left[g_l(Z_l^{(i)}), \sum_{j=1}^{N} \overline{U}_{l+1}(Z_{l+1}^{(j)})\overline{w}_{ij}\right] \tag{A.1}$$

by defining the weights

$$w_{ij} := \frac{p(Z_{l+1}^{(j)}|Z_l^{(i)})}{\sum_{m=1}^{N} p(Z_{l+1}^{(j)}|Z_l^{(m)})}, \tag{A.2}$$

where l is fixed and suppressed. Let us further abbreviate

$$\mathcal{E}[f](x) = \mathbb{E}\left[f(Z_{l+1})|\, Z_l = x\right] = \int f(y)p(y|x)dy$$

for a generic Borel function $f \geq 0$. Using

$$\widetilde{U}_l\left(Z_l^{(i)}\right) = \mathbb{1}_{\left|Z_l^{(i)}-x_0\right|\leq R} \max\left[g_l(Z_l^{(i)}), \mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(i)})\right],$$

(A.1) and $|\max(a,b) - \max(a,c)| \leq |b-c|$, we thus get

$$\begin{aligned}
&\left|\overline{U}_l - \widetilde{U}_l\right|_N := \frac{1}{N}\sum_{i=1}^N \left|\overline{U}_l(Z_l^{(i)}) - \widetilde{U}_l(Z_l^{(i)})\right| \leq \\
&\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\left|Z_l^{(i)}-x_0\right|\leq R} \left|\sum_{j=1}^N \overline{U}_{l+1}(Z_{l+1}^{(j)})w_{ij} - \mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(i)})\right| \\
&\leq \left|\overline{U}_{l+1} - \widetilde{U}_{l+1}\right|_N + \mathcal{R}_{l+1} \qquad \text{(A.3)}
\end{aligned}$$

with

$$\mathcal{R}_{l+1} = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\left|Z_l^{(i)}-x_0\right|\leq R} \left|\sum_{j=1}^N \widetilde{U}_{l+1}(Z_{l+1}^{(j)})w_{ij} - \mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(i)})\right|,$$

where we have used the fact that the weights in (A.2) sum up to one. One thus gets by iterating (A.3)

$$\left|\overline{U}_k - \widetilde{U}_k\right|_N \leq \sum_{l=k}^{L-1} \mathcal{R}_{l+1}, \qquad \text{(A.4)}$$

since $\overline{U}_L - \widetilde{U}_L = 0$. Let us now introduce

$$w_{ij}^\circ := \frac{1}{N}\frac{p(Z_{l+1}^{(j)}|Z_l^{(i)})}{p_{l+1}(Z_{l+1}^{(j)}|x_0)}, \qquad \text{(A.5)}$$

and consider the generic term

$$\begin{aligned}
\mathcal{R}_{l+1} &= \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\left|Z_l^{(i)}-x_0\right|\leq R} \left|\sum_{j=1}^N \widetilde{U}_{l+1}(Z_{l+1}^{(j)})w_{ij} - \mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(i)})\right| \\
&\leq \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\left|Z_l^{(i)}-x_0\right|\leq R} \sum_{j=1}^N \widetilde{U}_{l+1}(Z_{l+1}^{(j)}) \left|w_{ij} - w_{ij}^\circ\right| \\
&+ \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\left|Z_l^{(i)}-x_0\right|\leq R} \left|\sum_{j=1}^N \left(w_{ij}^\circ \widetilde{U}_{l+1}(Z_{l+1}^{(j)}) - \frac{1}{N}\mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(i)})\right)\right| \\
&=: \mathrm{Term}_1 + \mathrm{Term}_2.
\end{aligned}$$

We have

$$\mathbb{E}\left[\mathcal{R}_{l+1}\right] \leq \mathbb{E}\left[\mathrm{Term}_1\right] + \mathbb{E}\left[\mathrm{Term}_2\right].$$

While the first term Term_1 is small as (w_{ij}) are close to (w°_{ij}), the second one Term_2 tends to 0 as $N \to \infty$ by the law of large numbers. Indeed, due to (1.7) one has

$$\mathrm{Term}_1 \leq \frac{G_R}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{1}_{\left|Z_l^{(i)} - x_0\right| \leq R} \mathbb{1}_{\left|Z_{l+1}^{(j)} - x_0\right| \leq R} \left| w_{ij} - w^\circ_{ij} \right|,$$

and due to (A.2) and (A.5) we may write

$$\begin{aligned} \left| w_{ij} - w^\circ_{ij} \right| &= \left| \frac{p(Z_{l+1}^{(j)} | Z_l^{(i)})}{\sum_{m=1}^{N} p(Z_{l+1}^{(j)} | Z_l^{(m)})} - \frac{1}{N} \frac{p(Z_{l+1}^{(j)} | Z_l^{(i)})}{p_{l+1}(Z_{l+1}^{(j)} | x_0)} \right| \\ &= \frac{p(Z_{l+1}^{(j)} | Z_l^{(i)})}{\sum_{m=1}^{N} p(Z_{l+1}^{(j)} | Z_l^{(m)})} \left| 1 - \frac{\frac{1}{N} \sum_{m=1}^{N} p(Z_{l+1}^{(j)} | Z_l^{(m)})}{p_{l+1}(Z_{l+1}^{(j)} | x_0)} \right| \end{aligned}$$

to obtain

$$\mathrm{Term}_1 \leq \frac{G_R}{N} \sum_{j=1}^{N} \mathbb{1}_{\left|Z_{l+1}^{(j)} - x_0\right| \leq R} \left| 1 - \frac{\frac{1}{N} \sum_{m=1}^{N} p(Z_{l+1}^{(j)} | Z_l^{(m)})}{p_{l+1}(Z_{l+1}^{(j)} | x_0)} \right|.$$

The expectation of the random variable inside of the above sum is independent of j. So by taking $j = 1$ and splitting off the with $Z^{(1)}$ correlating term due to $m = 1$, one gets

$$\begin{aligned} \mathbb{E}\left[\mathrm{Term}_1\right] &\leq \frac{G_R}{N} \mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{(1)} - x_0\right| \leq R} \left| \sum_{m=1}^{N} \left(1 - \frac{p(Z_{l+1}^{(1)} | Z_l^{(m)})}{p_{l+1}(Z_{l+1}^{(1)} | x_0)} \right) \right| \right] \\ &\leq \frac{G_R}{N} D_{R,l} + \frac{G_R}{N} \mathbb{E}\left[\left| \sum_{m=2}^{N} \mathbb{1}_{\left|Z_{l+1}^{(1)} - x_0\right| \leq R} \left(1 - \frac{p(Z_{l+1}^{(1)} | Z_l^{(m)})}{p_{l+1}(Z_{l+1}^{(1)} | x_0)} \right) \right| \right] \end{aligned}$$

with

$$D_{R,l} := \mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{(1)} - x_0\right| \leq R} \left| 1 - \frac{p(Z_{l+1}^{(1)} | Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{(1)} | x_0)} \right| \right].$$

Now consider the i.i.d. random variables,

$$\eta_m^{(l+1)} := \mathbb{1}_{\left|Z_{l+1}^{(1)} - x_0\right| \leq R} \left(1 - \frac{p(Z_{l+1}^{(1)} | Z_l^{(m)})}{p_{l+1}(Z_{l+1}^{(1)} | x_0)} \right), \quad m = 2, ..., N. \tag{A.6}$$

It can be verified by conditioning on $Z_{l+1}^{(1)}$ that these have zero mean. Then by applying Jensen's inequality to the square-root, using the independence of the random variables (A.6), and that the latter variables are identically distributed with zero mean, we derive

$$\mathbb{E}\left| \sum_{m=2}^{N} \eta_m^{(l+1)} \right| \leq \sqrt{\mathbb{E}\left(\sum_{m=2}^{N} \eta_m^{(l+1)} \right)^2} = E_{R,l} \sqrt{N}$$

with

$$E_{R,l}^2 := \mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{(1)} - x_0\right| \leq R} \left| 1 - \frac{p(Z_{l+1}^{(1)} | Z_l^{(2)})}{p_{l+1}(Z_{l+1}^{(1)} | x_0)} \right|^2 \right].$$

Finally we get for Term_1,

$$\mathbb{E}[\mathrm{Term}_1] \leq \frac{G_R D_{R,l}}{N} + \frac{G_R E_{R,l}}{\sqrt{N}}.$$

Concerning Term_2, let us write

$$\begin{aligned}\mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(i)}) &= \int \widetilde{U}_{l+1}(y)\frac{p(y|Z_l^{(i)})}{p_{l+1}(y|x_0)}p_{l+1}(y|x_0)dy \\ &= \mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}^{0,x_0})\frac{p(Z_{l+1}^{0,x_0}|Z_l^{(i)})}{p_{l+1}(Z_{l+1}^{0,x_0}|x_0)}\right],\end{aligned}$$

where Z^{0,x_0} is an independent dummy trajectory. We thus have

$$\begin{aligned}\mathbb{E}\left[\text{Term}_2\right] \leq \mathbb{E}\left[\mathbb{1}_{\left|Z_l^{(1)}-x_0\right|\leq R}\left|\left(w_{11}^{\circ}\widetilde{U}_{l+1}(Z_{l+1}^{(1)}) - \frac{1}{N}\mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(1)})\right)\right|\right] \\ + \mathbb{E}\left[\left|\sum_{j=2}^{N}\zeta_j^{(l+1)}\right|\right],\end{aligned}$$

where for $j = 2, ..., N$, the random variables

$$\begin{aligned}\zeta_j^{(l+1)} &:= \mathbb{1}_{\left|Z_l^{(1)}-x_0\right|\leq R}\left(w_{1j}^{\circ}\widetilde{U}_{l+1}(Z_{l+1}^{(j)}) - \frac{1}{N}\mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(1)})\right) \\ &= \frac{\mathbb{1}_{\left|Z_l^{(1)}-x_0\right|\leq R}}{N}\left(\frac{p(Z_{l+1}^{(j)}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{(j)}|x_0)}\widetilde{U}_{l+1}(Z_{l+1}^{(j)}) - \mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}^{0,x_0})\frac{p(Z_{l+1}^{0,x_0}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{0,x_0}|x_0)}\right]\right)\end{aligned}$$

are i.i.d. with zero mean. We so have by the Jensen's inequality again,

$$\mathbb{E}\left[\left|\sum_{j=2}^{N}\zeta_j^{(l+1)}\right|\right] \leq \sqrt{N\mathsf{Var}\left(\zeta_2^{(l+1)}\right)} \leq F_{R,l}G_R/\sqrt{N},$$

where

$$F_{R,l}^2 = \mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{(2)}-x_0\right|\leq R}\left|\frac{p(Z_{l+1}^{(2)}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{(2)}|x_0)}\right|^2\right] = \int\int_{|y-x_0|\leq R}\frac{p^2(y|x)}{p_{l+1}(y|x_0)}p_l(x|x_0)\,dxdy.$$

Secondly, by (A.5) one has

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\left|Z_l^{(1)}-x_0\right|\leq R}\left|\left(w_{11}^{\circ}\widetilde{U}_{l+1}(Z_{l+1}^{(1)}) - \frac{1}{N}\mathcal{E}[\widetilde{U}_{l+1}](Z_l^{(1)})\right)\right|\right] \\ &\leq \frac{1}{N}\mathbb{E}\left[\mathbb{1}_{\left|Z_l^{(1)}-x_0\right|\leq R}\frac{p(Z_{l+1}^{(1)}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{(1)}|x_0)}\widetilde{U}_{l+1}(Z_{l+1}^{(1)})\right] \\ &+ \frac{1}{N}\mathbb{E}\left[\mathbb{1}_{\left|Z_l^{(1)}-x_0\right|\leq R}\mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}^{0,x_0})\frac{p(Z_{l+1}^{0,x_0}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{0,x_0}|x_0)}\right]\right] \\ &\leq \frac{G_R}{N}\mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{(1)}-x_0\right|\leq R}\frac{p(Z_{l+1}^{(1)}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{(1)}|x_0)}\right] \\ &+ \frac{G_R}{N}\mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{0,x_0}-x_0\right|\leq R}\frac{p(Z_{l+1}^{0,x_0}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{0,x_0}|x_0)}\right] =: \frac{G_R}{N}H_{R,l},\end{aligned}$$

where the latter inequality follows from (1.7) and the fact that $\widetilde{U}_{l+1}$ vanishes outside the ball B_R.

Combining the above estimates, we get for Term_2

$$\mathbb{E}[\mathrm{Term}_2] \leq \frac{F_{R,l}G_R}{\sqrt{N}} + \frac{G_R}{N}H_{R,l}.$$

Thus we have expressed our bounds for $\mathbb{E}[\mathrm{Term}_1]$ and $\mathbb{E}[\mathrm{Term}_2]$ in terms of the quantities $D_{R,l}$, $E_{R,l}$, $F_{R,l}$, $H_{R,l}$, and G_R. Furthermore, it is easy to see that using (1.15)

$$\begin{aligned} D_{R,l} &\leq 1 + \mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{(1)}-x_0\right|\leq R}\frac{p(Z_{l+1}^{(1)}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{(1)}|x_0)}\right] \\ &= 1 + \int p_l(x|x_0)\,dx \int_{|y-x_0|\leq R}\frac{p^2(y|x)}{p_{l+1}(y|x_0)}dy \\ &\leq 1 + F_R^2. \end{aligned}$$

Similarly, it follows that $E_{R,l}^2 \leq 2 + 2F_R^2$, and that $H_{R,l} \leq 1 + F_R^2$ due to

$$\mathbb{E}\left[\mathbb{1}_{\left|Z_{l+1}^{0,x_0}-x_0\right|\leq R}\frac{p(Z_{l+1}^{0,x_0}|Z_l^{(1)})}{p_{l+1}(Z_{l+1}^{0,x_0}|x_0)}\right] \leq 1.$$

By now taking the expectation in (A.4) and gathering all together we obtain,

$$\mathbb{E}\left[\left|\overline{U}_k - \widetilde{U}_k\right|_N\right] \leq (L-k)G_R\left(\frac{\sqrt{2+2F_R^2}+F_R}{\sqrt{N}} + \frac{2+2F_R^2}{N}\right). \tag{A.7}$$

By next taking $k = 0$ and assuming that N is taken such that $(1 + F_R)/\sqrt{N} < 1$, Proposition 4 follows.

A.1.5 Proof of Proposition 6

In order to achieve a required accuracy $\varepsilon > 0$, let us take R and N large enough such that both error terms in (1.16) are equal to $\varepsilon/2$. Hence, we first take

$$R_{\varepsilon,d} = (8\alpha L)^{1/2}\log^{1/2}\frac{Lc_g\varkappa\left(1 + c_Z + c_Z\left|x_0\right| + c_Z\sqrt{d\alpha L}\right)2^{1+d/4}}{\varepsilon},$$

that is $R \nearrow \infty$ when $d + \varepsilon^{-1} \nearrow \infty$. Then take, with $\asymp$ denoting asymptotic equivalence for $R \nearrow \infty$ up to some natural constant,

$$\begin{aligned} N_\varepsilon \asymp L^2c_g^2\varkappa\,(e/\alpha)^{d/2}\,d^{-d/2}R_\varepsilon^{d+2}\varepsilon^{-2} &\asymp \alpha c_g^2\varkappa\,(8e/d)^{d/2}\,L^{d/2+3} \\ &\times \varepsilon^{-2}\log^{d/2+1}\frac{L\left(1 + c_Z + c_Z\left|x_0\right| + c_Z\sqrt{d\alpha L}\right)2^{1+d/4}c_g\varkappa}{\varepsilon}. \end{aligned}$$

Thus, the computational work load (complexity) is given by

$$\begin{aligned} c_f^{(d)}N_\varepsilon^2 L \leq c_1\alpha^2c_g^4\varkappa^2c_f^{(d)}\,(8e/d)^d\,L^{d+7} \\ \times \varepsilon^{-4}\log^{d+2}\frac{L\left(1 + c_Z + c_Z\left|x_0\right| + c_Z\sqrt{d\alpha L}\right)2^{1+d/4}c_g\varkappa}{\varepsilon} \end{aligned} \tag{A.8}$$

where c_1 is a natural constant. Now let us write

$$d^{-d}\log^{d+2}\frac{L\left(1+c_Z+c_Z\left|x_0\right|+c_Z\sqrt{d\alpha L}\right)2^{1+d/4}c_g\varkappa}{\varepsilon}$$
$$=d^2\log^{d+2}\left[\frac{L^{1/d}\left(1+c_Z+c_Z\left|x_0\right|+c_Z\sqrt{d\alpha L}\right)^{1/d}2^{1/d+1/4}\left(c_g\varkappa\right)^{1/d}}{\varepsilon^{1/d}}\right].$$

Then, using the elementary estimate $\left(a+b\sqrt{d}\right)^{1/d}\leq ae^{b/a}$, for $a,b>0$, $d\geq 1$, and assuming that $\varepsilon<1$, (A.8) implies (1.17).

A.1.6 Proof of Proposition 9

On the one hand one has

$$\begin{aligned}U_{t_l}^{\circ}(X_{t_l})-U_{t_l}(\overline{X}_{t_l})&=\operatorname*{esssup}_{\tau\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[g(\tau,X_\tau)\right]-\operatorname*{esssup}_{\overline{\tau}\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[g(\overline{\tau},\overline{X}_{\overline{\tau}})\right]\\&\leq\operatorname*{esssup}_{\tau\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[g(\tau,X_\tau)-g(\tau,\overline{X}_\tau)\right]\\&\leq\operatorname*{esssup}_{\tau\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[\left|g(\tau,X_\tau)-g(\tau,\overline{X}_\tau)\right|\right],\end{aligned}$$

and on the other one has similarly

$$\begin{aligned}U_{t_l}(\overline{X}_{t_l})-U_{t_l}^{\circ}(X_{t_l})&=\operatorname*{esssup}_{\overline{\tau}\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[g(\overline{\tau},\overline{X}_{\overline{\tau}})\right]-\operatorname*{esssup}_{\tau\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[g(\tau,X_\tau)\right]\\&\leq\operatorname*{esssup}_{\overline{\tau}\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[g(\overline{\tau},\overline{X}_{\overline{\tau}})-g(\tau,X_\tau)\right]\\&\leq\operatorname*{esssup}_{\tau\in\mathcal{T}_{l,L}}\mathbb{E}_{\mathcal{F}_{t_l}}\left[\left|g(\tau,X_\tau)-g(\tau,\overline{X}_\tau)\right|\right].\end{aligned}$$

Hence we get

$$\begin{aligned}\mathbb{E}\left[\left|U_{t_l}^{\circ}(X_{t_l})-U_{t_l}(\overline{X}_{t_l})\right|\right]&\leq\mathbb{E}\left[\sup_{0\leq s\leq T}\left|g(s,X_s)-g(s,\overline{X}_s)\right|\right]\\&\leq L_g\mathbb{E}\left[\sup_{0\leq s\leq T}\left|X_s-\overline{X}_s\right|\right]\leq C^{\text{Euler}}\sqrt{h},\end{aligned}$$

due to the strong order of the Euler scheme, with L_g being some Lipschitz constant for g.

Appendix B

B.1 Proof of Proposition 14

The following derivation is largely borrowed from [46] and is repeated here for completeness. We begin by substituting $\tilde{\theta}_k$ into (2.3) to obtain

$$\begin{aligned}\tilde{\theta}_{k+1} &= (\mathrm{I}-\beta_k A_{11})\theta_k - \beta_k A_{12} w_k - \theta^\star + \beta_k b_1 - \beta_k V_{k+1} \\ &= (\mathrm{I}-\beta_k A_{11})\tilde{\theta}_k - \beta_k A_{11}\theta^\star - \beta_k A_{12}(\tilde{w}_k + w^\star - C_{k-1}\tilde{\theta}_k) + \beta_k b_1 - \beta_k V_{k+1} \\ &= (\mathrm{I}-\beta_k(A_{11} - A_{12}A_{22}^{-1}A_{21} - A_{12}L_k))\tilde{\theta}_k - \beta_k A_{12}\tilde{w}_k - \beta_k(A_{12}w^\star + A_{11}\theta^\star - b_1) - \beta_k V_{k+1}.\end{aligned}$$

Notice that

$$A_{12}w^\star + A_{11}\theta^\star - b_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})\theta^\star + A_{12}A_{22}^{-1}b_2 - b_1 = 0.$$

The above yields

$$\tilde{\theta}_{k+1} = (\mathrm{I}-\beta_k B_{11}^k)\tilde{\theta}_k - \beta_k A_{12}\tilde{w}_k - \beta_k V_{k+1}. \tag{B.1}$$

Next, we observe that

$$\begin{aligned}w_{k+1} - w^\star &= (\mathrm{I}-\gamma_k A_{22})w_k - \gamma_k A_{21}\theta_k - w^\star + \gamma_k b_2 - \gamma_k W_{k+1} \\ &= (\mathrm{I}-\gamma_k A_{22})(w_k - w^\star) - \gamma_k A_{22}w^\star - \gamma_k A_{21}\theta_k + \gamma_k b_2 - \gamma_k W_{k+1} \\ &= (\mathrm{I}-\gamma_k A_{22})(w_k - w^\star) - \gamma_k A_{21}(\theta_k - \theta^\star) - \gamma_k W_{k+1}\end{aligned}$$

Substitute $\tilde{w}_k$ into (2.4) and using (2.23) yield:

$$\begin{aligned}&\tilde{w}_{k+1} = (\mathrm{I}-\gamma_k A_{22})(w_k - w^\star) - \gamma_k A_{21}\tilde{\theta}_k + C_k\tilde{\theta}_{k+1} - \gamma_k W_{k+1} \\ &= (\mathrm{I}-\gamma_k A_{22})\tilde{w}_k - \big((\mathrm{I}-\gamma_k A_{22})C_{k-1} + \gamma_k A_{21}\big)\tilde{\theta}_k + C_k\big((\mathrm{I}-\beta_k B_{11}^k)\tilde{\theta}_k - \beta_k A_{12}\tilde{w}_k\big) \\ &\quad - \beta_k C_k V_{k+1} - \gamma_k W_{k+1} \\ &= (\mathrm{I}-\gamma_k B_{22}^k)\tilde{w}_k - \Big(C_{k-1} - \gamma_k(A_{22}C_{k-1} - A_{21}) - C_k(\mathrm{I}-\beta_k B_{11}^k)\Big)\tilde{\theta}_k - \beta_k C_k V_{k+1} - \gamma_k W_{k+1}\end{aligned}$$

We observe that

$$\begin{aligned}&C_{k-1} - \gamma_k(A_{22}C_{k-1} - A_{21}) - C_k(\mathrm{I}-\beta_k B_{11}^k) \\ &= L_k + A_{22}^{-1}A_{21} - (L_{k+1} + A_{22}^{-1}A_{21})(\mathrm{I}-\beta_k B_{11}^k) - \gamma_k(A_{22}C_{k-1} - A_{21}) \\ &= L_k - (L_k - \gamma_k A_{22}L_k + \beta_k A_{22}^{-1}A_{21}B_{11}^k) - \beta_k A_{22}^{-1}A_{21}B_{11}^k - \gamma_k(A_{22}C_{k-1} - A_{21}) \\ &= \gamma_k A_{22}L_k - \gamma_k(A_{22}(L_k + A_{22}^{-1}A_{21}) - A_{21}) = 0.\end{aligned}$$

The above yields

$$\tilde{w}_{k+1} = (\mathrm{I}-\gamma_k B_{22}^k)\tilde{w}_k - \beta_k C_k V_{k+1} - \gamma_k W_{k+1}. \tag{B.2}$$

B.2 Detailed Proofs for Section 2.3

Before we proceed to proving the main results of Section 2.3, we first study a few properties of the two timescale linear SA scheme.

To facilitate our discussions next, we define the constant:

$$\mathrm{C}_\infty := \sqrt{\lambda_{\mathsf{min}}(Q_\Delta)^{-1}\lambda_{\mathsf{max}}(Q_{22})}\,\mathrm{L}_\infty + \|A_{22}^{-1}A_{21}\|, \tag{B.3}$$

where $\|C_k\| \le \mathrm{C}_\infty$ for any $k \ge 0$. Then, as we have $\theta_k\theta_k^\top \preceq 2\tilde{\theta}_k\tilde{\theta}_k^\top + 2\theta^\star(\theta^\star)^\top$, it holds

$$\|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| \le 2\big\{\mathrm{M}_k^{\tilde{\theta}} + \|\theta^\star(\theta^\star)^\top\|\big\}, \quad \|\mathbb{E}\left[w_k w_k^\top\right]\| \le 3\big\{\mathrm{M}_k^{\tilde{w}} + \mathrm{M}_k^{\tilde{\theta}}\,\mathrm{C}_\infty^2 + \|w^\star(w^\star)^\top\|\big\}$$

The noise terms V_k, W_k can then be estimated in terms of the transformed variables $\tilde{\theta}_k, \tilde{w}_k$ and their variances $\mathrm{M}_k^{\tilde{\theta}}, \mathrm{M}_k^{\tilde{w}}$. In particular, combining with A12 yields

$$\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\| \le \tilde{m}_V(1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}), \quad \|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| \le \tilde{m}_W(1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}) \tag{B.4}$$

$$\|\mathbb{E}\left[V_{k+1}W_{k+1}^\top\right]\| \le \tilde{m}_{VW}(1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}) \tag{B.5}$$

where

$$\begin{aligned}
&\frac{\tilde{m}_V}{m_V} = \frac{\tilde{m}_W}{m_W} = (1 + 2\|\theta^\star(\theta^\star)^\top\| + 3\|w^\star(w^\star)^\top\|) \vee (2 + 3\,\mathrm{C}_\infty^2) \vee 3 \\
&\tilde{m}_{VW} = \frac{\sqrt{d_\theta d_w}}{2}\,(\tilde{m}_W + \tilde{m}_V)
\end{aligned} \tag{B.6}$$

We also define a few constants related to the matrices Q_Δ, Q_{22} associated with the Hurwitz matrices Δ, A_{22} in (2.8). Set $p_\Delta := \lambda_{\mathsf{min}}^{-1}(Q_\Delta)\lambda_{\mathsf{max}}(Q_\Delta)$, $p_{22} := \lambda_{\mathsf{min}}^{-1}(Q_{22})\lambda_{\mathsf{max}}(Q_{22})$, $p_{22,\Delta} := \sqrt{p_{22}p_\Delta}$. Moreover, for any $a > 0$, we set

$$\varrho^a := \frac{2}{a}\varsigma\max\{1, a_{22}/(4a_\Delta)\} \vee \frac{4}{a}\,(\varsigma)^3. \tag{B.7}$$

Next, we study the contraction properties of $\mathrm{I} - \beta_k B_{11}^k$ and $\mathrm{I} - \gamma_k B_{22}^k$ that appear in the transformed two timescale SA (2.23),(2.24). Using (2.9), we observe that

$$\begin{aligned}
\|\mathrm{I} - \beta_k B_{11}^k\|_{Q_\Delta} = \|\mathrm{I} - \beta_k\Delta + \beta_k A_{12}L_k\|_{Q_\Delta} &\le \|\mathrm{I} - \beta_k\Delta\|_{Q_\Delta} + \beta_k\|A_{12}\|_{Q_{22},Q_\Delta}\|L_k\|_{Q_\Delta,Q_{22}} \\
&\le (1 - \beta_k a_\Delta) + \beta_k\|A_{12}\|_{Q_{22},Q_\Delta}\|L_k\|_{Q_\Delta,Q_{22}}\,.
\end{aligned}$$

Recalling that $\|L_k\|_{Q_\Delta,Q_{22}} \le \mathrm{L}_\infty$, the above inequality yields

$$\|\mathrm{I} - \beta_k B_{11}^k\|_{Q_\Delta} \le 1 - (1/2)\beta_k a_\Delta\,. \tag{B.8}$$

Since $\|\mathrm{I} - \gamma_k B_{22}^k\|_{Q_{22}} \le \|\mathrm{I} - \gamma_k A_{22}\|_{Q_{22}} + \beta_k\|C_k A_{12}\|_{Q_{22}}$, we obtain the contraction:

$$\begin{aligned}
\|\mathrm{I} - \gamma_k B_{22}^k\|_{Q_{22}} &\le 1 - \gamma_k a_{22} + \beta_k(\mathrm{L}_\infty + \|A_{22}^{-1}A_{21}\|_{Q_\Delta,Q_{22}})\|A_{12}\|_{Q_{22},Q_\Delta} \\
&\le 1 - (1/2)\gamma_k a_{22}.
\end{aligned} \tag{B.9}$$

The last inequality is due to $\kappa \le (a_{22}/2)\{(\mathrm{L}_\infty + \|A_{22}^{-1}A_{21}\|_{Q_\Delta,Q_{22}})\|A_{12}\|_{Q_{22},Q_\Delta}\}^{-1}$. Lastly, the following quantities will be used throughout the analysis:

$$\begin{aligned}
&\Gamma_{m:n}^{(1)} := \prod_{i=m}^{n}(\mathrm{I} - \beta_i B_{11}^i), \quad \Gamma_{m:n}^{(2)} := \prod_{i=m}^{n}(\mathrm{I} - \gamma_i B_{22}^i), \\
&G_{m:n}^{(1)} := \prod_{i=m}^{n}\Big(1 - (1/2)\beta_i a_\Delta\Big), \quad G_{m:n}^{(2)} := \prod_{i=m}^{n}\Big(1 - (1/2)\gamma_i a_{22}\Big).
\end{aligned}$$

As a convention, we define $\Gamma^{(1)}_{m:n} = \Gamma^{(2)}_{m:n} = \mathrm{I}$ if $m > n$. In particular, for any $n, m \geq 0$, we observe the following bound on the operator norm of $\Gamma^{(1)}_{m:n}$,

$$\|\Gamma^{(1)}_{m:n}\| = \sqrt{p_\Delta}\|\Gamma^{(1)}_{m:n}\|_{Q_\Delta} \leq \sqrt{p_\Delta}\prod_{i=m}^{n}\|\,\mathrm{I} - \beta_i B^i_{11}\|_{Q_\Delta} \leq \sqrt{p_\Delta} G^{(1)}_{m:n} \tag{B.10}$$

Similarly, we have $\|\Gamma^{(1)}_{m:n}\| \leq \sqrt{p_{22}} G^{(2)}_{m:n}$. Lastly, we define

$$\Sigma_k := \mathbb{E}\left[\tilde{w}_k \tilde{w}_k^\top\right], \quad \Omega_k := \mathbb{E}\left[\tilde{\theta}_k \tilde{w}_k^\top\right], \quad \Theta_k := \mathbb{E}\left[\tilde{\theta}_k \tilde{\theta}_k^\top\right],$$

whose operator norms correspond to $\mathrm{M}^{\tilde{w}}_k, \mathrm{M}^{\tilde{\theta},\tilde{w}}_k, \mathrm{M}^{\tilde{\theta}}_k$, respectively.

B.2.1 Detailed Proof of Theorem 12

This subsection provides proofs to the propositions stated in Section 2.3.1, as well as providing detailed steps in establishing Theorem 12.

Bounding $\mathrm{M}^{\tilde{w}}_k$ (Proof of Proposition 15) Using (2.24), as the noise terms are martingale, we get

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\tilde{w}_{k+1}\tilde{w}_{k+1}^\top\right] &= (\mathrm{I} - \gamma_k B^k_{22})\tilde{w}_k\tilde{w}_k^\top(\mathrm{I} - \gamma_k B^k_{22})^\top + \gamma_k^2 \mathbb{E}^{\mathcal{F}_k}\left[W_{k+1}W_{k+1}^\top\right] \\ &+ \beta_k^2 C_k \mathbb{E}^{\mathcal{F}_k}\left[V_{k+1}V_{k+1}^\top\right] C_k^\top + \beta_k\gamma_k\left(\mathbb{E}^{\mathcal{F}_k}\left[W_{k+1}V_{k+1}^\top\right] C_k^\top + C_k \mathbb{E}^{\mathcal{F}_k}\left[V_{k+1}W_{k+1}^\top\right]\right).\end{aligned} \tag{B.11}$$

Repeatedly applying (B.11) and taking the total expectation on both sides show

$$\Sigma_{k+1} = \Gamma^{(2)}_{0:k}\Sigma_0(\Gamma^{(2)}_{0:k})^\top + \sum_{j=0}^{k}\Gamma^{(2)}_{j+1:k} D_{j+1}(\Gamma^{(2)}_{j+1:k})^\top, \tag{B.12}$$

where

$$D_{k+1} = \gamma_k^2\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right] + \beta_k^2 C_k \mathbb{E}\left[V_{k+1}V_{k+1}^\top\right] C_k^\top + \beta_k\gamma_k\big(\mathbb{E}\left[W_{k+1}V_{k+1}^\top\right] C_k^\top + C_k\mathbb{E}\left[V_{k+1}W_{k+1}^\top\right]\big).$$

Using Lemma 39, we observe that

$$\gamma_k\beta_k\|\mathbb{E}\left[W_{k+1}V_{k+1}^\top\right] C_k^\top\| \leq \frac{\sqrt{d_\theta d_w}}{2} C_\infty\big(\gamma_k^2\|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| + \beta_k^2\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\|\big), \tag{B.13}$$

Let $K_C := \max\{C_\infty^2, 1\} + \sqrt{d_\theta d_w} C_\infty$, we have

$$\begin{aligned}\|D_{k+1}\| &\leq \gamma_k^2\Big(1 + C_\infty\sqrt{d_\theta d_w}\Big)\|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| + \beta_k^2 C_\infty\Big(C_\infty + \sqrt{d_\theta d_w}\Big)\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\| \\ &\leq K_C\Big(\gamma_k^2\big\{\tilde{m}_V + \tilde{m}_V\,\mathrm{M}^{\tilde{\theta}}_k + \tilde{m}_V\,\mathrm{M}^{\tilde{w}}_k\big\} + \beta_k^2\big\{\tilde{m}_W + \tilde{m}_W\,\mathrm{M}^{\tilde{\theta}}_k + \tilde{m}_W\,\mathrm{M}^{\tilde{w}}_k\big\}\Big)\end{aligned} \tag{B.14}$$

where the last inequality is due to (B.4). Taking the operator norm on both sides of (B.12) yields

$$\mathrm{M}^{\tilde{w}}_{k+1} \leq p_{22}\Big\{\left(G^{(2)}_{0:k}\right)^2\mathrm{M}^{\tilde{w}}_0 + K_C\sum_{j=0}^{k}\left(G^{(2)}_{j+1:k}\right)^2(\gamma_j^2\tilde{m}_V + \beta_j^2\tilde{m}_W)\big\{1 + \mathrm{M}^{\tilde{\theta}}_j + \mathrm{M}^{\tilde{w}}_j\big\}\Big\}. \tag{B.15}$$

Using that $\beta_k \leq \kappa\gamma_k$ one writes

$$\mathrm{M}_{k+1}^{\tilde{w}} \leq \mathrm{C}_0^{\tilde{w}\prime}\big(G_{0:k}^{(2)}\big)^2 + c_0\mathrm{C}_1^{\tilde{w}\prime}\sum_{j=0}^{k}\gamma_j^2\big(G_{j+1:k}^{(2)}\big)^2(1+\mathrm{M}_j^{\tilde{\theta}}) + \mathrm{C}_2^{\tilde{w}\prime}\sum_{j=0}^{k}\gamma_j^2\big(G_{j+1:k}^{(2)}\big)^2\,\mathrm{M}_j^{\tilde{w}} \quad \text{(B.16)}$$

where $c_0 = \tilde{m}_V + \kappa^2\tilde{m}_W$, $\mathrm{C}_0^{\tilde{w}\prime} = p_{22}\,\mathrm{M}_0^{\tilde{w}}$, $\mathrm{C}_1^{\tilde{w}\prime} = p_{22}K_C$, and $\mathrm{C}_2^{\tilde{w}\prime} = p_{22}c_0$. Define:

$$\tilde{\mathrm{U}}_k = \mathrm{C}_0^{\tilde{w}\prime}\big(G_{0:k-1}^{(2)}\big)^2 + c_0\mathrm{C}_1^{\tilde{w}\prime}\sum_{j=0}^{k-1}\gamma_j^2\big(G_{j+1:k-1}^{(2)}\big)^2(1+\mathrm{M}_j^{\tilde{\theta}}) + \mathrm{C}_2^{\tilde{w}\prime}\sum_{j=0}^{k-1}\gamma_j^2\big(G_{j+1:k-1}^{(2)}\big)^2\tilde{\mathrm{U}}_j,$$

It is easily seen that the sequence $(\tilde{\mathrm{U}}_k)_{k\geq 0}$ is given by the following recursion

$$\tilde{\mathrm{U}}_{k+1} = (1-a_{22}\gamma_k/2)^2\tilde{\mathrm{U}}_k + c_0\mathrm{C}_1^{\tilde{w}\prime}\gamma_k^2(1+\mathrm{M}_k^{\tilde{\theta}}) + \mathrm{C}_2^{\tilde{w}\prime}\gamma_k^2\tilde{\mathrm{U}}_k, \quad \tilde{\mathrm{U}}_0 = \mathrm{C}_0^{\tilde{w}\prime}.$$

Since the step size was chosen such that $\gamma_k(\mathrm{C}_2^{\tilde{w}\prime} + (a_{22}^2/4)) \leq \frac{a_{22}}{2}$ [cf. (2.26)], we have

$$\tilde{\mathrm{U}}_{k+1} \leq (1-a_{22}\gamma_k/2)\tilde{\mathrm{U}}_k + \mathrm{C}_1^{\tilde{w}\prime}\gamma_k^2(c_0 + c_1\,\mathrm{M}_k^{\tilde{\theta}})$$

which implies

$$\tilde{\mathrm{U}}_{k+1} \leq \mathrm{C}_0\,G_{0:k}^{(2)} + c_0\mathrm{C}_1^{\tilde{w}\prime}\sum_{j=0}^{k}\gamma_j^2(1+\mathrm{M}_j^{\tilde{\theta}})G_{j+1:k}^{(2)}. \quad \text{(B.17)}$$

Observe that $\mathrm{M}_k^{\tilde{w}} \leq \tilde{\mathrm{U}}_k$. Applying Corollary 30 shows that $\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)} \leq \varrho^{a_{22}/2}\gamma_{k+1}$, we get

$$\boxed{\mathrm{M}_{k+1}^{\tilde{w}} \leq \mathrm{C}_0^{\tilde{w}}\,G_{0:k}^{(2)} + \mathrm{C}_1^{\tilde{w}}\,\gamma_{k+1} + \mathrm{C}_2^{\tilde{w}}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{\theta}},} \quad \text{(B.18)}$$

where we recall $K_C := \max\{C_\infty^2, 1\} + \sqrt{d_\theta d_w}C_\infty$, and

$$\mathrm{C}_0^{\tilde{w}} := p_{22}\,\mathrm{M}_0^{\tilde{w}}, \quad \mathrm{C}_1^{\tilde{w}} := p_{22}(\tilde{m}_V + \kappa^2\tilde{m}_W)K_C\varrho^{a_{22}/2}, \quad \mathrm{C}_2^{\tilde{w}} := p_{22}K_C(\tilde{m}_V + \kappa^2\tilde{m}_W). \quad \text{(B.19)}$$

This concludes the proof for Proposition 15.

Bounding $\mathrm{M}_k^{\tilde{\theta},\tilde{w}}$ (Proof of Proposition 16) We proceed by observing the following recursion of Ω_k:

$$\begin{aligned}\Omega_{k+1} = (\mathrm{I}-\beta_k B_{11}^k)\Omega_k(\mathrm{I}-\gamma_k B_{22}^k)^\top - \beta_k A_{12}\Sigma_k(\mathrm{I}-\gamma_k B_{22}^k)^\top \\ + \beta_k\gamma_k\mathbb{E}\left[V_{k+1}W_{k+1}^\top\right] + \beta_k^2\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]C_k^\top.\end{aligned} \quad \text{(B.20)}$$

Repeatedly applying the recursion gives

$$\Omega_{k+1} = \Gamma_{0:k}^{(1)}\Omega_0\left(\Gamma_{0:k}^{(2)}\right)^\top - \sum_{j=0}^{k}\beta_j\Gamma_{j+1:k}^{(1)}A_{12}\Sigma_j\left(\Gamma_{j:k}^{(2)}\right)^\top \quad \text{(B.21)}$$

$$+\sum_{j=0}^{k}\beta_j\gamma_j\Gamma_{j+1:k}^{(1)}\mathbb{E}\left[V_{j+1}W_{j+1}^\top\right]\left(\Gamma_{j+1:k}^{(2)}\right)^\top + \sum_{j=0}^{k}\beta_j^2\Gamma_{j+1:k}^{(1)}\mathbb{E}\left[V_{j+1}V_{j+1}^\top\right]C_j^\top\left(\Gamma_{j+1:k}^{(2)}\right)^\top. \quad \text{(B.22)}$$

The contraction properties (B.9), (B.8) result in

$$\mathrm{M}_{k+1}^{\tilde{\theta},\tilde{w}} \leq p_{22,\Delta}\Big\{ G_{0:k}^{(1)} G_{0:k}^{(2)}\, \mathrm{M}_0^{\tilde{\theta},\tilde{w}} + \|A_{12}\| \sum_{j=0}^{k} \beta_j G_{j+1:k}^{(1)} G_{j:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \Big\} \tag{B.23}$$

$$+ p_{22,\Delta}\Big\{ \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)} \|\mathbb{E}\left[V_{j+1} W_{j+1}^\top\right]\| + \mathrm{C}_\infty \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} G_{j+1:k}^{(2)} \|\mathbb{E}\left[V_{j+1} V_{j+1}^\top\right]\| \Big\} \tag{B.24}$$

Applying (B.18), we bound the third last term of (B.23) as

$$\sum_{j=0}^{k} \beta_j G_{j+1:k}^{(1)} G_{j:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \leq \sum_{j=0}^{k} \beta_j G_{j+1:k}^{(1)} G_{j:k}^{(2)} \big(\mathrm{C}_0^{\tilde{w}}\, G_{0:j-1}^{(2)} + \mathrm{C}_1^{\tilde{w}}\, \gamma_j + \mathrm{C}_2^{\tilde{w}} \sum_{i=0}^{j-1} \gamma_i^2 G_{i+1:j-1}^{(2)}\, \mathrm{M}_i^{\tilde{\theta}} \big)$$

$$\leq \frac{2\, \mathrm{C}_0^{\tilde{w}}\, G_{0:k}^{(2)}}{a_\Delta} + \mathrm{C}_1^{\tilde{w}} \sum_{j=0}^{k} \beta_j G_{j:k}^{(2)} \gamma_j + \mathrm{C}_2^{\tilde{w}} \sum_{j=0}^{k} \beta_j G_{j+1:k}^{(1)} G_{j:k}^{(2)} \sum_{i=0}^{j-1} \gamma_i^2 G_{i+1:j-1}^{(2)}\, \mathrm{M}_i^{\tilde{\theta}} \tag{B.25}$$

where we have used Lemma 29 and $G_{j+1:k}^{(1)} \leq 1$ in the last inequality. Applying Corollary 30 and Lemma 31, A10 to the second and the last term on the right hand side, respectively, we obtain the following upper bound:

$$\sum_{j=0}^{k} \beta_j G_{j+1:k}^{(1)} G_{j:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \leq \frac{2\, \mathrm{C}_0^{\tilde{w}}\, G_{0:k}^{(2)}}{a_\Delta} + \mathrm{C}_1^{\tilde{w}}\, \varrho^{a_{22}/2} \beta_{k+1} + \frac{2\, \mathrm{C}_2^{\tilde{w}}}{a_\Delta} \sum_{i=0}^{k} \gamma_i^2 G_{i+1:k}^{(2)}\, \mathrm{M}_i^{\tilde{\theta}}\,. \tag{B.26}$$

Applying (B.5), we bound the second last term of (B.23) as

$$\sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)} \|\mathbb{E}\left[V_{j+1} W_{j+1}^\top\right]\| \leq \tilde{m}_{VW} \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)} \big(1 + \mathrm{M}_j^{\tilde{\theta}} + \mathrm{M}_j^{\tilde{w}}\big)$$

$$\leq \tilde{m}_{VW} \Big\{ \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(2)} + \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{\theta}} + \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \Big\}$$

$$\leq \tilde{m}_{VW} \Big\{ \varrho^{a_{22}/2} \beta_{k+1} + \kappa \sum_{j=0}^{k} \gamma_j^2 G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{\theta}} + \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \Big\} \tag{B.27}$$

where the last inequality applied Corollary 30 again. We observe

$$\sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \leq \frac{\gamma_0}{1 - \gamma_0 a_{22}/2} \sum_{j=0}^{k} \beta_j G_{j+1:k-1}^{(1)} G_{j:k-1}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \tag{B.28}$$

Thirdly, we repeat the calculations above and exploit $\beta_k \leq \kappa \gamma_k$ to bound

$$\sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} G_{j+1:k}^{(2)} \|\mathbb{E}\left[V_{j+1} V_{j+1}^\top\right]\| \leq \tilde{m}_V \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} G_{j+1:k}^{(2)} \big(1 + \mathrm{M}_j^{\tilde{\theta}} + \mathrm{M}_j^{\tilde{w}}\big)$$

$$\leq \tilde{m}_V \Big\{ \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} + \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{\theta}} + \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \Big\} \tag{B.29}$$

$$\leq \tilde{m}_V \Big\{ \kappa \varrho^{a_{22}/2} \beta_{k+1} + \kappa^2 \sum_{j=0}^{k} \gamma_j^2 G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{\theta}} + \kappa \sum_{j=0}^{k} \beta_j \gamma_j G_{j+1:k}^{(1)} G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{w}} \Big\}.$$

Combining (B.26), (B.27), (B.28), (B.29), we conclude that

$$\boxed{\mathrm{M}_{k+1}^{\tilde{\theta},\tilde{w}} \leq \mathrm{C}_0^{\tilde{\theta},\tilde{w}}\, G_{0:k}^{(2)} + \mathrm{C}_1^{\tilde{\theta},\tilde{w}}\, \beta_{k+1} + \mathrm{C}_2^{\tilde{\theta},\tilde{w}} \sum_{j=0}^{k} \gamma_j^2 G_{j+1:k}^{(2)}\, \mathrm{M}_j^{\tilde{\theta}}} \tag{B.30}$$

where

$$\begin{aligned}
\mathrm{C}_0^{\tilde{\theta},\tilde{w}} &:= p_{22,\Delta}\Big(\mathrm{M}_0^{\tilde{\theta},\tilde{w}} + \|A_{12}\| \frac{2\,\mathrm{C}_0^{\tilde{w}}}{a_\Delta} + (\tilde{m}_{VW} + \kappa C_\infty \tilde{m}_V) \frac{2\gamma_0\, \mathrm{C}_0^{\tilde{w}}}{a_\Delta(1-\gamma_0 a_{22}/2)} \Big), \\
\mathrm{C}_1^{\tilde{\theta},\tilde{w}} &:= p_{22,\Delta} \varrho^{a_{22}/2} \Big(\mathrm{C}_1^{\tilde{w}} \left(\|A_{12}\| + \frac{\gamma_0}{1-\gamma_0 a_{22}/2} (\tilde{m}_{VW} + C_\infty \kappa \tilde{m}_V) \right) + \tilde{m}_{VW} + C_\infty \kappa \tilde{m}_V \Big), \\
\mathrm{C}_2^{\tilde{\theta},\tilde{w}} &:= p_{22,\Delta} \Big(\frac{2\,\mathrm{C}_2^{\tilde{w}}}{a_\Delta} \big(\|A_{12}\| + \frac{\gamma_0}{1-\gamma_0 a_{22}/2} (\tilde{m}_{VW} + C_\infty \kappa \tilde{m}_V) \big) + \kappa \big(\tilde{m}_{VW} + C_\infty \kappa \tilde{m}_V \big) \Big).
\end{aligned} \tag{B.31}$$

This concludes the proof of Proposition 16.

Bounding $\mathrm{M}_k^{\tilde{\theta}}$ (Proof of Proposition 17) We observe the following recursion:

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_k}\left[\tilde{\theta}_{k+1}\tilde{\theta}_{k+1}^\top\right] &= (\mathrm{I} - \beta_k B_{11}^k) \mathbb{E}^{\mathcal{F}_k}\left[\tilde{\theta}_k \tilde{\theta}_k^\top\right] (\mathrm{I} - \beta_k B_{11}^k)^\top + \beta_k^2 A_{12} \mathbb{E}^{\mathcal{F}_k}\left[\tilde{w}_k \tilde{w}_k^\top\right] A_{12}^\top \\
&\quad + \beta_k^2 \mathbb{E}^{\mathcal{F}_k}\left[V_{k+1} V_{k+1}^\top\right] - \beta_k \Big((\mathrm{I} - \beta_k B_{11}^k) \mathbb{E}^{\mathcal{F}_k}\left[\tilde{\theta}_k \tilde{w}_k^\top\right] A_{12}^\top + A_{12} \mathbb{E}^{\mathcal{F}_k}\left[\tilde{w}_k \tilde{\theta}_k^\top\right] (\mathrm{I} - \beta_k B_{11}^k)^\top \Big)
\end{aligned}$$

Taking total expectations and evaluating the recursion gives

$$\begin{aligned}
\Theta_{k+1} &= \Gamma_{0:k}^{(1)} \Theta_0 (\Gamma_{0:k}^{(1)})^\top + \sum_{j=0}^{k} \beta_j^2 \Gamma_{j+1:k}^{(1)} (A_{12} \Sigma_j A_{12}^\top + \mathbb{E}\left[V_{j+1} V_{j+1}^\top\right]) \big(\Gamma_{j+1:k}^{(1)}\big)^\top \\
&\quad - \sum_{j=0}^{k} \beta_j \Gamma_{j+1:k}^{(1)} ((\mathrm{I} - \beta_j B_{11}^j) \Omega_j A_{12}^\top + A_{12} \Omega_j^\top (\mathrm{I} - \beta_j B_{11}^j)^\top) \big(\Gamma_{j+1:k}^{(1)}\big)^\top
\end{aligned}$$

The above implies

$$\begin{aligned}
\mathrm{M}_{k+1}^{\tilde{\theta}} &\leq p_\Delta \Big\{ (G_{0:k}^{(1)})^2\, \mathrm{M}_0^{\tilde{\theta}} + 2\|A_{12}\| \sum_{j=0}^{k} \beta_j (G_{j+1:k}^{(1)})^2 (1 - \beta_j a_\Delta/2)\, \mathrm{M}_j^{\tilde{\theta},\tilde{w}} \Big\} \\
&\quad + p_\Delta \sum_{j=0}^{k} \beta_j^2 (G_{j+1:k}^{(1)})^2 \big(\|A_{12}\|^2\, \mathrm{M}_j^{\tilde{w}} + \|\mathbb{E}\left[V_{j+1} V_{j+1}^\top\right]\| \big)
\end{aligned} \tag{B.32}$$

Applying (B.4) and Corollary 30 yield

$$\begin{aligned}
\mathrm{M}_{k+1}^{\tilde{\theta}} &\leq p_\Delta \Big\{ (G_{0:k}^{(1)})^2\, \mathrm{M}_0^{\tilde{\theta}} + \tilde{m}_V \varrho^{a_\Delta/2} \beta_{k+1} + \tilde{m}_V \sum_{j=0}^{k} \beta_j^2 (G_{j+1:k}^{(1)})^2\, \mathrm{M}_j^{\tilde{\theta}} \Big\} \\
&\quad + 2p_\Delta \|A_{12}\| \sum_{j=0}^{k} \beta_j G_{j:k}^{(1)} G_{j+1:k}^{(1)}\, \mathrm{M}_j^{\tilde{\theta},\tilde{w}} + p_\Delta \big(\|A_{12}\|^2 + \tilde{m}_V \big) \sum_{j=0}^{k} \beta_j^2 (G_{j+1:k}^{(1)})^2\, \mathrm{M}_j^{\tilde{w}},
\end{aligned} \tag{B.33}$$

Applying (B.30), we can bound the second last term in (B.33) as

$$\sum_{j=0}^{k}\beta_j G^{(1)}_{j:k}G^{(1)}_{j+1:k}\,\mathrm{M}_j^{\tilde\theta,\tilde w}\le\sum_{j=0}^{k}\beta_j G^{(1)}_{j:k}G^{(1)}_{j+1:k}\Big(\,\mathrm{C}_0^{\tilde\theta,\tilde w}\,G^{(2)}_{0:j-1}+\mathrm{C}_1^{\tilde\theta,\tilde w}\,\beta_j+\mathrm{C}_2^{\tilde\theta,\tilde w}\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta}\Big)$$
$$\le\frac{2\,\mathrm{C}_0^{\tilde\theta,\tilde w}}{a_\Delta}G^{(1)}_{0:k}+\mathrm{C}_1^{\tilde\theta,\tilde w}\,\varrho^{a_{22}/2}\beta_{k+1}+\mathrm{C}_2^{\tilde\theta,\tilde w}\sum_{j=0}^{k}\beta_j G^{(1)}_{j:k}G^{(1)}_{j+1:k}\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta}$$

where the second inequality is derived using Corollary 30. To bound the last term in the above we start from the following observation. Indeed, taking into account definition of β_∞ in (2.26), we get $(1-\beta_\ell a_\Delta/2)^{-1}\le 1+\beta_\ell a_\Delta$. This inequality and assumption A10-2 yield that

$$\begin{aligned}&\frac{\gamma_{\ell-1}}{\gamma_\ell}\frac{1-\gamma_\ell a_{22}/2}{1-\beta_\ell a_\Delta/2}\le(1+\epsilon_\gamma\gamma_\ell)(1-\gamma_\ell a_{22}/2)(1+\beta_\ell a_\Delta)\\&\le 1-\gamma_\ell\Big\{a_{22}/2-a_\Delta\kappa-\epsilon\Big\}+\epsilon_\gamma\gamma_\ell^2\Big\{\kappa a_\Delta-a_{22}/2\Big\}\le 1-(1/8)a_{22}\gamma_\ell,\end{aligned}\tag{B.34}$$

since $\kappa_\infty\le(1/4)a_{22}/a_\Delta$, see (2.21). We observe the following chain

$$\begin{aligned}&\sum_{j=0}^{k}\beta_j G^{(1)}_{j:k}G^{(1)}_{j+1:k}\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta}=\sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde\theta}\sum_{j=i+1}^{k}\beta_j G^{(1)}_{j:k}G^{(1)}_{j+1:k}G^{(2)}_{i+1:j-1}\\&=\sum_{i=0}^{k-1}\gamma_i^2 G^{(1)}_{i+1:k}\,\mathrm{M}_i^{\tilde\theta}\sum_{j=i+1}^{k}\beta_j G^{(1)}_{j+1:k}\frac{G^{(2)}_{i+1:j-1}}{G^{(1)}_{i+1:j-1}}\overset{(a)}{\le}\sum_{i=0}^{k-1}\beta_i\gamma_i G^{(1)}_{i+1:k}\,\mathrm{M}_i^{\tilde\theta}\sum_{j=i+1}^{k}\gamma_{j-1}\prod_{\ell=i+1}^{j-1}\frac{\gamma_{\ell-1}}{\gamma_\ell}\frac{G^{(2)}_{i+1:j-1}}{G^{(1)}_{i+1:j-1}}\\&\overset{(b)}{\le}\sum_{i=0}^{k-1}\beta_i\gamma_i G^{(1)}_{i+1:k}\,\mathrm{M}_i^{\tilde\theta}\sum_{j=i+1}^{k}\gamma_{j-1}\prod_{\ell=i+1}^{j-1}(1-(1/8)\gamma_\ell a_{22})\overset{(c)}{\le}\frac{8\varsigma}{a_{22}}\sum_{i=0}^{k-1}\beta_i\gamma_i G^{(1)}_{i+1:k}\,\mathrm{M}_i^{\tilde\theta}.\end{aligned}\tag{B.35}$$

where (a) is due to $\beta_j\le\beta_i$ and $G^{(1)}_{j+1:k}\le 1$, (b) is due to (B.34), (c) is due to A10-1 and $\sum_{j=i+1}^{k}\gamma_j\prod_{\ell=i+1}^{j-1}(1-\gamma_\ell\tilde a)\le(8/a_{22})$ for any i,k.

Moreover, applying (B.18), we can bound the last term of (B.33) as:

$$\sum_{j=0}^{k}\beta_j^2(G^{(1)}_{j+1:k})^2\,\mathrm{M}_j^{\tilde w}\le\sum_{j=0}^{k}\beta_j^2(G^{(1)}_{j+1:k})^2\Big(\,\mathrm{C}_0^{\tilde w}\,G^{(2)}_{0:j-1}+\mathrm{C}_1^{\tilde w}\,\gamma_j+\mathrm{C}_2^{\tilde w}\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta}\Big)$$
$$\le\frac{\mathrm{C}_0^{\tilde w}\,G^{(1)}_{0:k}}{1-\beta_0 a_\Delta/2}\sum_{j=0}^{k}\beta_j^2 G^{(1)}_{j+1:k}+\gamma_0\,\mathrm{C}_1^{\tilde w}\sum_{j=0}^{k}\beta_j^2 G^{(1)}_{j+1:k}+\mathrm{C}_2^{\tilde w}\sum_{j=0}^{k}\beta_j^2(G^{(1)}_{j+1:k})^2\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta}$$
$$\le\Big(\frac{\mathrm{C}_0^{\tilde w}\,G^{(1)}_{0:k}}{1-\beta_0 a_\Delta/2}+\gamma_0\,\mathrm{C}_1^{\tilde w}\Big)\varrho^{a_{22}/2}\beta_{k+1}+\mathrm{C}_2^{\tilde w}\sum_{j=0}^{k}\beta_j^2(G^{(1)}_{j+1:k})^2\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta},$$

where the last inequality is due to Corollary 30. In addition, similar to (B.35), we can derive the bound

$$\sum_{j=0}^{k}\beta_j^2(G^{(1)}_{j+1:k})^2\sum_{i=0}^{j-1}\gamma_i^2 G^{(2)}_{i+1:j-1}\,\mathrm{M}_i^{\tilde\theta}\le\frac{(8\varsigma)/a_{22}}{1-\beta_0 a_\Delta/2}\sum_{i=0}^{k}\beta_i^2\gamma_i G^{(1)}_{i+1:k}\,\mathrm{M}_i^{\tilde\theta}$$

Substituting the above inequalities into (B.33) leads to

$$\boxed{\mathrm{M}_{k+1}^{\tilde{\theta}} \leq \mathrm{C}_0^{\tilde{\theta}} G_{0:k}^{(1)} + \mathrm{C}_1^{\tilde{\theta}} \beta_{k+1} + \mathrm{C}_2^{\tilde{\theta}} \sum_{j=0}^{k} \gamma_j \beta_j G_{j+1:k}^{(1)} \mathrm{M}_j^{\tilde{\theta}}} \tag{B.36}$$

where

$$\begin{aligned}
\mathrm{C}_0^{\tilde{\theta}} &:= p_\Delta \Big(\mathrm{M}_0^{\tilde{\theta}} + \frac{4\|A_{12}\| \, \mathrm{C}_0^{\tilde{\theta},\tilde{w}}}{a_\Delta} \Big), \\
\mathrm{C}_1^{\tilde{\theta}} &:= p_\Delta \left\{ \tilde{m}_V \varrho^{a_\Delta/2} + 2\|A_{12}\| \, \mathrm{C}_1^{\tilde{\theta},\tilde{w}} \, \varrho^{a_{22}/2} + (\|A_{12}\|^2 + \tilde{m}_V) \Big(\gamma_0 \, \mathrm{C}_1^{\tilde{w}} + \frac{\mathrm{C}_0^{\tilde{w}}}{1 - \beta_0 a_\Delta/2} \Big) \varrho^{a_{22}/2} \right\}, \\
\mathrm{C}_2^{\tilde{\theta}} &:= p_\Delta \left\{ \frac{16\varsigma \|A_{12}\| \, \mathrm{C}_2^{\tilde{\theta},\tilde{w}}}{a_{22}} + \tilde{m}_V + (\|A_{12}\|^2 + \tilde{m}_V) \frac{8 \, \mathrm{C}_2^{\tilde{w}} \, \varsigma / a_{22}}{1 - \beta_0 a_\Delta/2} \right\}.
\end{aligned} \tag{B.37}$$

This completes the proof for Proposition 17.

Completing the Proof of Theorem 12 We complete the proof by analyzing the convergence rate of $\mathrm{M}_k^{\tilde{\theta}}$ using (B.36). Consider the following recursion which upper bounds $\mathrm{M}_k^{\tilde{\theta}}$:

$$\mathrm{U}_{k+1} = \mathrm{C}_0^{\tilde{\theta}} G_{0:k}^{(1)} + \mathrm{C}_1^{\tilde{\theta}} \beta_{k+1} + \mathrm{C}_2^{\tilde{\theta}} \sum_{j=0}^{k} \gamma_j \beta_j G_{j+1:k}^{(1)} \mathrm{U}_j,$$

where we have set $\mathrm{U}_0 = \mathrm{C}_0^{\tilde{\theta}}$. Observe that

$$\begin{aligned}
\mathrm{U}_{k+1} - (1 - \beta_k a_\Delta/2) \, \mathrm{U}_k &= \mathrm{C}_1^{\tilde{\theta}} (\beta_{k+1} - (1 - \beta_k a_\Delta/2)\beta_k) + \mathrm{C}_2^{\tilde{\theta}} \gamma_k \beta_k \, \mathrm{U}_k \\
\iff \mathrm{U}_{k+1} &= (1 - \beta_k (a_\Delta/2 - \mathrm{C}_2^{\tilde{\theta}} \gamma_k)) \, \mathrm{U}_k + \mathrm{C}_1^{\tilde{\theta}} (\beta_{k+1} - \beta_k + \beta_k^2 a_\Delta/2)
\end{aligned}$$

Since $\gamma_k \leq \gamma_0 \leq \frac{a_\Delta}{4 \, \mathrm{C}_2^{\tilde{\theta}}}$, we have

$$\mathrm{U}_{k+1} \leq (1 - \beta_k a_\Delta/4) \, \mathrm{U}_k + \mathrm{C}_1^{\tilde{\theta}} \, \beta_k^2 a_\Delta/2$$

Evaluating the recursion gives

$$\mathrm{U}_{k+1} \leq \prod_{\ell=0}^{k} (1 - \beta_\ell a_\Delta/4) \, \mathrm{U}_0 + \mathrm{C}_1^{\tilde{\theta}} (a_\Delta/2) \sum_{j=0}^{k} \beta_j^2 \prod_{\ell=j+1}^{k} (1 - \beta_\ell a_\Delta/4)$$

Applying Corollary 30 shows $\sum_{j=0}^{k} \beta_j^2 \prod_{\ell=j+1}^{k} (1 - \beta_\ell a_\Delta/4) \leq \varrho^{a_\Delta/4} \beta_{k+1}$. Lastly, observing that $\mathrm{M}_k^{\tilde{\theta}} \leq \mathrm{U}_k$ gives

$$\boxed{\mathrm{M}_{k+1}^{\tilde{\theta}} \leq \mathrm{C}_0^{\tilde{\theta}} \prod_{\ell=0}^{k} \Big(1 - \beta_\ell \frac{a_\Delta}{4} \Big) + \mathrm{C}_1^{\tilde{\theta}} \, \varrho^{a_\Delta/4} \frac{a_\Delta}{2} \beta_{k+1}.} \tag{B.38}$$

To finish the proof of (2.14), we observe (i) the constant $\mathrm{C}_0^{\tilde{\theta}} \leq \mathrm{C}_0^{\tilde{\theta},\mathsf{mtg}} \mathrm{V}_0$ for some constant $\mathrm{C}_0^{\tilde{\theta},\mathsf{mtg}}$, (ii) the inequality that $\mathbb{E}\,[\|\theta_k - \theta^\star\|^2] \leq d_\theta \, \mathrm{M}_k^{\tilde{\theta}}$, and (iii) setting the constant $\mathrm{C}_1^{\tilde{\theta},\mathsf{mtg}} := \mathrm{C}_1^{\tilde{\theta}} \, \varrho^{a_\Delta/4} (a_\Delta/2)$.

Our last endeavor is to prove (2.15). Observe that the tracking error $\widehat{w}_k := w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)$ may be represented as

$$\begin{aligned}\widehat{w}_k &= w_k - w^\star + w^\star - A_{22}^{-1}(b_2 - A_{21}\theta_k)\\ &= \tilde{w}_k - C_{k-1}\tilde{\theta}_k + A_{22}^{-1}\big((b_2 - A_{21}\theta^\star) - (b_2 - A_{21}\theta_k)\big) = \tilde{w}_k - L_k\tilde{\theta}_k\end{aligned}$$

using the definitions in (2.22). This leads to the following estimate of $\mathrm{M}_k^{\widehat{w}} := \|\mathbb{E}\left[\widehat{w}_k\widehat{w}_k^\top\right]\|$:

$$\mathrm{M}_{k+1}^{\widehat{w}} \le 2\,\mathrm{M}_{k+1}^{\tilde{w}} + 2\|L_{k+1}\|^2\,\mathrm{M}_{k+1}^{\tilde{\theta}} \le 2\,\mathrm{M}_{k+1}^{\tilde{w}} + 2\,\mathrm{L}_\infty^2\,\frac{\lambda_{\mathsf{max}}(Q_{22})}{\lambda_{\mathsf{min}}(Q_\Delta)}\,\mathrm{M}_{k+1}^{\tilde{\theta}} \tag{B.39}$$

In particular, substituting (B.38) into (B.18), we obtain:

$$\begin{aligned}\mathrm{M}_{k+1}^{\tilde{w}} &\le \mathrm{C}_0^{\tilde{w}}\,G_{0:k}^{(2)} + \mathrm{C}_1^{\tilde{w}}\,\gamma_{k+1} + \mathrm{C}_2^{\tilde{w}}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\Big\{\,\mathrm{C}_0^{\tilde{\theta}}\prod_{\ell=0}^{j-1}\Big(1-\beta_\ell\frac{a_\Delta}{4}\Big) + \mathrm{C}_1^{\tilde{\theta}}\,\varrho^{a_\Delta/4}\frac{a_\Delta}{2}\beta_j\Big\}\\ &\le \mathrm{C}_0^{\tilde{w}}\,G_{0:k}^{(2)} + \gamma_{k+1}\Bigg\{\,\mathrm{C}_1^{\tilde{w}} + \mathrm{C}_1^{\tilde{\theta}}\,\varrho^{a_\Delta/4}\frac{a_\Delta}{2}\,\mathrm{C}_2^{\tilde{w}}\,\varrho^{a_{22}/2} + \frac{\mathrm{C}_2^{\tilde{w}}\,\mathrm{C}_0^{\tilde{\theta}}\,\varrho^{a_{22}/2}}{1-\beta_0 a_\Delta/4}\prod_{\ell=0}^{k}\Big(1-\beta_\ell\frac{a_\Delta}{4}\Big)\Bigg\}\end{aligned}$$

where the last inequality is due to the observation $G_{j+1:k}^{(2)} \le \prod_{i=j+1}^{k}(1-\gamma_i a_{22}/4)^2$ and the application of Corollary 30. Furthermore using $G_{0:k}^{(2)} \le \prod_{\ell=0}^{k}(1-\beta_\ell a_\Delta/4)$ and applying (B.39) gives

$$\boxed{\mathrm{M}_{k+1}^{\widehat{w}} \le \mathrm{C}_0^{w}\prod_{\ell=0}^{k}(1-\beta_\ell a_\Delta/4) + \mathrm{C}_1^{\widehat{w},\mathsf{mtg}}\gamma_{k+1},} \tag{B.40}$$

where

$$\begin{aligned}\mathrm{C}_0^{w} &:= 2\Big\{\,\mathrm{L}_\infty^2\,\frac{\lambda_{\mathsf{max}}(Q_{22})}{\lambda_{\mathsf{min}}(Q_\Delta)}\,\mathrm{C}_0^{\tilde{\theta}} + \frac{\varrho^{a_{22}/2}\,\mathrm{C}_2^{\tilde{w}}\,\mathrm{C}_0^{\tilde{\theta}}}{1-\beta_0 a_\Delta/4} + \mathrm{C}_0^{\tilde{w}}\,\Big\}\\ \mathrm{C}_1^{\widehat{w},\mathsf{mtg}} &:= 2\Big\{\kappa\,\mathrm{L}_\infty^2\,\frac{\lambda_{\mathsf{max}}(Q_{22})}{\lambda_{\mathsf{min}}(Q_\Delta)}\varrho^{a_\Delta/4}\frac{a_\Delta}{2}\,\mathrm{C}_1^{\tilde{\theta}} + \mathrm{C}_1^{\tilde{w}} + \varrho^{a_\Delta/4}\frac{a_\Delta}{2}\,\mathrm{C}_2^{\tilde{w}}\,\varrho^{a_{22}/2}\,\mathrm{C}_1^{\tilde{\theta}}\,\Big\}\end{aligned} \tag{B.41}$$

We conclude the proof for Theorem 12 by observing that $\mathrm{C}_0^{w} \le \mathrm{C}_0^{\widehat{w},\mathsf{mtg}}\mathrm{V}_0$ for some constant $\mathrm{C}_0^{\widehat{w},\mathsf{mtg}}$.

B.2.2 Detailed Proofs of Theorem 13

To facilitate our discussions next, define a few additional constants as:

$$\begin{aligned}&\widetilde{G}_{m:n}^{(1)} := \prod_{i=m}^{n}(1-\beta_i a_\Delta/4), \quad \widetilde{G}_{m:n}^{(2)} := \prod_{i=m}^{n}(1-\gamma_i a_{22}/4)\\ &\mathrm{B}_{11,\infty} := \|\Delta\| + \sqrt{\lambda_{\mathsf{min}}(Q_\Delta)^{-1}\lambda_{\mathsf{max}}(Q_{22})}\,\mathrm{L}_\infty\,\|A_{12}\|, \quad \mathrm{B}_{22,\infty} := \kappa\,\mathrm{C}_\infty\,\|A_{12}\| + \|A_{22}\|.\end{aligned} \tag{B.42}$$

Before we begin the proof, notice by observing the form of (2.32) that that A12 is satisfied by the Markovian noise through setting

$$m_V = \overline{\mathrm{b}} \vee (3\overline{\mathrm{A}}),\ m_W = \overline{\mathrm{b}} \vee (3\overline{\mathrm{A}}),$$

and furthermore (B.4) is satisfied with $\tilde{m}_V, \tilde{m}_W, \tilde{m}_{VW}$ defined in (B.6) and the above m_V, m_W. Moreover, for $i = 0, 1$, the second order moments of the decomposed noise satisfy:

$$\|\mathbb{E}\left[V_{k+1}^{(i)}(V_{k+1}^{(i)})^\top\right]\| \leq \tilde{m}_V^{(i)}(1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}),\ \|\mathbb{E}\left[W_{k+1}^{(i)}(W_{k+1}^{(i)})^\top\right]\| \leq \tilde{m}_W^{(i)}(1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}), \tag{B.43}$$

$$\|\mathbb{E}\left[V_{k+1}^{(i)}(W_{k+1}^{(i)})^\top\right]\| \leq \tilde{m}_{VW}^{(i)}(1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}}), \tag{B.44}$$

for some constants $\tilde{m}_V^{(i)}$, $\tilde{m}_W^{(i)}$, $\tilde{m}_{VW}^{(i)}$, $i = 1, 2$. We proceed with the proof for Theorem 13 as follows.

Bounding $\mathrm{M}_k^{\tilde{w}}$ (Proof of Lemma 18 and Proposition 19) Repeating the analysis that leaded to (B.16) and using the martingale property of $V_{k+1}^{(0)}, W_{k+1}^{(0)}$ shows that

$$\|\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| \leq (G_{0:k}^{(2)})^2 p_{22}\,\mathrm{M}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_0 \sum_{j=0}^{k} \gamma_j^2 (G_{j+1:k}^{(2)})^2 (1 + \mathrm{M}_j^{\tilde{w}} + \mathrm{M}_j^{\tilde{\theta}}), \tag{B.45}$$

where

$$\widetilde{\mathrm{C}}_0 = p_{22}\Big[\{\mathrm{C}_\infty^2 \vee 1 + \sqrt{d_w d_\theta}\,\mathrm{C}_\infty\}\{\tilde{m}_V + \kappa^2 \tilde{m}_W^{(0)} \vee \tilde{m}_V^{\theta} + \kappa^2 \tilde{m}_W^{(0)}\}\Big] \vee \big[\tilde{m}_V^{(0)} + \kappa^2 \tilde{m}_W^{(0)}\big]. \tag{B.46}$$

Our next endeavor is to bound $\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right]$. Evaluating the recursion in (2.33) gives

$$\tilde{w}_{k+1}^{(1)} = \Gamma_{0:k}^{(2)} \tilde{w}_0^{(1)} + \sum_{j=0}^{k} \gamma_j \Gamma_{j+1:k}^{(2)} (W_{j+1}^{(1)} + C_j V_{j+1}^{(1)}) \tag{B.47}$$

Set $\widetilde{\psi}_j^{b_i} := \psi_j^{b_i} + \Psi_j^{A_{i1}}\theta^\star + \Psi_j^{A_{i2}} w^\star$ for $i = 1, 2$. Using the definitions, the combined noise has the following expression

$$\begin{aligned} W_{j+1}^{(1)} + C_j V_{j+1}^{(1)} &= \big(\widetilde{\psi}_j^{b_2} - \widetilde{\psi}_{j+1}^{b_2}\big) + C_j\big(\widetilde{\psi}_j^{b_1} - \widetilde{\psi}_{j+1}^{b_1}\big) + \Big\{\Psi_j^{A_{22}} - \Psi_{j+1}^{A_{22}} + C_j\big(\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}}\big)\Big\}\tilde{w}_j \\ &+ \Big\{\Psi_j^{A_{21}} - \Psi_{j+1}^{A_{21}} - (\Psi_j^{A_{22}} - \Psi_{j+1}^{A_{22}})C_{j-1} + C_j(\Psi_j^{A_{11}} - \Psi_{j+1}^{A_{11}}) - C_j(\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}})C_{j-1}\Big\}\tilde{\theta}_j \end{aligned} \tag{B.48}$$

Upon some algebra manipulations that are detailed in Appendix B.2.2, we deduce that the combined noise may be decomposed as:

$$\begin{aligned} W_{j+1}^{(1)} + C_j V_{j+1}^{(1)} \equiv\ & \psi_j^{WV} - \psi_{j+1}^{WV} + \widetilde{\Psi}_j^{WV,\tilde{\theta}}\tilde{\theta}_j + \widetilde{\Psi}_j^{WV,\tilde{w}}\tilde{w}_j \\ & + \big(\Upsilon_j^{WV,\tilde{\theta}}\tilde{\theta}_j - \Upsilon_{j+1}^{WV,\tilde{\theta}}\tilde{\theta}_{j+1}\big) + \big(\Upsilon_j^{WV,\tilde{w}}\tilde{w}_j - \Upsilon_{j+1}^{WV,\tilde{w}}\tilde{w}_{j+1}\big) \\ & + \Phi^{WV,\tilde{\theta}}\big(\tilde{\theta}_{j+1} - \tilde{\theta}_j\big) + \Phi^{WV,\tilde{w}}\big(\tilde{w}_{j+1} - \tilde{w}_j\big), \end{aligned} \tag{B.49}$$

where it holds that

$$\|\psi_j^{WV}\| \vee \|\Upsilon_j^{WV,\tilde{\theta}}\| \vee \|\Upsilon_j^{WV,\tilde{w}}\| \vee \|\Phi^{WV,\tilde{w}}\| \vee \|\Phi^{WV,\tilde{\theta}}\| \leq \mathrm{E}_0^{WV},\ \ \|\widetilde{\Psi}_j^{WV,\tilde{\theta}}\| \vee \|\widetilde{\Psi}_j^{WV,\tilde{w}}\| \leq \mathrm{E}_0^{WV}\,\gamma_j, \tag{B.50}$$

with

$$\mathrm{E}_0^{WV} := \max\{\overline{\mathrm{b}}(1+\mathrm{C}_\infty), \overline{\mathrm{A}}(1+2\,\mathrm{C}_\infty+\mathrm{C}_\infty^2), \overline{\mathrm{A}}C_2^U\varrho^{a_{22}/2}(1+\mathrm{C}_\infty)(1+\varsigma)\}. \tag{B.51}$$

Let us bound the second term in (B.47) one by one as follows. Using Lemma 28, we obtain

$$\begin{aligned}
&\sum_{j=0}^{k}\gamma_j\Gamma_{j+1:k}^{(2)}\Big(\psi_j^{WV}-\psi_{j+1}^{WV}+\big(\Upsilon_j^{WV,\tilde{\theta}}\tilde{\theta}_j-\Upsilon_{j+1}^{WV,\tilde{\theta}}\tilde{\theta}_{j+1}\big)+\big(\Upsilon_j^{WV,\tilde{w}}\tilde{w}_j-\Upsilon_{j+1}^{WV,\tilde{w}}\tilde{w}_{j+1}\big)\Big)\\
&=\gamma_0\Gamma_{1:k}^{(2)}\big(\psi_0^{WV}+\Upsilon_0^{WV,\tilde{\theta}}\tilde{\theta}_0+\Upsilon_0^{WV,\tilde{w}}\tilde{w}_0\big)-\gamma_k\big(\psi_{k+1}^{WV}+\Upsilon_{k+1}^{WV,\tilde{\theta}}\tilde{\theta}_{k+1}+\Upsilon_{k+1}^{WV,\tilde{w}}\tilde{w}_{k+1}\big)\\
&+\sum_{j=1}^{k}\big(\gamma_j^2B_{22}^j\Gamma_{j+1:k}^{(2)}+(\gamma_j-\gamma_{j-1})\Gamma_{j:k}^{(2)}\big)\big(\psi_j^{WV}+\Upsilon_j^{WV,\tilde{\theta}}\tilde{\theta}_j+\Upsilon_j^{WV,\tilde{w}}\tilde{w}_j\big),
\end{aligned}$$

Secondly,

$$\sum_{j=0}^{k}\gamma_j\Gamma_{j+1:k}^{(2)}\Phi^{WV,\tilde{\theta}}\big(\tilde{\theta}_{j+1}-\tilde{\theta}_j\big)=-\sum_{j=0}^{k}\gamma_j\beta_j\Gamma_{j+1:k}^{(2)}\Phi^{WV,\tilde{\theta}}\big(A_{12}\tilde{w}_{j+1}+V_{j+1}\big)$$

$$\sum_{j=0}^{k}\gamma_j\Gamma_{j+1:k}^{(2)}\Phi^{WV,\tilde{w}}\big(\tilde{w}_{j+1}-\tilde{w}_j\big)=-\sum_{j=0}^{k}\gamma_j^2\Gamma_{j+1:k}^{(2)}\Phi^{WV,\tilde{w}}\big(W_{j+1}+C_jV_{j+1}\big)$$

As a consequence of (B.43)–(B.44), we have

$$\mathbb{E}\left[\|A_{12}\tilde{w}_{j+1}+V_{j+1}\|^2\right]\le\tilde{m}_{\Delta\tilde{\theta}}\big(1+\mathrm{M}_j^{\tilde{w}}+\mathrm{M}_j^{\tilde{\theta}}\big),\ \ \mathbb{E}\left[\|W_{j+1}+C_jV_{j+1}\|^2\right]\le\tilde{m}_{\Delta\tilde{w}}\big(1+\mathrm{M}_j^{\tilde{w}}+\mathrm{M}_j^{\tilde{\theta}}\big) \tag{B.52}$$

where

$$\tilde{m}_{\Delta\tilde{\theta}}:=2\big\{\|A_{12}\|^2+\tilde{m}_V\big\},\quad \tilde{m}_{\Delta\tilde{w}}:=2(\tilde{m}_W+\mathrm{C}_\infty\,\tilde{m}_V).$$

Noting that $\tilde{w}_0^{(1)}=0$, taking Euclidean norm on both sides of (B.47) yields

$$\begin{aligned}
\|\tilde{w}_{k+1}^{(1)}\|&\le\sqrt{p_{22}}\Big\{\mathrm{E}_0^{WV}\left[G_{1:k}^{(2)}\gamma_0(1+\|\tilde{\theta}_0\|+\|\tilde{w}_0\|)+\gamma_k(1+\|\tilde{\theta}_{k+1}\|+\|\tilde{w}_{k+1}\|)\right]\Big\}\\
&+\sqrt{p_{22}}\,\mathrm{E}_0^{WV}\Big\{\sum_{j=1}^{k}G_{j+1:k}^{(2)}(\gamma_j^2+\|\gamma_j^2B_{22}^j+(\gamma_j-\gamma_{j-1})(\mathrm{I}-B_{22}^j)\|)(1+\|\tilde{w}_j\|+\|\tilde{\theta}_j\|)\Big\}\\
&+\sqrt{p_{22}}\,\mathrm{E}_0^{WV}\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}(\kappa\|A_{12}\tilde{w}_{j+1}+V_{j+1}\|+\|W_{j+1}+C_jV_{j+1}\|)
\end{aligned} \tag{B.53}$$

Note that for any sequence $(b_j)_{j\ge0}$, the following inequality holds:

$$\Big(\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}b_j\Big)^2\le\big(\sum_{i=0}^{k}\gamma_i^2G_{i+1:k}^{(2)}\big)\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}b_j^2\le\gamma_{k+1}\varrho^{a_{22}/2}\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}b_j^2, \tag{B.54}$$

where the first inequality is due to Jensen's inequality and the second inequality is due to Corollary 30. Using $\|B_{22}^j\|\le\mathrm{B}_{22,\infty}$, $|\gamma_j-\gamma_{j-1}|\le\frac{a_{22}}{8}\gamma_j^2$ [cf. it is a direct consequence

of A10-2 and the fact $\gamma_j \le \gamma_{j-1}$] and applying the above inequality to (B.53) yields

$$\begin{aligned}\|\tilde{w}_{k+1}^{(1)}\|^2 &\le 9p_{22}(\mathrm{E}_0^{WV})^2\Big\{(G_{1:k}^{(2)})^2\gamma_0^2(1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|)^2+\gamma_k^2(1+\|\tilde{\theta}_{k+1}\|^2+\|\tilde{w}_{k+1}\|^2)\Big\}\\ &+9p_{22}(\mathrm{E}_0^{WV})^2(\mathrm{B}_{22,\infty}+\frac{a_{22}}{8}+1)^2\varrho^{a_{22}/2}\gamma_{k+1}\Big\{\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}(1+\|\tilde{w}_j\|^2+\|\tilde{\theta}_j\|^2)\Big\}\\ &+9p_{22}(\mathrm{E}_0^{WV})^2\varrho^{a_{22}/2}\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}\big(\kappa\|A_{12}\tilde{w}_{j+1}+V_{j+1}\|^2+\|W_{j+1}+C_jV_{j+1}\|^2\big)\end{aligned} \tag{B.55}$$

Using the fact $\mathbb{E}\left[\|\tilde{w}_k\|^2\right] \le d_w\|\mathbb{E}\left[[]\,\tilde{w}_k\tilde{w}_k^\top\right]\|$, $\mathbb{E}\left[\|\tilde{\theta}_k\|^2\right] \le d_\theta\|\mathbb{E}\left[\tilde{\theta}_k\tilde{\theta}_k^\top\right]\|$ (cf. Corollary 38), taking the expectation on both sides yields

$$\begin{aligned}&\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right] \le 9p_{22}(\mathrm{E}_0^{WV})^2\Big\{(G_{1:k}^{(2)})^2\gamma_0^2(1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|)^2+\gamma_k^2(1+d_\theta\,\mathrm{M}_{k+1}^{\tilde{\theta}}+d_w\,\mathrm{M}_{k+1}^{\tilde{w}})\Big\}\\ &+9p_{22}(\mathrm{E}_0^{WV})^2(\mathrm{B}_{22,\infty}+\frac{a_{22}}{8}+1)^2\varrho^{a_{22}/2}\gamma_{k+1}\Big\{\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}(1+d_\theta\,\mathrm{M}_j^{\tilde{\theta}}+d_w\,\mathrm{M}_j^{\tilde{w}})\Big\}\\ &+9p_{22}(\mathrm{E}_0^{WV})^2(\kappa\tilde{m}_{\Delta\tilde{\theta}}+\tilde{m}_{\Delta\tilde{w}})\varrho^{a_{22}/2}\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}\big(1+\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}}\big)\end{aligned}$$

The above simplifies to

$$\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right] \le \widetilde{\mathrm{C}}_1(G_{0:k}^{(2)})^2+\widetilde{\mathrm{C}}_2\gamma_k^2(\mathrm{M}_{k+1}^{\tilde{\theta}}+\mathrm{M}_{k+1}^{\tilde{w}})+\widetilde{\mathrm{C}}_3\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2G_{j+1:k}^{(2)}(\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}})+\widetilde{\mathrm{C}}_4\gamma_k^2 \tag{B.56}$$

where we have used $\gamma_{k+1} \le \gamma_k$ and defined

$$\begin{aligned}\widetilde{\mathrm{C}}_1 &= 9p_{22}(\mathrm{E}_0^{WV})^2(1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|)^2(\gamma_0/(1-\gamma_0a_{22}/2))^2,\\ \widetilde{\mathrm{C}}_2 &= 9p_{22}(\mathrm{E}_0^{WV})^2(d_\theta\vee d_w),\\ \widetilde{\mathrm{C}}_3 &= 9p_{22}(\mathrm{E}_0^{WV})^2\varrho^{a_{22}/2}\big[(d_\theta\vee d_w)(\mathrm{B}_{22,\infty}+\frac{a_{22}}{8}+1)^2+\big(\kappa\tilde{m}_{\Delta\tilde{\theta}}+\tilde{m}_{\Delta\tilde{w}}\big)\big],\\ \widetilde{\mathrm{C}}_4 &= \widetilde{\mathrm{C}}_2+\varrho^{a_{22}/2}\widetilde{\mathrm{C}}_3.\end{aligned} \tag{B.57}$$

Notice that the intermediate results (B.45), (B.56) lead to Lemma 18.

Compared to (B.45), an important feature of the bound (B.56) is that the latter contains an extra γ_k factor. This indicates that the iterate $\tilde{w}_{k+1}^{(1)}$ driven by Markovian noise decays at a faster rate. As we will demonstrate below, the effect of the additional Markov noise is thus negligible compared to the martingale noise driven terms.

As the operator norm $\|\cdot\|$ is convex, applying Jensen's inequality yields

$$\mathrm{M}_{k+1}^{\tilde{w}} \le 2\|\mathbb{E}\left[\tilde{w}_{k+1}^{(1)}(\tilde{w}_{k+1}^{(1)})^\top\right]\|+2\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| \le 2\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right]+2\|\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\|$$

Substituting (B.45) and (B.56) gives

$$\begin{aligned}\mathrm{M}_{k+1}^{\tilde{w}} &\leq 2\Big\{\widetilde{\mathrm{C}}_1(G_{0:k}^{(2)})^2 + \widetilde{\mathrm{C}}_2\gamma_k^2(\mathrm{M}_{k+1}^{\tilde{w}} + \mathrm{M}_{k+1}^{\tilde{\theta}}) + \widetilde{\mathrm{C}}_3\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}(\mathrm{M}_j^{\tilde{w}} + \mathrm{M}_j^{\tilde{\theta}}) + \widetilde{\mathrm{C}}_4\gamma_k^2\Big\} \\ &\quad + 2\Big\{p_{22}(G_{0:k}^{(2)})^2\,\mathrm{M}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}\gamma_{k+1} + \widetilde{\mathrm{C}}_0\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}(\mathrm{M}_j^{\tilde{w}} + \mathrm{M}_j^{\tilde{\theta}})\Big\}\end{aligned} \tag{B.58}$$

The assumption on step size in (2.34) guarantees $2\widetilde{\mathrm{C}}_2\gamma_k^2 \leq (1/2)$, which further implies

$$\begin{aligned}\mathrm{M}_{k+1}^{\tilde{w}} &\leq 4\Big\{\widetilde{\mathrm{C}}_1(G_{0:k}^{(2)})^2 + \widetilde{\mathrm{C}}_2\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}(\mathrm{M}_j^{\tilde{w}} + \mathrm{M}_j^{\tilde{\theta}}) + \widetilde{\mathrm{C}}_4\gamma_k^2\Big\} \\ &\quad + 4\Big\{p_{22}(G_{0:k}^{(2)})^2\,\mathrm{M}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}\gamma_{k+1} + \widetilde{\mathrm{C}}_0\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}(\mathrm{M}_j^{\tilde{w}} + \mathrm{M}_j^{\tilde{\theta}})\Big\}\end{aligned} \tag{B.59}$$

Like in the proof of Theorem 12, we set

$$\mathrm{U}_{k+1} = G_{0:k}^{(2)}(\widetilde{\mathrm{C}}_1 + p_{22}\,\mathrm{M}_0^{\tilde{w}}) + \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}\gamma_{k+1} + \sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\big(\widetilde{\mathrm{C}}_3\gamma_j + \widetilde{\mathrm{C}}_0\big)(\mathrm{U}_j + \mathrm{M}_j^{\tilde{\theta}}) \tag{B.60}$$

with $\mathrm{U}_0 = \widetilde{\mathrm{C}}_1 + p_{22}\,\mathrm{M}_0^{\tilde{w}}$. Through evaluating the recursion, we observe that for any $k \geq 0$, it holds

$$\begin{aligned}\mathrm{M}_{k+1}^{\tilde{w}} &\leq 4\Big\{\,\mathrm{U}_{k+1} + \sum_{j=1}^{k+1}\gamma_{j-1}^2 G_{j:k}^{(2)}\big(\widetilde{\mathrm{C}}_2\,\mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_4\big)\Big\} \\ &\leq 4\Big\{\,\mathrm{U}_{k+1} + \gamma_k^2(\widetilde{\mathrm{C}}_2\,\mathrm{M}_{k+1}^{\tilde{\theta}} + \widetilde{\mathrm{C}}_4) + \sum_{j=1}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\big(\widetilde{\mathrm{C}}_2\,\mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_4\big)\Big\}\end{aligned} \tag{B.61}$$

where the last inequality is due to A10-2 which guarantees that $\gamma_{j-1}^2 G_{j:k}^{(2)} \leq \gamma_j^2 G_{j+1:k}$. Moreover, the sequence U_{k+1} can be expressed as follows:

$$\begin{aligned}\mathrm{U}_{k+1} - (1 - \gamma_k a_{22}/2)\,\mathrm{U}_k &= \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}(\gamma_{k+1} - \gamma_k(1 - \gamma_k a_{22}/2)) + \gamma_k^2\big(\widetilde{\mathrm{C}}_3\gamma_k + \widetilde{\mathrm{C}}_0\big)(\mathrm{U}_k + \mathrm{M}_k^{\tilde{\theta}}) \\ &\leq \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}(a_{22}/2)\gamma_k^2 + \gamma_k^2\big(\widetilde{\mathrm{C}}_3\gamma_k + \widetilde{\mathrm{C}}_0\big)(\mathrm{U}_k + \mathrm{M}_k^{\tilde{\theta}})\end{aligned} \tag{B.62}$$

As the step size satisfies $\gamma_k\big(\widetilde{\mathrm{C}}_3\gamma_0 + \widetilde{\mathrm{C}}_0\big) \leq \frac{a_{22}}{4}$, we get

$$\begin{aligned}&\mathrm{U}_{k+1} \leq (1 - \gamma_k a_{22}/4)\,\mathrm{U}_k + \gamma_k^2\big(\widetilde{\mathrm{C}}_3\gamma_0 + \widetilde{\mathrm{C}}_0\big)\,\mathrm{M}_k^{\tilde{\theta}} + \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}(a_{22}/2)\gamma_k^2 \\ &\implies \mathrm{U}_{k+1} \leq \widetilde{G}_{0:k}^{(2)}\,\mathrm{U}_0 + \sum_{j=0}^{k}\gamma_j^2\widetilde{G}_{j+1:k}^{(2)}\Big\{\big(\widetilde{\mathrm{C}}_3\gamma_0 + \widetilde{\mathrm{C}}_0\big)\,\mathrm{M}_j^{\tilde{\theta}} + (\widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}(a_{22}/2))\Big\}.\end{aligned} \tag{B.63}$$

Substituting the above into (B.61) yields

$$\begin{aligned}\mathrm{M}_{k+1}^{\tilde{w}} &\leq 4\Big\{\widetilde{G}_{0:k}^{(2)}\,\mathrm{U}_0 + \sum_{j=0}^{k}\gamma_j^2\widetilde{G}_{j+1:k}^{(2)}\big\{\big(\widetilde{\mathrm{C}}_3\gamma_0 + \widetilde{\mathrm{C}}_0\big)\,\mathrm{M}_j^{\tilde{\theta}} + (\widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}(a_{22}/2))\big\}\Big\} \\ &\quad + 4\Big\{\gamma_k^2(\widetilde{\mathrm{C}}_2\,\mathrm{M}_{k+1}^{\tilde{\theta}} + \widetilde{\mathrm{C}}_4) + \sum_{j=1}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\big(\widetilde{\mathrm{C}}_2\,\mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_4\big)\Big\}\end{aligned} \tag{B.64}$$

Finally, using the fact that $\gamma_k^2 \leq \varsigma\gamma_{k+1}$ yields

$$\boxed{\mathrm{M}_{k+1}^{\tilde{w}} \leq \widetilde{G}_{0:k}^{(2)}\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\gamma_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{w}}\sum_{j=0}^{k}\gamma_j^2\widetilde{G}_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}}\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde{\theta}}\,.} \tag{B.65}$$

where

$$\widetilde{\mathrm{C}}_0^{\tilde{w}} := 4(\widetilde{\mathrm{C}}_1 + p_{22}\,\mathrm{M}_0^{\tilde{w}}),\quad \widetilde{\mathrm{C}}_1^{\tilde{w}} := 4(\widetilde{\mathrm{C}}_4(\varsigma + \varrho^{a_{22}/2}) + \widetilde{\mathrm{C}}_0(\varrho^{a_{22}/2})^2(a_{22}/2)) \tag{B.66}$$
$$\widetilde{\mathrm{C}}_2^{\tilde{w}} := 4(\widetilde{\mathrm{C}}_3\gamma_0 + \widetilde{\mathrm{C}}_2 + \widetilde{\mathrm{C}}_0),\quad \widetilde{\mathrm{C}}_3^{\tilde{w}} := 4\widetilde{\mathrm{C}}_2.$$

This concludes the proof for Proposition 19.

Before we proceed, we need to bound $\|\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\|$ and $\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right]$ as well. Substituting (B.65) into (B.45) yields

$$\begin{aligned}\|\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| &\leq (G_{0:k}^{(2)})^2 p_{22}\,\mathrm{M}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_0\sum_{j=0}^{k}\gamma_j^2(G_{j+1:k}^{(2)})^2(1+\mathrm{M}_j^{\tilde{\theta}}) \\ &+ \widetilde{\mathrm{C}}_0\sum_{j=0}^{k}\gamma_j^2(G_{j+1:k}^{(2)})^2\Big(\widetilde{\mathrm{C}}_0^{\tilde{w}}\widetilde{G}_{0:j-1}^{(2)} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\gamma_j + \widetilde{\mathrm{C}}_2^{\tilde{w}}\sum_{i=0}^{j-1}\gamma_i^2\widetilde{G}_{i+1:j-1}^{(2)}\,\mathrm{M}_i^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}}\gamma_{j-1}^2\,\mathrm{M}_j^{\tilde{\theta}}\Big)\end{aligned} \tag{B.67}$$

We observe

$$\begin{aligned}&\sum_{j=0}^{k}\gamma_j^2(G_{j+1:k}^{(2)})^2\sum_{i=0}^{j-1}\gamma_i^2\widetilde{G}_{i+1:j-1}^{(2)}\,\mathrm{M}_i^{\tilde{\theta}} = \sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde{\theta}}\sum_{j=i+1}^{k}\gamma_j^2(G_{j+1:k}^{(2)})^2\widetilde{G}_{i+1:j-1}^{(2)} \\ &\overset{(a)}{\leq} \frac{1}{1-\gamma_0 a_{22}/4}\sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde{\theta}}\,\widetilde{G}_{i+1:k}^{(2)}\sum_{j=i+1}^{k}\gamma_j^2 G_{j+1:k}^{(2)} \overset{(b)}{\leq} \frac{\varrho^{a_{22}/2}\gamma_{k+1}}{1-\gamma_0 a_{22}/4}\sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde{\theta}}\,\widetilde{G}_{i+1:k}^{(2)}\end{aligned}$$

where (a) is due to $G_{j+1:k}^{(2)} \leq \widetilde{G}_{j+1:k}^{(2)}$ and (b) is due to Corollary 30. As such, combining terms in (B.67) yields:

$$\boxed{\|\mathbb{E}\left[\tilde{w}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| \leq \widetilde{\mathrm{C}}_0^{\tilde{w}'}\widetilde{G}_{0:k}^{(2)} + \widetilde{\mathrm{C}}_1^{\tilde{w}'}\gamma_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{w}'}\sum_{j=0}^{k}\gamma_j^2\widetilde{G}_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{\theta}},} \tag{B.68}$$

where

$$\begin{aligned}\widetilde{\mathrm{C}}_0^{\tilde{w}'} &:= p_{22}\,\mathrm{M}_0^{\tilde{\theta}},\quad \widetilde{\mathrm{C}}_1^{\tilde{w}'} := \widetilde{\mathrm{C}}_0\varrho^{a_{22}/2}\big(1 + \widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\big) \\ \widetilde{\mathrm{C}}_2^{\tilde{w}'} &:= \widetilde{\mathrm{C}}_0\Big(1 + \widetilde{\mathrm{C}}_3^{\tilde{w}} + \widetilde{\mathrm{C}}_2^{\tilde{w}}\varrho^{a_{22}/2}\frac{\gamma_0}{1-\gamma_0 a_{22}/4}\Big)\end{aligned} \tag{B.69}$$

Similarly, we can compute the bound for $\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right]$ as follows. Using (B.56):

$$\begin{aligned}\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right] &\leq \widetilde{\mathrm{C}}_1(G_{0:k}^{(2)})^2 + \widetilde{\mathrm{C}}_2\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{\theta}} \\ &+ \widetilde{\mathrm{C}}_2\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde{w}} + \widetilde{\mathrm{C}}_3\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{w}}\end{aligned} \tag{B.70}$$

Notice that

$$\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{w}} \le \sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\Big(\widetilde{\mathrm{C}}_0^{\tilde{w}}\widetilde{G}_{0:j-1}^{(2)} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\gamma_j + \widetilde{\mathrm{C}}_2^{\tilde{w}}\sum_{i=0}^{j-1}\gamma_i^2\widetilde{G}_{i+1:j-1}^{(2)}\,\mathrm{M}_i^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}}\gamma_{j-1}^2\,\mathrm{M}_j^{\tilde{\theta}}\Big)$$

$$\le \varrho^{a_{22}/2}(\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\gamma_0)\gamma_{k+1} + \widetilde{\mathrm{C}}_3^{\tilde{w}}\sum_{j=0}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_2^{\tilde{w}}\sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde{\theta}}\sum_{j=i+1}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\widetilde{G}_{i+1:j-1}^{(2)} \tag{B.71}$$

Since $(1-\gamma a_{22}/2) \le (1-\gamma a_{22}/4)^2$ for any $\gamma > 0$, we have $G_{j+1:k}^{(2)} \le (\widetilde{G}_{j+1:k}^{(2)})^2$, therefore together with Corollary 30 it yields

$$\sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde{\theta}}\sum_{j=i+1}^{k}\gamma_j^2 G_{j+1:k}^{(2)}\widetilde{G}_{i+1:j-1}^{(2)} \le \frac{\varrho^{a_{22}/4}\gamma_{k+1}}{1-\gamma_0 a_{22}/4}\sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde{\theta}}\,\widetilde{G}_{i+1:k}^{(2)}. \tag{B.72}$$

Collecting terms and substituting them in (B.70) yield

$$\boxed{\mathbb{E}\left[\|\tilde{w}_{k+1}^{(1)}\|^2\right] \le \widetilde{\mathrm{C}}_0^{\tilde{w}''}\widetilde{G}_{0:k}^{(2)} + \widetilde{\mathrm{C}}_1^{\tilde{w}''}\gamma_{k+1}^2 + \widetilde{\mathrm{C}}_2^{\tilde{w}''}\gamma_{k+1}\sum_{j=0}^{k}\gamma_j^2\widetilde{G}_{j+1:k}^{(2)}\,\mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}''}\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde{\theta}},} \tag{B.73}$$

where we use again the fact that $\gamma_k^2 \le \varsigma\gamma_{k+1}$ and

$$\widetilde{\mathrm{C}}_0^{\tilde{w}''} := \widetilde{\mathrm{C}}_1 + \gamma_0^2\widetilde{\mathrm{C}}_2\widetilde{\mathrm{C}}_0^{\tilde{w}}, \quad \widetilde{\mathrm{C}}_1^{\tilde{w}''} := \widetilde{\mathrm{C}}_3\varrho^{a_{22}/2}(\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}}\gamma_0) + \varsigma\widetilde{\mathrm{C}}_2\widetilde{\mathrm{C}}_1^{\tilde{w}}$$

$$\widetilde{\mathrm{C}}_2^{\tilde{w}''} := \widetilde{\mathrm{C}}_3\Big(1 + \widetilde{\mathrm{C}}_3^{\tilde{w}} + \widetilde{\mathrm{C}}_2^{\tilde{w}}\frac{\varrho^{a_{22}/4}\gamma_0}{1-\gamma_0 a_{22}/4}\Big) + \varsigma\widetilde{\mathrm{C}}_2\widetilde{\mathrm{C}}_2^{\tilde{w}}, \quad \widetilde{\mathrm{C}}_3^{\tilde{w}''} := \widetilde{\mathrm{C}}_2(1+\gamma_0^2\widetilde{\mathrm{C}}_3^{\tilde{w}}).$$

Bounding the Cross Term (Proof of Lemma 20) Our next endeavor is to bound the cross variance between the *martingale noise* driven terms $\tilde{w}_{k+1}^{(0)}$ and $\tilde{\theta}_{k+1}^{(0)}$. Here, the steps involved are similar to those in bounding $\mathrm{M}_k^{\tilde{\theta},\tilde{w}}$ in the proof of Theorem 12. Particularly, in a similar vein as the derivation of (B.23), we obtain

$$\|\mathbb{E}\left[\tilde{\theta}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| \le p_{22,\Delta}\Big\{G_{0:k}^{(2)}\,\mathrm{M}_0^{\tilde{\theta},\tilde{w}} + \|A_{12}\|\sum_{j=0}^{k}\beta_j G_{j+1:k}^{(1)}G_{j:k}^{(2)}\|\mathbb{E}\left[\tilde{w}_j^{(0)}(\tilde{w}_j^{(0)})^\top\right]\|\Big\} \tag{B.74}$$

$$+ p_{22,\Delta}\Big\{\sum_{j=0}^{k}\beta_j\gamma_j G_{j+1:k}^{(1)}G_{j+1:k}^{(2)}\|\mathbb{E}\left[V_{j+1}^{(0)}(W_{j+1}^{(0)})^\top\right]\| + \mathrm{C}_\infty\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}G_{j+1:k}^{(2)}\|\mathbb{E}\left[V_{j+1}^{(0)}(V_{j+1}^{(0)})^\top\right]\|\Big\} \tag{B.75}$$

By observing that $G_{j:k}^{(2)} \le \widetilde{G}_{j:k}^{(2)}$, we have

$$\|\mathbb{E}\left[\tilde{\theta}_{k+1}^{(0)}(\tilde{w}_{k+1}^{(0)})^\top\right]\| \le p_{22,\Delta}\widetilde{G}_{0:k}^{(2)}\,\mathrm{M}_0^{\tilde{\theta},\tilde{w}} + p_{22,\Delta}\|A_{12}\|\sum_{j=0}^{k}\beta_j G_{j+1:k}^{(1)}G_{j:k}^{(2)}\|\mathbb{E}\left[\tilde{w}_j^{(0)}(\tilde{w}_j^{(0)})^\top\right]\|$$

$$+ p_{22,\Delta}\sum_{j=0}^{k}\beta_j G_{j+1:k}^{(1)}G_{j+1:k}^{(2)}\big(\tilde{m}_{VW}^{(0)}\gamma_j + \tilde{m}_V^{(0)}\,\mathrm{C}_\infty\,\beta_j\big)\big(1 + \mathrm{M}_j^{\tilde{\theta}} + \mathrm{M}_j^{\tilde{w}}\big) \tag{B.76}$$

When combined with (B.65), (B.68), it can be verified using similar steps as in deriving (B.26) that:

$$\sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(2)}_{j:k} \| \mathbb{E}\left[\tilde{w}^{(0)}_j (\tilde{w}^{(0)}_j)^\top\right] \| \leq \frac{2\widetilde{\mathrm{C}}^{\tilde{w}'}_0 \widetilde{G}^{(2)}_{0:k}}{a_\Delta} + \widetilde{\mathrm{C}}^{\tilde{w}'}_1 \varrho^{a_{22}/4} \beta_{k+1} + \frac{2\widetilde{\mathrm{C}}^{\tilde{w}'}_2}{a_\Delta} \sum_{i=0}^{k} \gamma_i^2 \widetilde{G}^{(2)}_{i+1:k} \mathrm{M}^{\tilde{\theta}}_i,$$

$$\sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(2)}_{j:k} \mathrm{M}^{\tilde{w}}_j \leq \frac{2\widetilde{\mathrm{C}}^{\tilde{w}}_0 \widetilde{G}^{(2)}_{0:k}}{a_\Delta} + \widetilde{\mathrm{C}}^{\tilde{w}}_1 \varrho^{a_{22}/4} \beta_{k+1} + \Big(\frac{2\widetilde{\mathrm{C}}^{\tilde{w}}_2}{a_\Delta} + \widetilde{\mathrm{C}}^{\tilde{w}}_3\Big) \sum_{i=0}^{k} \gamma_i^2 \widetilde{G}^{(2)}_{i+1:k} \mathrm{M}^{\tilde{\theta}}_i,$$

Substituting the above into (B.76) gives:

$$\boxed{\| \mathbb{E}\left[\tilde{\theta}^{(0)}_{k+1} (\tilde{w}^{(0)}_{k+1})^\top\right] \| \leq \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_0 \widetilde{G}^{(2)}_{0:k} + \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_1 \beta_{k+1} + \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_2 \sum_{j=0}^{k} \gamma_j^2 \widetilde{G}^{(2)}_{j+1:k} \mathrm{M}^{\tilde{\theta}}_j,} \tag{B.77}$$

where

$$\begin{aligned}
\widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_0 &:= p_{22,\Delta}\Big(\mathrm{M}^{\tilde{\theta},\tilde{w}}_0 + \frac{\widetilde{\mathrm{C}}^{\tilde{w}'}_0 \|A_{12}\| + \widetilde{\mathrm{C}}^{\tilde{w}}_0 \big(\tilde{m}^{(0)}_{VW} \gamma_0 + \tilde{m}^{(0)}_V \mathrm{C}_\infty \beta_0\big)}{a_\Delta/2} \Big) \\
\widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_1 &:= p_{22,\Delta} \varrho^{a_{22}/4} \Big(\widetilde{\mathrm{C}}^{\tilde{w}'}_1 \|A_{12}\| + \widetilde{\mathrm{C}}^{\tilde{w}}_1 \big(\tilde{m}^{(0)}_{VW} \gamma_0 + \tilde{m}^{(0)}_V \mathrm{C}_\infty \beta_0\big) \Big) \\
\widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_2 &:= p_{22,\Delta} \Big\{ \frac{2\widetilde{\mathrm{C}}^{\tilde{w}'}_2}{a_\Delta} \|A_{12}\| + \Big(\frac{2\widetilde{\mathrm{C}}^{\tilde{w}}_2}{a_\Delta} + \widetilde{\mathrm{C}}^{\tilde{w}}_3\Big) \big(\tilde{m}^{(0)}_{VW} \gamma_0 + \tilde{m}^{(0)}_V \mathrm{C}_\infty \beta_0\big) \Big\}
\end{aligned} \tag{B.78}$$

Notice that this concludes the proof of Lemma 20.

Bounding $\mathrm{M}^{\tilde{\theta}}_k$ (Proof of Proposition 21) Like in the proof of Theorem 12, we begin by bounding $\| \mathbb{E}\left[\tilde{w}^{(0)}_k (\tilde{w}^{(0)}_k)^\top\right] \|$ as follows. Evaluating the recursion in (2.33) and following the derivations that lead to (B.32), we obtain

$$\begin{aligned}
\| \mathbb{E}\left[\tilde{\theta}^{(0)}_{k+1} (\tilde{\theta}^{(0)}_{k+1})^\top\right] \| \leq p_\Delta \Big\{ (G^{(1)}_{0:k})^2 \mathrm{M}^{\tilde{\theta}}_0 + 2\|A_{12}\| \sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(1)}_{j:k} \| \mathbb{E}\left[\tilde{\theta}^{(0)}_j (\tilde{w}^{(0)}_j)^\top\right] \| \Big\} \\
+ p_\Delta \Big\{ \sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \big(\|A_{12}\|^2 \| \mathbb{E}\left[\tilde{w}^{(0)}_j (\tilde{w}^{(0)}_j)^\top\right] \| + \| \mathbb{E}\left[V^{(0)}_{j+1} (V^{(0)}_{j+1})^\top\right] \| \big) \Big\}
\end{aligned} \tag{B.79}$$

We apply (B.77) and note that

$$\begin{aligned}
&\sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(1)}_{j:k} \| \mathbb{E}\left[\tilde{\theta}^{(0)}_j (\tilde{w}^{(0)}_j)^\top\right] \| \\
&\leq \sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(1)}_{j:k} \Big(\widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_0 \widetilde{G}^{(2)}_{0:j-1} + \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_1 \beta_j + \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_2 \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}^{(2)}_{i+1:j-1} \mathrm{M}^{\tilde{\theta}}_i \Big) \\
&\overset{(a)}{\leq} \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_0 \frac{G^{(1)}_{0:k}}{a_\Delta/2} + \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_1 \varrho^{a_\Delta/2} \beta_{k+1} + \widetilde{\mathrm{C}}^{\tilde{\theta},\tilde{w}}_2 \sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(1)}_{j:k} \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}^{(2)}_{i+1:j-1} \mathrm{M}^{\tilde{\theta}}_i
\end{aligned}$$

where (a) is due to the observation that $1-\gamma_j a_{22}/4 \le 1-\beta_j a_\Delta/2$ and the application of Lemma 29. Moreover, by a slight modification of (B.35), we have

$$\sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(1)}_{j:k} \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}^{(2)}_{i+1:j-1} \, \mathrm{M}_i^{\tilde{\theta}} \le \frac{16\varsigma}{a_{22}} \sum_{i=0}^{k-1} \beta_i \gamma_i G^{(1)}_{i+1:k} \, \mathrm{M}_i^{\tilde{\theta}} \tag{B.80}$$

Therefore,

$$\begin{aligned} &\sum_{j=0}^{k} \beta_j G^{(1)}_{j+1:k} G^{(1)}_{j:k} \| \mathbb{E}\left[\tilde{\theta}_j^{(0)} (\tilde{w}_j^{(0)})^\top\right] \| \\ &\le \widetilde{\mathrm{C}}_0^{\tilde{\theta},\tilde{w}} \frac{G^{(1)}_{0:k}}{a_\Delta/2} + \widetilde{\mathrm{C}}_1^{\tilde{\theta},\tilde{w}} \varrho^{a_\Delta/2} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{\theta},\tilde{w}} \frac{16\varsigma}{a_{22}} \sum_{i=0}^{k-1} \beta_i \gamma_i G^{(1)}_{i+1:k} \, \mathrm{M}_i^{\tilde{\theta}} \end{aligned} \tag{B.81}$$

Similarly, we apply (B.68), (B.80) and note that

$$\begin{aligned} &\sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \| \mathbb{E}\left[\tilde{w}_j^{(0)} (\tilde{w}_j^{(0)})^\top\right] \| \\ &\le \sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \Big\{ \widetilde{\mathrm{C}}_0^{\tilde{w}'} \widetilde{G}^{(2)}_{0:j-1} + \widetilde{\mathrm{C}}_1^{\tilde{w}'} \gamma_j + \widetilde{\mathrm{C}}_2^{\tilde{w}'} \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}^{(2)}_{i+1:j-1} \, \mathrm{M}_i^{\tilde{\theta}} \Big\} \\ &\le (\widetilde{\mathrm{C}}_0^{\tilde{w}'} + \widetilde{\mathrm{C}}_1^{\tilde{w}'} \gamma_0) \varrho^{a_\Delta/2} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{w}'} \frac{\beta_0 (16\varsigma/a_{22})}{1-\beta_0 a_\Delta/2} \sum_{i=0}^{k-1} \beta_i \gamma_i G^{(1)}_{i+1:k} \, \mathrm{M}_i^{\tilde{\theta}} \end{aligned} \tag{B.82}$$

Finally, we obtain that

$$\begin{aligned} &\sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \| \mathbb{E}\left[V_{j+1}^{(0)} (V_{j+1}^{(0)})^\top\right] \| \le \tilde{m}_V^{(0)} \sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \big(1 + \mathrm{M}_j^{\tilde{\theta}} + \mathrm{M}_j^{\tilde{w}}\big) \\ &\le \tilde{m}_V^{(0)} \Big\{ \varrho^{a_\Delta/2} \beta_{k+1} + \sum_{j=0}^{k} \beta_j^2 G^{(1)}_{j+1:k} \, \mathrm{M}_j^{\tilde{\theta}} + \sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \, \mathrm{M}_j^{\tilde{w}} \Big\} \end{aligned} \tag{B.83}$$

Using the bound in (B.65) and the derivations in (B.82), we have

$$\begin{aligned} \sum_{j=0}^{k} \beta_j^2 (G^{(1)}_{j+1:k})^2 \, \mathrm{M}_j^{\tilde{w}} &\le (\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}} \gamma_0) \varrho^{a_\Delta/2} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{w}} \frac{\beta_0 (16\varsigma/a_{22})}{1-\beta_0 a_\Delta/2} \sum_{i=0}^{k-1} \beta_i \gamma_i G^{(1)}_{i+1:k} \, \mathrm{M}_i^{\tilde{\theta}} \\ &\quad + \widetilde{\mathrm{C}}_3^{\tilde{w}} \gamma_0 \beta_0 \sum_{i=0}^{k} \beta_i \gamma_i G^{(1)}_{i+1:k} \, \mathrm{M}_i^{\tilde{\theta}} \end{aligned}$$

Combining the above results, we obtain that

$$\| \mathbb{E}\left[\tilde{\theta}_{k+1}^{(0)} (\tilde{\theta}_{k+1}^{(0)})^\top\right] \| \le \widetilde{\mathrm{C}}_0^{(0)} G^{(1)}_{0:k} + \widetilde{\mathrm{C}}_1^{(0)} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{(0)} \sum_{j=0}^{k} \beta_j \gamma_j G^{(1)}_{j+1:k} \, \mathrm{M}_j^{\tilde{\theta}}, \tag{B.84}$$

where

$$\widetilde{\mathrm{C}}_0^{(0)} = p_\Delta \Big(\mathrm{M}_0^{\tilde{\theta}} + \widetilde{\mathrm{C}}_0^{\tilde{\theta},\tilde{w}} \frac{4}{a_\Delta/4} \|A_{12}\| \Big), \tag{B.85}$$

$$\widetilde{\mathrm{C}}_1^{(0)} = p_\Delta \varrho^{a_\Delta/2}\Big(2\|A_{12}\|\widetilde{\mathrm{C}}_1^{\tilde\theta,\tilde w} + \|A_{12}\|^2(\widetilde{\mathrm{C}}_0^{\tilde w'} + \widetilde{\mathrm{C}}_1^{\tilde w'}\gamma_0) + \tilde m_V^{(0)}(\widetilde{\mathrm{C}}_0^{\tilde w} + \widetilde{\mathrm{C}}_1^{\tilde w}\gamma_0)\Big),$$

$$\widetilde{\mathrm{C}}_2^{(0)} = p_\Delta\Big\{2\|A_{12}\|\frac{\widetilde{\mathrm{C}}_2^{\tilde\theta,\tilde w}}{a_{22}} + \|A_{12}\|^2\frac{\beta_0(16\varsigma/a_{22})}{1-\beta_0 a_\Delta/2} + \tilde m_V^{(0)}\Big(\widetilde{\mathrm{C}}_2^{\tilde w}\frac{\beta_0(16\varsigma/a_{22})}{1-\beta_0 a_\Delta/2} + \widetilde{\mathrm{C}}_3^{\tilde w}\gamma_0\beta_0\Big)\Big\}$$

To bound the term $\mathbb{E}\left[\|\tilde\theta_{k+1}^{(1)}\|^2\right]$, we proceed by considering the following decomposition:

$$\tilde\theta_{k+1}^{(1)} = \underbrace{\Gamma_{0:k}^{(1)}\tilde\theta_0^{(1)} + \sum_{j=0}^k \beta_j\Gamma_{j+1:k}^{(1)}A_{12}\tilde w_j^{(1)}}_{=\tilde\theta_{k+1}^{(1,0)}} + \underbrace{\sum_{j=0}^k \beta_j\Gamma_{j+1:k}^{(1)}V_{j+1}^{(1)}}_{=\tilde\theta_{k+1}^{(1,1)}} \tag{B.86}$$

As $\tilde\theta_0^{(1)} = 0$, we observe that

$$\|\tilde\theta_{k+1}^{(1,0)}\| \le \sqrt{p_\Delta}\|A_{12}\|\sum_{j=0}^k \beta_j G_{j+1:k}^{(1)}\|\tilde w_j^{(1)}\| \tag{B.87}$$

Taking square on both sides and applying the Jensen's inequality (B.54) yields

$$\mathbb{E}\left[\|\tilde\theta_{k+1}^{(1,0)}\|^2\right] \le \frac{p_\Delta\|A_{12}\|^2}{a_\Delta/2}\sum_{j=0}^k \beta_j G_{j+1:k}^{(1)}\mathbb{E}\left[\|\tilde w_j^{(1)}\|^2\right] \tag{B.88}$$

Applying (B.73) gives

$$\begin{aligned}\mathbb{E}\left[\|\tilde\theta_{k+1}^{(1,0)}\|^2\right] &\le \frac{p_\Delta\|A_{12}\|^2}{a_\Delta/2}\sum_{j=0}^k \beta_j G_{j+1:k}^{(1)}\Big(\widetilde{\mathrm{C}}_0^{\tilde w''}\widetilde G_{0:j-1}^{(2)} + \widetilde{\mathrm{C}}_3^{\tilde w''}\gamma_{j-1}^2\,\mathrm{M}_j^{\tilde\theta} + \widetilde{\mathrm{C}}_1^{\tilde w''}\gamma_j^2\Big)\\ &+ \frac{p_\Delta\|A_{12}\|^2}{a_\Delta/2}\widetilde{\mathrm{C}}_2^{\tilde w''}\sum_{j=0}^k \beta_j\gamma_j G_{j+1:k}^{(1)}\sum_{i=0}^{j-1}\gamma_i^2\widetilde G_{i+1:j-1}^{(2)}\,\mathrm{M}_i^{\tilde\theta}\end{aligned} \tag{B.89}$$

Let us bound the right hand side one by one, we observe

$$\sum_{j=0}^k \beta_j G_{j+1:k}^{(1)}\widetilde G_{0:j-1}^{(2)} \le \sum_{j=0}^k \beta_j(\widetilde G_{j+1:k}^{(1)})^2\widetilde G_{0:j-1}^{(2)} \le \frac{G_{0:k}^{(1)}}{1-\beta_0 a_\Delta/2}\sum_{j=0}^k \beta_j\widetilde G_{j+1:k}^{(1)} \le \frac{(4/a_\Delta)G_{0:k}^{(1)}}{1-\beta_0 a_\Delta/2}$$

$$\sum_{j=0}^k \beta_j G_{j+1:k}^{(1)}\gamma_{j-1}^2\,\mathrm{M}_j^{\tilde\theta} \le \rho_0\sum_{j=0}^k \beta_j^2 G_{j+1:k}^{(1)}\,\mathrm{M}_j^{\tilde\theta},\quad \sum_{j=0}^k \beta_j G_{j+1:k}^{(1)}\gamma_j^2 \le \rho_0\varrho^{a_\Delta/2}\beta_{k+1}$$

where the last two inequalities are due to $\gamma_{j-1}^2 \le \rho_0\beta_j$, see B8. In addition, using the fact $G_{m:n}^{(1)} \le (\widetilde G_{m:n}^{(1)})^2$, we have

$$\begin{aligned}&\sum_{j=0}^k \beta_j\gamma_j G_{j+1:k}^{(1)}\sum_{i=0}^{j-1}\gamma_i^2\widetilde G_{i+1:j-1}^{(2)}\,\mathrm{M}_i^{\tilde\theta} = \sum_{i=0}^{k-1}\gamma_i^2\,\mathrm{M}_i^{\tilde\theta}\sum_{j=i+1}^k \beta_j\gamma_j G_{j+1:k}^{(1)}\widetilde G_{i+1:j-1}^{(2)}\\ &\le \rho_0\sum_{i=0}^{k-1}\beta_i\,\mathrm{M}_i^{\tilde\theta}\sum_{j=i+1}^k \beta_j\gamma_j(\widetilde G_{j+1:k}^{(1)})^2\widetilde G_{i+1:j-1}^{(2)} \le \frac{\rho_0}{1-\beta_0 a_\Delta/4}\sum_{i=0}^{k-1}\beta_i\,\mathrm{M}_i^{\tilde\theta}\,\widetilde G_{i+1:k}^{(1)}\sum_{j=i+1}^k \beta_j\gamma_j\widetilde G_{j+1:k}^{(1)}\\ &\le \frac{\varrho^{a_\Delta/4}\rho_0}{1-\beta_0 a_\Delta/4}\sum_{i=0}^{k-1}\beta_i^2\,\mathrm{M}_i^{\tilde\theta}\,\widetilde G_{i+1:k}^{(1)}\end{aligned}$$

Substituting these back into (B.89) yields

$$\mathbb{E}\left[\|\tilde{\theta}_{k+1}^{(1,0)}\|^2\right] \le \widetilde{\mathrm{C}}_0^{(1,0)}\widetilde{G}_{0:k}^{(1)} + \widetilde{\mathrm{C}}_1^{(1,0)}\beta_{k+1} + \widetilde{\mathrm{C}}_2^{(1,0)}\sum_{i=0}^{k}\beta_i^2\,\mathrm{M}_i^{\tilde{\theta}}\,\widetilde{G}_{i+1:k}^{(1)}, \tag{B.90}$$

where

$$\widetilde{\mathrm{C}}_0^{(1,0)} = \widetilde{\mathrm{C}}_0^{\tilde{w}''}\frac{8p_\Delta\|A_{12}\|^2}{a_\Delta^2(1-\beta_0 a_\Delta/2)},\quad \widetilde{\mathrm{C}}_1^{(1,0)} = \widetilde{\mathrm{C}}_1^{\tilde{w}''}\frac{p_\Delta\|A_{12}\|^2}{a_\Delta/2}\rho_0\varrho^{a_\Delta/2},$$
$$\widetilde{\mathrm{C}}_2^{(1,0)} = \frac{p_\Delta\|A_{12}\|^2}{a_\Delta/2}\rho_0\Big(\widetilde{\mathrm{C}}_3^{\tilde{w}''} + \widetilde{\mathrm{C}}_2^{\tilde{w}''}\frac{\varrho^{a_\Delta/4}}{1-\beta_0 a_\Delta/4}\Big).$$

Next, we bound $\mathbb{E}\left[\|\tilde{\theta}_k^{(1,1)}\|^2\right]$. Set $\widetilde{\psi}_j^{b_1} := \psi_j^{b_1} + \Psi_j^{A_{11}}\theta^\star + \Psi_j^{A_{12}}w^\star$, upon some algebraic manipulations (details in Appendix B.2.2) we observe the following decomposition

$$\begin{aligned} V_{j+1}^{(1)} \equiv{}& \widetilde{\psi}_j^{b_1} - \widetilde{\psi}_{j+1}^{b_1} + \big(\Psi_j^{A_{11}}\tilde{\theta}_j - \Psi_{j+1}^{A_{11}}\tilde{\theta}_{j+1}\big) + \big(\Psi_j^{A_{12}}\tilde{w}_j - \Psi_{j+1}^{A_{12}}\tilde{w}_{j+1}\big) \\ &+ \Psi_j^{A_{11}}\big(\tilde{\theta}_{j+1} - \tilde{\theta}_j\big) + \Psi_j^{A_{12}}\big(\tilde{w}_{j+1} - \tilde{w}_j\big), \end{aligned} \tag{B.91}$$

and from B7 we have

$$\|\widetilde{\psi}_j^{b_1}\| \vee \|\Psi_j^{A_{11}}\| \vee \|\Psi_j^{A_{12}}\| \le \mathrm{E}_0^V := \overline{\mathrm{A}} \vee (\overline{\mathrm{b}} + \overline{\mathrm{A}}(\|\theta^\star\| + \|w^\star\|)). \tag{B.92}$$

Applying Lemma 28, we can show

$$\begin{aligned} &\sum_{j=0}^{k}\beta_j\Gamma_{j+1:k}^{(1)}\Big(\widetilde{\psi}_j^{b_1} - \widetilde{\psi}_{j+1}^{b_1} + \big(\Psi_j^{A_{11}}\tilde{\theta}_j - \Psi_{j+1}^{A_{11}}\tilde{\theta}_{j+1}\big) + \big(\Psi_j^{A_{12}}\tilde{w}_j - \Psi_{j+1}^{A_{12}}\tilde{w}_{j+1}\big)\Big) \\ &= \beta_0\Gamma_{1:k}^{(1)}\big(\widetilde{\psi}_0^{b_1} + \Psi_0^{A_{11}}\tilde{\theta}_0 + \Psi_0^{A_{12}}\tilde{w}_0\big) - \beta_k\big(\widetilde{\psi}_{k+1}^{b_1} + \Psi_{k+1}^{A_{11}}\tilde{\theta}_{k+1} + \Psi_{k+1}^{A_{12}}\tilde{w}_{k+1}\big) \\ &\quad + \sum_{j=1}^{k}\big(\beta_j^2 B_{11}^k\Gamma_{j+1:k}^{(1)} + (\beta_j - \beta_{j-1})\Gamma_{j:k}^{(1)}\big)\big(\widetilde{\psi}_j^{b_1} + \Psi_j^{A_{11}}\tilde{\theta}_j + \Psi_j^{A_{12}}\tilde{w}_j\big). \end{aligned} \tag{B.93}$$

Moreover,

$$\sum_{j=0}^{k}\beta_j\Gamma_{j+1:k}^{(1)}\Psi_j^{A_{11}}\big(\tilde{\theta}_{j+1} - \tilde{\theta}_j\big) = -\sum_{j=0}^{k}\beta_j^2\Gamma_{j+1:k}^{(1)}\Psi_j^{A_{11}}(A_{12}\tilde{w}_j + W_{j+1}) \tag{B.94}$$

$$\sum_{j=0}^{k}\beta_j\Gamma_{j+1:k}^{(1)}\Psi_j^{A_{12}}\big(\tilde{w}_{j+1} - \tilde{w}_j\big) = -\sum_{j=0}^{k}\beta_j\gamma_j\Gamma_{j+1:k}^{(1)}\Psi_j^{A_{12}}(W_{j+1} + C_jV_{j+1}) \tag{B.95}$$

The above inequalities allow us to upper bound $\|\tilde{\theta}_{k+1}^{(1,1)}\|$. Note that as $|\beta_j - \beta_{j-1}| \le \frac{a_\Delta}{16}\beta_j^2$ [cf. A10], we have

$$\begin{aligned} \|\tilde{\theta}_{k+1}^{(1,1)}\| \le{}& \sqrt{p_\Delta}\,\mathrm{E}_0^V\left\{\frac{G_{0:k}^{(1)}}{1-\beta_0 a_\Delta/2}(1 + \|\tilde{w}_0\| + \tilde{\theta}_0\|) + \beta_k(1 + \|\tilde{\theta}_{k+1}\| + \|\tilde{w}_{k+1}\|)\right\} \\ &+ \sqrt{p_\Delta}\,\mathrm{E}_0^V(\mathrm{B}_{11,\infty} + a_\Delta/16)\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}\big(1 + \|\tilde{\theta}_j\| + \|\tilde{w}_j\|\big) \\ &+ \sqrt{p_\Delta}\sum_{j=0}^{k}G_{j+1:k}^{(1)}\big(\beta_j^2\|A_{12}\tilde{w}_j + W_{j+1}\| + \beta_j\gamma_j\|W_{j+1} + C_jV_{j+1}\|\big), \end{aligned} \tag{B.96}$$

Applying the Jensen's inequality (B.54) and taking square on both sides give

$$\begin{aligned}
\|\tilde{\theta}_{k+1}^{(1,1)}\|^2 &\leq 7p_\Delta(\mathrm{E}_0^V)^2\Big\{(G_{0:k}^{(1)})^2\Big(\frac{1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|}{1-\beta_0 a_\Delta/2}\Big)^2+\beta_k^2(1+\|\tilde{\theta}_{k+1}\|^2+\|\tilde{w}_{k+1}\|^2)\Big\}\\
&+7p_\Delta(\mathrm{E}_0^V)^2(\mathrm{B}_{11,\infty}+a_\Delta/16)^2\varrho^{a_\Delta/2}\beta_{k+1}\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}(1+\|\tilde{\theta}_j\|^2+\|\tilde{w}_j\|^2)\Big\}\\
&+7p_\Delta\varrho^{a_\Delta/2}\Big\{\beta_{k+1}\sum_{j=0}^{k}G_{j+1:k}^{(1)}\beta_j^2\|A_{12}\tilde{w}_j+W_{j+1}\|^2+\gamma_{k+1}\sum_{j=0}^{k}G_{j+1:k}^{(1)}\beta_j\gamma_j\|W_{j+1}+C_jV_{j+1}\|^2\Big\},
\end{aligned} \tag{B.97}$$

Note the subtle difference that the last term takes γ_{k+1}. Taking expectation on both sides leads to

$$\begin{aligned}
\mathbb{E}\Big[\|\tilde{\theta}_{k+1}^{(1,1)}\|^2\Big] &\leq 7p_\Delta\Big\{(G_{0:k}^{(1)})^2\Big(\frac{1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|}{1-\beta_0 a_\Delta/2}\Big)^2+\beta_k^2(1+d_\theta\,\mathrm{M}_{k+1}^{\tilde{\theta}}+d_w\,\mathrm{M}_{k+1}^{\tilde{w}})\Big\}\\
&\quad+7p_\Delta(\mathrm{E}_0^V)^2(\mathrm{B}_{11,\infty}+a_\Delta/16)^2\varrho^{a_\Delta/2}\beta_{k+1}\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}(1+d_\theta\,\mathrm{M}_j^{\tilde{\theta}}+d_w\,\mathrm{M}_j^{\tilde{w}})\Big\}\\
&\quad+7p_\Delta\varrho^{a_\Delta/2}\sum_{j=0}^{k}G_{j+1:k}^{(1)}(\beta_0\beta_j^2\tilde{m}_{\Delta\tilde{\theta}}+\beta_j\gamma_j^2\tilde{m}_{\Delta\tilde{w}})(1+\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}}),
\end{aligned} \tag{B.98}$$

where we have used $\beta_{k+1}\leq\beta_0$ and $\gamma_{k+1}\leq\gamma_j$. Again, using the bound $\gamma_j^2\leq\rho_0\beta_j$ from B8, we can simplify the above inequality into

$$\begin{aligned}
\mathbb{E}\Big[\|\tilde{\theta}_{k+1}^{(1,1)}\|^2\Big] &\leq 7p_\Delta\Big\{(G_{0:k}^{(1)})^2\Big(\frac{1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|}{1-\beta_0 a_\Delta/2}\Big)^2+\beta_k^2(1+d_\theta\,\mathrm{M}_{k+1}^{\tilde{\theta}}+d_w\,\mathrm{M}_{k+1}^{\tilde{w}})\Big\}\\
&+7p_\Delta(\mathrm{E}_0^V)^2(\mathrm{B}_{11,\infty}+a_\Delta/16)^2\varrho^{a_\Delta/2}\beta_{k+1}\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}(1+d_\theta\,\mathrm{M}_j^{\tilde{\theta}}+d_w\,\mathrm{M}_j^{\tilde{w}})\Big\}\\
&+7p_\Delta\varrho^{a_\Delta/2}(\beta_0\tilde{m}_{\Delta\tilde{\theta}}+\rho_0\tilde{m}_{\Delta\tilde{w}})\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}(1+\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}})\\
&\leq \widehat{\mathrm{C}}_0^{(1,1)}(G_{0:k}^{(1)})^2+\widehat{\mathrm{C}}_1^{(1,1)}\beta_k^2(1+\mathrm{M}_{k+1}^{\tilde{\theta}}+\mathrm{M}_{k+1}^{\tilde{w}})+\widehat{\mathrm{C}}_2^{(1,1)}\sum_{j=0}^{k}\beta_j^2 G_{j+1:k}^{(1)}(\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}})+\widehat{\mathrm{C}}_3^{(1,1)}\beta_{k+1},
\end{aligned} \tag{B.99}$$

where

$$\widehat{\mathrm{C}}_0^{(1,1)}=7p_\Delta\Big(\frac{1+\|\tilde{w}_0\|+\|\tilde{\theta}_0\|}{1-\beta_0 a_\Delta/2}\Big)^2,\quad \widehat{\mathrm{C}}_1^{(1,1)}=7p_\Delta(d_\theta\vee d_w)$$

$$\widehat{\mathrm{C}}_2^{(1,1)}=7p_\Delta\varrho^{a_\Delta/2}\big\{(d_\theta\vee d_w)(\mathrm{E}_0^V)^2(\mathrm{B}_{11,\infty}+a_\Delta/16)^2\beta_0+(\beta_0\tilde{m}_{\Delta\tilde{\theta}}+\rho_0\tilde{m}_{\Delta\tilde{w}})\big\}$$

$$\widehat{\mathrm{C}}_3^{(1,1)}=7p_\Delta(\varrho^{a_\Delta/2})^2\big((\mathrm{E}_0^V)^2(\mathrm{B}_{11,\infty}+a_\Delta/16)^2+\beta_0\tilde{m}_{\Delta\tilde{\theta}}+\rho_0\tilde{m}_{\Delta\tilde{w}}\big)$$

Observe that

$$\sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} \mathrm{M}_j^{\tilde{w}} \leq \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} \Big\{ \widetilde{\mathrm{C}}_0^{\tilde{w}} \widetilde{G}_{0:j-1}^{(2)} + \widetilde{\mathrm{C}}_1^{\tilde{w}} \gamma_j + \widetilde{\mathrm{C}}_2^{\tilde{w}} \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}_{i+1:j-1}^{(2)} \mathrm{M}_i^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}} \gamma_{j-1}^2 \mathrm{M}_j^{\tilde{\theta}} \Big\}$$
$$\leq \big(\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}} \gamma_0\big) \varrho^{a_\Delta/2} \beta_{k+1} + \widetilde{\mathrm{C}}_3^{\tilde{w}} \gamma_0^2 \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} \mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_2^{\tilde{w}} \sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}_{i+1:j-1}^{(2)} \mathrm{M}_i^{\tilde{\theta}}. \tag{B.100}$$

Furthermore, using $G_{j+1:k}^{(1)} \leq (\widetilde{G}_{j+1:k}^{(1)})^2$ and $\widetilde{G}_{i+1:j-1}^{(2)} \leq \widetilde{G}_{i+1:j-1}^{(1)}$, we have

$$\begin{aligned} &\sum_{j=0}^{k} \beta_j^2 G_{j+1:k}^{(1)} \sum_{i=0}^{j-1} \gamma_i^2 \widetilde{G}_{i+1:j-1}^{(2)} \mathrm{M}_i^{\tilde{\theta}} = \sum_{i=0}^{k-1} \gamma_i^2 \mathrm{M}_i^{\tilde{\theta}} \sum_{j=i+1}^{k} \beta_j^2 G_{j+1:k}^{(1)} \widetilde{G}_{i+1:j-1}^{(2)} \\ &\leq \sum_{i=0}^{k-1} \gamma_i^2 \mathrm{M}_i^{\tilde{\theta}} \sum_{j=i+1}^{k} \beta_j^2 (\widetilde{G}_{j+1:k}^{(1)})^2 \widetilde{G}_{i+1:j-1}^{(1)} \leq \frac{\varrho^{a_\Delta/4} \beta_{k+1}}{1 - \beta_0 a_\Delta/4} \sum_{i=0}^{k-1} \gamma_i^2 \mathrm{M}_i^{\tilde{\theta}} \widetilde{G}_{i+1:k}^{(1)} \end{aligned} \tag{B.101}$$

Moreover, through applying $\widetilde{G}_{i+1:j-1}^{(2)} \leq \widetilde{G}_{i+1:j-1}^{(1)}$ and $\beta_k \leq \beta_j$ for any $j \leq k$, we have

$$\beta_k^2 \mathrm{M}_{k+1}^{\tilde{w}} \leq \beta_0^2 \widetilde{\mathrm{C}}_0^{\tilde{w}} \widetilde{G}_{0:k}^{(1)} + \varsigma \gamma_0 \widetilde{\mathrm{C}}_1^{\tilde{w}} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{w}} \gamma_0^2 \sum_{j=0}^{k} \beta_j^2 \widetilde{G}_{j+1:k}^{(1)} \mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{\tilde{w}} \gamma_0^2 \mathrm{M}_{k+1}^{\tilde{\theta}},$$

where we have used $\beta_k^2 \leq \varsigma \beta_{k+1}$. The above results simplify (B.99) into

$$\mathbb{E}\left[\|\tilde{\theta}_{k+1}^{(1,1)}\|^2\right] \leq \widetilde{\mathrm{C}}_0^{(1,1)} \widetilde{G}_{0:k}^{(1)} + \widetilde{\mathrm{C}}_1^{(1,1)} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{(1,1)} \sum_{j=0}^{k} \beta_j^2 \widetilde{G}_{j+1:k}^{(1)} \mathrm{M}_j^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{(1,1)} \beta_k^2 \mathrm{M}_{k+1}^{\tilde{\theta}}, \tag{B.102}$$

where

$$\begin{aligned} \widetilde{\mathrm{C}}_0^{(1,1)} &= \widehat{\mathrm{C}}_0^{(1,1)} + \beta_0^2 \widehat{\mathrm{C}}_1^{(1,1)} \widetilde{\mathrm{C}}_0^{\tilde{w}}, \\ \widetilde{\mathrm{C}}_1^{(1,1)} &= \widehat{\mathrm{C}}_1^{(1,1)} (1 + \varsigma \gamma_0 \widetilde{\mathrm{C}}_1^{\tilde{w}}) + \widehat{\mathrm{C}}_3^{(1,1)} + \widehat{\mathrm{C}}_2^{(1,1)} \big(\widetilde{\mathrm{C}}_0^{\tilde{w}} + \widetilde{\mathrm{C}}_1^{\tilde{w}} \gamma_0\big) \varrho^{a_\Delta/2} \\ \widetilde{\mathrm{C}}_2^{(1,1)} &= \widehat{\mathrm{C}}_2^{(1,1)} \Big(1 + \widetilde{\mathrm{C}}_3^{\tilde{w}} \gamma_0^2 + \frac{\widetilde{\mathrm{C}}_2^{\tilde{w}} \varrho^{a_\Delta/4} \beta_0}{1 - \beta_0 a_\Delta/4}\Big) + \widehat{\mathrm{C}}_1^{(1,1)} \widetilde{\mathrm{C}}_2^{\tilde{w}} \gamma_0^2, \\ \widetilde{\mathrm{C}}_3^{(1,1)} &= \widehat{\mathrm{C}}_1^{(1,1)} (1 + \widetilde{\mathrm{C}}_3^{\tilde{w}} \gamma_0^2). \end{aligned} \tag{B.103}$$

Finally, combining (B.84), (B.90), (B.102) gives

$$\begin{aligned} \mathrm{M}_{k+1}^{\tilde{\theta}} &\leq 3\big(\|\mathbb{E}\left[\tilde{\theta}_{k+1}^{(0)} (\tilde{\theta}_{k+1}^{(0)})^\top\right]\| + \mathbb{E}\left[\|\tilde{\theta}_{k+1}^{(1,0)}\|^2\right] + \mathbb{E}\left[\|\tilde{\theta}_{k+1}^{(1,1)}\|^2\right]\big) \\ &\leq 3\Big\{\big(\widetilde{\mathrm{C}}_0^{(0)} + \widetilde{\mathrm{C}}_0^{(1,0)} + \widetilde{\mathrm{C}}_0^{(1,1)}\big) \widetilde{G}_{0:k}^{(1)} + \big(\widetilde{\mathrm{C}}_1^{(0)} + \widetilde{\mathrm{C}}_1^{(1,0)} + \widetilde{\mathrm{C}}_1^{(1,1)}\big) \beta_{k+1}\Big\} \\ &\quad + 3\Big\{\big(\widetilde{\mathrm{C}}_2^{(0)} + \widetilde{\mathrm{C}}_2^{(1,0)} + \widetilde{\mathrm{C}}_2^{(1,1)}\big) \sum_{i=0}^{k} \beta_i^2 \widetilde{G}_{i+1:k}^{(1)} \mathrm{M}_i^{\tilde{\theta}} + \widetilde{\mathrm{C}}_3^{(1,1)} \beta_k^2 \mathrm{M}_{k+1}^{\tilde{\theta}}\Big\} \end{aligned} \tag{B.104}$$

As we have $3\widetilde{\mathrm{C}}_3^{(1,1)} \beta_k^2 \leq 1/2$, we have

$$\boxed{\mathrm{M}_{k+1}^{\tilde{\theta}} \leq \widetilde{\mathrm{C}}_0^{\tilde{\theta}} \widetilde{G}_{0:k}^{(1)} + \widetilde{\mathrm{C}}_1^{\tilde{\theta}} \beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde{\theta}} \sum_{i=0}^{k} \beta_i^2 \widetilde{G}_{i+1:k}^{(1)} \mathrm{M}_i^{\tilde{\theta}},} \tag{B.105}$$

where

$$\begin{aligned}\widetilde{\mathrm{C}}_0^{\tilde\theta} &:= 6\big(\widetilde{\mathrm{C}}_0^{(0)} + \widetilde{\mathrm{C}}_0^{(1,0)} + \widetilde{\mathrm{C}}_0^{(1,1)}\big), \quad \widetilde{\mathrm{C}}_1^{\tilde\theta} := 6\big(\widetilde{\mathrm{C}}_1^{(0)} + \widetilde{\mathrm{C}}_1^{(1,0)} + \widetilde{\mathrm{C}}_1^{(1,1)}\big),\\ \widetilde{\mathrm{C}}_2^{\tilde\theta} &:= 6\big(\widetilde{\mathrm{C}}_2^{(0)} + \widetilde{\mathrm{C}}_2^{(1,0)} + \widetilde{\mathrm{C}}_2^{(1,1)}\big).\end{aligned} \tag{B.106}$$

This concludes the proof of Proposition 21.

Completing the Proof of Theorem 13 From (2.38) we can derive a bound for $\mathrm{M}_k^{\tilde\theta}$ as follows. Let $\tilde{\mathrm{U}}_0 = \widetilde{\mathrm{C}}_0^{\tilde\theta}$, observe the following equivalent forms of the recursion

$$\begin{aligned}&\tilde{\mathrm{U}}_{k+1} = \widetilde{\mathrm{C}}_0^{\tilde w}\widetilde{G}_{0:k}^{(1)} + \widetilde{\mathrm{C}}_1^{\tilde\theta}\beta_{k+1} + \widetilde{\mathrm{C}}_2^{\tilde\theta}\sum_{i=0}^{k}\beta_i^2\widetilde{G}_{i+1:k}^{(1)}\tilde{\mathrm{U}}_i\\ \iff &\tilde{\mathrm{U}}_{k+1} = (1-\beta_k a_\Delta/4)\tilde{\mathrm{U}}_k + \widetilde{\mathrm{C}}_1^{\tilde\theta}(\beta_{k+1} - \beta_k + \beta_k^2 a_\Delta/4) + \widetilde{\mathrm{C}}_2^{\tilde\theta}\beta_k^2\tilde{\mathrm{U}}_k\\ &\leq (1-\beta_k a_\Delta/8)\tilde{\mathrm{U}}_k + \widetilde{\mathrm{C}}_1^{\tilde\theta}\beta_k^2 a_\Delta/4\end{aligned}$$

where the last inequality is due to the fact $\beta_k\widetilde{\mathrm{C}}_2^{\tilde\theta} \leq a_\Delta/8$. Subsequently, we have

$$\begin{aligned}\tilde{\mathrm{U}}_{k+1} &\leq \prod_{i=0}^{k}(1-\beta_i a_\Delta/8)\tilde{\mathrm{U}}_0 + \frac{\widetilde{\mathrm{C}}_1^{\tilde\theta}a_\Delta}{4}\sum_{j=0}^{k}\gamma_j^2\prod_{i=j+1}^{k}(1-\beta_i a_\Delta/8)\\ &\leq \prod_{i=0}^{k}(1-\beta_i a_\Delta/8)\tilde{\mathrm{U}}_0 + \frac{\widetilde{\mathrm{C}}_1^{\tilde\theta}a_\Delta}{4}\varrho^{a_\Delta/8}\beta_{k+1},\end{aligned}$$

Observing that $\mathrm{M}_{k+1}^{\tilde\theta} \leq \tilde{\mathrm{U}}_{k+1}$, we obtain

$$\boxed{\mathrm{M}_{k+1}^{\tilde\theta} \leq \widetilde{\mathrm{C}}_0^{\tilde\theta}\prod_{i=0}^{k}(1-\beta_i a_\Delta/8) + \widetilde{\mathrm{C}}_1^{\tilde\theta}\frac{a_\Delta}{4}\varrho^{a_\Delta/8}\beta_{k+1},} \tag{B.107}$$

We obtain (2.18) by setting $\mathrm{C}_1^{\tilde\theta,\mathsf{mark}} = \frac{a_\Delta}{4}\widetilde{\mathrm{C}}_1^{\tilde\theta}\varrho^{a_\Delta/8}$ and observing $\widetilde{\mathrm{C}}_0^{\tilde\theta} \leq \mathrm{C}_0^{\tilde\theta,\mathsf{mark}}(1+\mathrm{V}_0)$ for some constant $\mathrm{C}_0^{\tilde\theta,\mathsf{mark}}$.

Finally, we bound the tracking error $\widehat{w}_k := w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)$ as follows. Similarly to the martingale noise case, we set $\mathrm{M}_k^{\widehat w} := \|\mathbb{E}\left[\widehat{w}_k\widehat{w}_k^\top\right]\|$ and observe:

$$\mathrm{M}_{k+1}^{\widehat w} \leq 2\,\mathrm{M}_{k+1}^{\tilde w} + 2\|L_{k+1}\|^2\,\mathrm{M}_{k+1}^{\tilde\theta} \leq 2\,\mathrm{M}_{k+1}^{\tilde w} + 2\,\mathrm{L}_\infty^2\,\frac{\lambda_{\mathsf{max}}(Q_{22})}{\lambda_{\mathsf{min}}(Q_\Delta)}\,\mathrm{M}_{k+1}^{\tilde\theta}$$

Substituting (B.107) into (B.65) gives

$$\begin{aligned}&\mathrm{M}_{k+1}^{\tilde w} \leq \widetilde{G}_{0:k}^{(2)}\widetilde{\mathrm{C}}_0^{\tilde w} + \widetilde{\mathrm{C}}_1^{\tilde w}\gamma_{k+1} + \widetilde{\mathrm{C}}_3^{\tilde w}\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde\theta} + \widetilde{\mathrm{C}}_2^{\tilde w}\sum_{j=0}^{k}\gamma_j^2\widetilde{G}_{j+1:k}^{(2)}\Big\{\widetilde{\mathrm{C}}_0^{\tilde\theta}\prod_{i=0}^{j-1}\Big(1-\beta_i\frac{a_\Delta}{8}\Big) + \widetilde{\mathrm{C}}_1^{\tilde\theta}\frac{a_\Delta}{4}\varrho^{a_\Delta/8}\beta_j\Big\}\\
&\overset{(a)}{\leq} \widetilde{G}_{0:k}^{(2)}\widetilde{\mathrm{C}}_0^{\tilde w} + \widetilde{\mathrm{C}}_1^{\tilde w}\gamma_{k+1} + \widetilde{\mathrm{C}}_3^{\tilde w}\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde\theta} + \widetilde{\mathrm{C}}_2^{\tilde w}\gamma_{k+1}\Big\{\frac{\mathrm{C}_0^{\tilde\theta}\,\varrho^{a_{22}/8}}{1-\beta_0 a_\Delta/8}\prod_{i=0}^{k}\Big(1-\beta_i\frac{a_\Delta}{8}\Big) + \widetilde{\mathrm{C}}_1^{\tilde\theta}\frac{a_\Delta}{4}\varrho^{a_\Delta/8}\varrho^{a_{22}/4}\Big\}\\
&\leq \Big\{\widetilde{\mathrm{C}}_1^{\tilde w} + \widetilde{\mathrm{C}}_2^{\tilde w}\widetilde{\mathrm{C}}_1^{\tilde\theta}\frac{a_\Delta}{4}\varrho^{a_\Delta/8}\varrho^{a_{22}/4}\Big\}\gamma_{k+1} + \Big\{\widetilde{\mathrm{C}}_0^{\tilde w} + \frac{\widetilde{\mathrm{C}}_2^{\tilde w}\,\mathrm{C}_0^{\tilde\theta}\,\varrho^{a_{22}/8}}{1-\beta_0 a_\Delta/8}\Big\}\prod_{i=0}^{k}\Big(1-\beta_i\frac{a_\Delta}{8}\Big) + \widetilde{\mathrm{C}}_3^{\tilde w}\gamma_k^2\,\mathrm{M}_{k+1}^{\tilde\theta}\end{aligned}$$

where we have used $\widetilde{G}^{(2)}_{j+1:k} \leq \big(\prod_{i=j+1}^{k}(1-\gamma_i a_{22}/8)\big)^2$ in (a). As such, together with (B.107) this gives

$$\boxed{\mathrm{M}^{\widehat{w}}_{k+1} \leq \widetilde{\mathrm{C}}^{\widehat{w}}_0 \prod_{\ell=0}^{k}\Big(1-\beta_\ell \frac{a_\Delta}{8}\Big) + \mathrm{C}^{\widehat{w},\mathsf{mark}}_1 \gamma_{k+1},} \tag{B.108}$$

where

$$\begin{aligned}
\widetilde{\mathrm{C}}^{\widehat{w}}_0 &:= 2\Bigg\{\widetilde{\mathrm{C}}^{\tilde{w}}_0 + \frac{\widetilde{\mathrm{C}}^{\tilde{w}}_2\,\mathrm{C}^{\tilde{\theta}}_0\,\varrho^{a_{22}/8}}{1-\beta_0 a_\Delta/8} + (1+\widetilde{\mathrm{C}}^{\tilde{w}}_3)\,\mathrm{L}^2_\infty\,\frac{\lambda_{\mathsf{max}}(Q_{22})}{\lambda_{\mathsf{min}}(Q_\Delta)}\widetilde{\mathrm{C}}^{\tilde{\theta}}_0\Bigg\} \\
\mathrm{C}^{\widehat{w},\mathsf{mark}}_1 &:= 2\Bigg\{\widetilde{\mathrm{C}}^{\tilde{w}}_1 + \widetilde{\mathrm{C}}^{\tilde{w}}_2\widetilde{\mathrm{C}}^{\tilde{\theta}}_1\frac{a_\Delta}{4}\varrho^{a_\Delta/8}\varrho^{a_{22}/4} + \kappa(1+\widetilde{\mathrm{C}}^{\tilde{w}}_3)\,\mathrm{L}^2_\infty\,\frac{\lambda_{\mathsf{max}}(Q_{22})}{\lambda_{\mathsf{min}}(Q_\Delta)}\widetilde{\mathrm{C}}^{\tilde{\theta}}_1\frac{a_\Delta}{4}\varrho^{a_\Delta/8}\Bigg\}
\end{aligned} \tag{B.109}$$

Similarly, as $\widetilde{\mathrm{C}}^{\widehat{w}}_0 \leq \mathrm{C}^{\widehat{w},\mathsf{mark}}_0(1+\mathrm{V}_0)$ for some constant $\mathrm{C}^{\widehat{w},\mathsf{mark}}_0$, the above yields (2.19). We conclude the proof of Theorem 13.

Auxiliary Results for the Markovian Noise Case

Lemma 28. *Let $(a_j)_{j\geq 0}$ be a sequence of d_θ-dimensional vectors. The following equality holds:*

$$\begin{aligned}
&\textstyle\sum_{j=0}^{k}\beta_j\Gamma^{(1)}_{j+1:k}(a_j - a_{j+1}) \\
&= \beta_0\Gamma^{(1)}_{1:k}a_0 - \beta_k a_{k+1} + \textstyle\sum_{j=1}^{k}\big(\beta_j^2 B^j_{11}\Gamma^{(1)}_{j+1:k} + (\beta_j-\beta_{j-1})\Gamma^{(1)}_{j:k}\big)a_j
\end{aligned} \tag{B.110}$$

Similarly, for $(b_j)_{j\geq 0}$ being a sequence of d_w-dimensional vectors, it holds:

$$\begin{aligned}
&\textstyle\sum_{j=0}^{k}\gamma_j\Gamma^{(2)}_{j+1:k}(b_j - b_{j+1}) \\
&= \gamma_0\Gamma^{(2)}_{1:k}b_0 - \gamma_k b_{k+1} + \textstyle\sum_{j=1}^{k}\big(\gamma_j^2 B^j_{22}\Gamma^{(2)}_{j+1:k} + (\gamma_j-\gamma_{j-1})\Gamma^{(1)}_{j:k}\big)b_j.
\end{aligned} \tag{B.111}$$

Proof. We only prove (B.110). Observe the following chain

$$\begin{aligned}
\sum_{j=0}^{k}\beta_j\Gamma^{(1)}_{j+1:k}(a_j - a_{j+1}) &= \sum_{j=0}^{k}\beta_j\Gamma^{(1)}_{j+1:k}a_j - \sum_{j=0}^{k}\beta_j\Gamma^{(1)}_{j+1:k}a_{j+1} \\
&= \beta_0\Gamma^{(1)}_{1:k}a_0 - \beta_k a_{k+1} + \sum_{j=1}^{k}\big(\beta_j\Gamma^{(1)}_{j+1:k} - \beta_{j-1}\Gamma^{(1)}_{j:k}\big)a_j
\end{aligned} \tag{B.112}$$

Using $\beta_j\Gamma^{(1)}_{j+1:k} - \beta_{j-1}\Gamma^{(1)}_{j:k} = \beta_j^2 B^j_{11}\Gamma^{(1)}_{j+1:k} + (\beta_j-\beta_{j-1})\Gamma^{(1)}_{j:k}$ concludes the proof. □

Derivation of Eq. (B.49) The decomposition is obtained through repeatedly adding/subtracting terms. Particularly, we observe that the individual terms can be expressed as:

$$\begin{aligned}
C_j\big(\widetilde{\psi}_j^{b_1} - \widetilde{\psi}_{j+1}^{b_1}\big) &= C_j\widetilde{\psi}_j^{b_1} - C_{j-1}\widetilde{\psi}_j^{b_1} + C_{j-1}\widetilde{\psi}_j^{b_1} - C_j\widetilde{\psi}_{j+1}^{b_1} \\
\big(\Psi_j^{A_{22}} - \Psi_{j+1}^{A_{22}}\big)\tilde{w}_j &= \Psi_j^{A_{22}}\tilde{w}_j - \Psi_{j+1}^{A_{22}}\tilde{w}_{j+1} + \Psi_{j+1}^{A_{22}}(\tilde{w}_{j+1} - \tilde{w}_j) \\
C_j\big(\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}}\big)\tilde{w}_j &= (C_j - C_{j-1})\Psi_j^{A_{12}}\tilde{w}_j + C_{j-1}\Psi_j^{A_{12}}(\tilde{w}_j - \tilde{w}_{j-1}) \\
&\quad + C_{j-1}\Psi_j^{A_{12}}\tilde{w}_{j-1} - C_j\Psi_{j+1}^{A_{12}}\tilde{w}_j \\
\big(\Psi_j^{A_{21}} - \Psi_{j+1}^{A_{21}}\big)\tilde{\theta}_j &= \Psi_j^{A_{21}}\tilde{\theta}_j - \Psi_{j+1}^{A_{21}}\tilde{\theta}_{j+1} + \Psi_{j+1}^{A_{21}}(\tilde{\theta}_{j+1} - \tilde{\theta}_j) \\
C_j\big(\Psi_j^{A_{11}} - \Psi_{j+1}^{A_{11}}\big)\tilde{\theta}_j &= (C_j - C_{j-1})\Psi_j^{A_{11}}\tilde{\theta}_j + C_{j-1}\Psi_j^{A_{11}}(\tilde{\theta}_j - \tilde{\theta}_{j-1}) \\
&\quad + C_{j-1}\Psi_j^{A_{11}}\tilde{\theta}_{j-1} - C_j\Psi_{j+1}^{A_{11}}\tilde{\theta}_j \\
\big(\Psi_j^{A_{22}} - \Psi_{j+1}^{A_{22}}\big)C_{j-1}\tilde{\theta}_j &= \Psi_j^{A_{22}}(C_{j-1} - C_{j-2})\tilde{\theta}_j + \Psi_j^{A_{22}}C_{j-2}(\tilde{\theta}_j - \tilde{\theta}_{j-1}) \\
&\quad + \Psi_j^{A_{22}}C_{j-2}\tilde{\theta}_{j-1} - \Psi_{j+1}^{A_{22}}C_{j-1}\tilde{\theta}_j \\
C_j(\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}})C_{j-1}\tilde{\theta}_j &= (C_j - C_{j-1})\Psi_j^{A_{12}}C_{j-1}\tilde{\theta}_j + C_{j-1}\Psi_j^{A_{12}}(C_{j-1} - C_{j-2})\tilde{\theta}_j \\
&\quad + C_{j-1}\Psi_j^{A_{12}}C_{j-2}(\tilde{\theta}_j - \tilde{\theta}_{j-1}) \\
&\quad + C_{j-1}\Psi_j^{A_{12}}C_{j-2}\tilde{\theta}_{j-1} - C_j\Psi_{j+1}^{A_{12}}C_{j-1}\tilde{\theta}_j.
\end{aligned}$$

Collecting terms on the right hand side of the above equations yields (B.49). Moreover, we the vectors/matrices that appear in (B.49) can be bounded as

$$\|\psi_j^{WV}\| \le \overline{\mathrm{b}}(1+\mathrm{C}_\infty), \quad \|\Upsilon_j^{WV,\tilde{\theta}}\| \le \overline{\mathrm{A}}(1+2\,\mathrm{C}_\infty+\mathrm{C}_\infty^2), \quad \|\Upsilon_j^{WV,\tilde{w}}\| \le \overline{\mathrm{A}}(1+\mathrm{C}_\infty)$$

$$\|\Phi^{WV,\tilde{\theta}}\| \le \overline{\mathrm{A}}(1+2\,\mathrm{C}_\infty+\mathrm{C}_\infty^2), \quad \|\Phi^{WV,\tilde{w}}\| \le \overline{\mathrm{A}}(1+\mathrm{C}_\infty)$$

$$\|\widetilde{\Psi}_j^{WV,\tilde{\theta}}\| \le \overline{\mathrm{A}}C_2^U\varrho^{a_{22}/2}(1+\mathrm{C}_\infty)(1+\varsigma)\gamma_j, \quad \|\widetilde{\Psi}_j^{WV,\tilde{w}}\| \le \overline{\mathrm{A}}C_2^U\varrho^{a_{22}/2}\gamma_j.$$

where the last inequality is due to Lemma 36 and we have used $\gamma_{j-1} \le \varsigma\gamma_j$ [cf. A10-1]. Consequently, we can establish the bounds on the matrix/vector norms by setting

$$\mathrm{E}_0^{WV} := \max\{\overline{\mathrm{b}}(1+\mathrm{C}_\infty), \overline{\mathrm{A}}(1+2\,\mathrm{C}_\infty+\mathrm{C}_\infty^2), \overline{\mathrm{A}}C_2^U\varrho^{a_{22}/2}(1+\mathrm{C}_\infty)(1+\varsigma)\}. \tag{B.113}$$

Derivation of Eq. (B.91) Setting $\widetilde{\psi}_j^{b_1} := \psi_j^{b_1} + \Psi_j^{A_{11}}\theta^\star + \Psi_j^{A_{12}}w^\star$, we observe

$$\begin{aligned}
V_{j+1}^{(1)} &= (\psi_j^{b_1} - \psi_{j+1}^{b_1}) + (\Psi_j^{A_{11}} - \Psi_{j+1}^{A_{11}})\theta_j + (\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}})w_j \\
&= \widetilde{\psi}_j^{b_1} - \widetilde{\psi}_{j+1}^{b_1} + (\Psi_j^{A_{11}} - \Psi_{j+1}^{A_{11}})\tilde{\theta}_j + (\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}})\tilde{w}_j
\end{aligned}$$

Similar to the previous paragraph, the decomposition is obtained through repeatedly adding/subtracting terms. We observe

$$\begin{aligned}
(\Psi_j^{A_{11}} - \Psi_{j+1}^{A_{11}})\tilde{\theta}_j &= \Psi_j^{A_{11}}\tilde{\theta}_j - \Psi_{j+1}^{A_{11}}\tilde{\theta}_{j+1} + \Psi_{j+1}^{A_{11}}(\tilde{\theta}_{j+1} - \tilde{\theta}_j) \\
(\Psi_j^{A_{12}} - \Psi_{j+1}^{A_{12}})\tilde{w}_j &= \Psi_j^{A_{12}}\tilde{w}_j - \Psi_{j+1}^{A_{12}}\tilde{w}_{j+1} + \Psi_{j+1}^{A_{12}}(\tilde{w}_{j+1} - \tilde{w}_j)
\end{aligned}$$

B.3 Detailed Proof of Theorem 22

Throughout this section we will use additional notations. We denote

$$\kappa_\ell := \frac{\beta_\ell}{\gamma_\ell}.$$

Let

$$\kappa_\infty^{\mathsf{exp}} := \kappa_\infty \wedge (1/2)(\|A_{12}\| \, \mathrm{C}_\infty \, \varrho^{a_{22}})^{-1} \tag{B.114}$$

and

$$\beta_\infty^{\mathsf{exp}} := \beta_\infty^{\mathsf{mtg}} \wedge 1/(4\|\Delta\|) \tag{B.115}$$

We assume $\beta_k \le \beta_\infty^{\mathsf{exp}}, \gamma_k \le \gamma_\infty^{\mathsf{mtg}}, \kappa_k \le \kappa \le \kappa_\infty^{\mathsf{exp}}$. Furthermore, let us define

$$\widetilde{\Gamma}^{(1)}_{m,n} := \prod_{j=m}^{n} (\mathrm{I} - \beta_j \Delta), \; \widetilde{\Gamma}^{(2)}_{m,n} := \prod_{j=m}^{n} (\mathrm{I} - \gamma_j A_{22}).$$

Using standard arguments we may bound operator norm of these matrices

$$\|\widetilde{\Gamma}^{(1)}_{m:n}\| \le \sqrt{p_\Delta} \prod_{j=m}^{n} (1 - a_\Delta \beta_j), \; \|\widetilde{\Gamma}^{(2)}_{m:n}\| \le \sqrt{p_{22}} \prod_{j=m}^{n} (1 - a_{22} \gamma_j) \tag{B.116}$$

We set quantities m_W, m_V from the assumption 12 to be equal to

$$m_W := m_V := \max(m_{VW}^{\mathsf{exp}}, \|\Sigma^{11}\|, \|\Sigma^{12}\|, \|\Sigma^{22}\|)$$

All conditions of Theorem 12 are satisfied. We will use this theorem in the following form

$$\mathrm{M}_k^{\tilde{\theta}} \le \mathrm{C}_{0,\theta}^{\mathsf{exp}} \prod_{\ell=0}^{k-1} (1 - (a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}_{1,\theta}^{\mathsf{exp}} \beta_k, \tag{B.117}$$

$$\mathrm{M}_k^{\tilde{w}} \le \mathrm{C}_{0,w}^{\mathsf{exp}} \prod_{\ell=0}^{k-1} (1 - (a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}_{1,w}^{\mathsf{exp}} \gamma_k, \tag{B.118}$$

where $\mathrm{C}_{0,\theta}^{\mathsf{exp}}, \mathrm{C}_{1,\theta}^{\mathsf{exp}}, \mathrm{C}_{0,w}^{\mathsf{exp}}$ and $\mathrm{C}_{1,w}^{\mathsf{exp}}$ denote corresponding constants from Theorem 12. Similarly to (B.4) and (B.5) we can define $\widetilde{m}_{VW}^{\mathsf{exp}}$. Hence, the following inequality holds

$$\|\mathbb{E}\left[V_j V_j^T\right]\| \vee \|\mathbb{E}\left[W_j W_j^T\right]\| \vee \|\mathbb{E}\left[V_j W_j^T\right]\| \le \widetilde{m}_{VW}^{\mathsf{exp}} (1 + \mathrm{M}_k^{\tilde{\theta}} + \mathrm{M}_k^{\tilde{w}})$$

Applying (2.23) and (2.24) (compare with [46][Formula 4.4]) we may write down the following expansion for $\tilde{\theta}_{k+1}$:

$$\tilde{\theta}_{k+1} = S_{k+1}^{(0)} + \ldots + S_{k+1}^{(6)}, \tag{B.119}$$

where

$$
\begin{aligned}
S^{(0)}_{k+1} &:= \widetilde{\Gamma}^{(1)}_{0:k}\tilde{\theta}_0;\\
S^{(1)}_{k+1} &:= \sum_{j=0}^{k}\beta_j\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}\widetilde{\Gamma}^{(2)}_{0:j}\tilde{w}_0;\\
S^{(2)}_{k+1} &:= \sum_{j=0}^{k}\beta_j\widetilde{\Gamma}^{(1)}_{j+1:k}\delta^{(1)}_j;\\
S^{(3)}_{k+1} &:= \sum_{j=0}^{k}\beta_j\widetilde{\Gamma}^{(1)}_{j+1:k}(V_{j+1}+A_{12}A_{22}^{-1}W_{j+1});\\
S^{(4)}_{k+1} &:= \sum_{j=0}^{k}\beta_j\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}\Bigg(\sum_{\ell=0}^{j-1}\beta_\ell\widetilde{\Gamma}^{(2)}_{\ell+1:j-1}\delta^{(2)}_\ell\Bigg);\\
S^{(5)}_{k+1} &:= \sum_{j=0}^{k}\beta_j\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}\Bigg(\sum_{\ell=0}^{j-1}\beta_\ell\widetilde{\Gamma}^{(2)}_{\ell+1:j}C_\ell V_{\ell+1}\Bigg);\\
S^{(6)}_{k+1} &:= \sum_{j=0}^{k}\beta_j\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}\sum_{\ell=0}^{j-1}\gamma_\ell\widetilde{\Gamma}^{(2)}_{\ell+1:j-1}W_{\ell+1} - \sum_{\ell=0}^{k}\beta_\ell\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}A_{22}^{-1}W_{\ell+1},
\end{aligned}
$$

where

$$
\delta^{(1)}_\ell := A_{12}L_\ell\tilde{\theta}_\ell,\ \delta^{(2)}_\ell := -C_kA_{12}\tilde{w}_\ell.
$$

We will group all terms in the expansion into 5 blocks, $S^{(0)}_{k+1}+S^{(1)}_{k+1}, S^{(2)}_{k+1}+S^{(5)}_{k+1}, S^{(3)}_{k+1}, S^{(4)}_{k+1}$ and $S^{(6)}_{k+1}$. It is easy to see that $S^{(0)}_{k+1}+S^{(1)}_{k+1}$ is uncorrelated with $S^{(3)}_{k+1}, S^{(6)}_{k+1}$ (moreover it is uncorrelated with $S^{(5)}_{k+1}$, but we ignore this fact). Since $\mathbb{E}\left[\left\|\tilde{\theta}_{k+1}\right\|^2\right] = \mathbb{E}\left[\mathrm{Tr}(\tilde{\theta}_{k+1}\tilde{\theta}_{k+1}^\top)\right]$ and by the linearity of trace using expansion we show

$$
\mathbb{E}\left[\left\|\tilde{\theta}_{k+1}\right\|^2\right] = \mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}(S^{(3)}_{k+1})^\top)\right] + J'_{k+1}, \tag{B.120}
$$

where for J'_{k+1} we will use the following crude estimate

$$
\begin{aligned}
|J'_{k+1}| \le\ & 3\mathbb{E}\left[\mathrm{Tr}(\{S^{(0)}_{k+1}+S^{(1)}_{k+1}\}\{S^{(0)}_{k+1}+S^{(1)}_{k+1}\}^\top)\right]\\
&+ 5\mathbb{E}\left[\mathrm{Tr}(\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top) + 5\mathbb{E}\left[\mathrm{Tr}(S^{(6)}_{k+1}(S^{(6)}_{k+1})^\top)\right]\right]\\
&+ 2\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}(S^{(6)}_{k+1})^\top)\right] + 2\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top)\right]\\
&+ 5\mathbb{E}\left[\mathrm{Tr}(S^{(4)}_{k+1}(S^{(4)}_{k+1})^\top)\right] + 2\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}(S^{(4)}_{k+1})^\top)\right]
\end{aligned}
$$

Using martingale property and definition of Σ we rewrite the term $\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}(S^{(3)}_{k+1})^\top)\right]$ as follows

$$
\begin{aligned}
\mathrm{Tr}(\mathbb{E}\left[S^{(3)}_{k+1}(S^{(3)}_{k+1})^\top\right]) &= \sum_{j=0}^{k}\beta_j^2\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}\Sigma[\widetilde{\Gamma}^{(1)}_{j+1:k}]^\top)\\
&\quad+\sum_{j=0}^{k}\beta_j^2\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}(\Sigma_j-\Sigma)[\widetilde{\Gamma}^{(1)}_{j+1:k}]^\top)
\end{aligned} \tag{B.121}
$$

where

$$\Sigma_j := \mathbb{E}\left[V_j V_j^\top\right] + A_{12}A_{22}^{-1}\mathbb{E}\left[W_j W_j^\top\right] A_{22}^{-\top}A_{12}^\top + \mathbb{E}\left[V_j W_j^\top\right] A_{22}^{-\top}A_{12}^\top + A_{12}A_{22}^{-1}\mathbb{E}\left[W_j V_j^\top\right]$$

Leading term in (B.121) For lower bound of the first term in (B.121) we will use the following fact. Since for any $s \in [j+1, k]$

$$(\mathrm{I} - \beta_s\Delta)^\top(\mathrm{I} - \beta_s\Delta) = \mathrm{I} - \beta_s(\Delta + \Delta^\top) + \beta_s^2\Delta^\top\Delta \succeq (1 - 2\beta_s\|\Delta\|)\,\mathrm{I},$$

we obtain using Lemma 29 (and remark after this lemma)

$$\sum_{j=0}^{k} \beta_j^2 \operatorname{Tr}(\widetilde{\Gamma}_{j+1:k}^{(1)}\Sigma[\widetilde{\Gamma}_{j+1:k}^{(1)}]^\top) \geq \beta_{k+1}\operatorname{Tr}\Sigma \sum_{j=0}^{k}\beta_j \prod_{\ell=j+1}^{k}(1 - 2\beta_\ell\|\Delta\|) \geq \mathrm{C}_3^{\mathsf{exp}}\beta_{k+1}\operatorname{Tr}\Sigma \tag{B.122}$$

where

$$\mathrm{C}_3^{\mathsf{exp}} := 1/(8\|\Delta\|) \tag{B.123}$$

and we used $\beta_\infty^{\mathsf{exp}} \leq 1/(4\|\Delta\|)$ and $k \geq k_0^{\mathrm{exp}}$. To obtain upper bound we apply von Neumann trace inequality (i.e. $\operatorname{Tr}(AB) \leq \sum_{j=1}^n a_j b_j$, where $\{a_j\}$ and $\{b_j\}$ are non-increasing sequences of eigenvalues of Hermitian matrices A and B resp.) and Lemma 30

$$\sum_{j=0}^{k} \beta_j^2 \operatorname{Tr}(\widetilde{\Gamma}_{j+1:k}^{(1)}\Sigma[\widetilde{\Gamma}_{j+1:k}^{(1)}]^\top) \leq p_\Delta \operatorname{Tr}(\Sigma)\sum_{j=0}^{k}\beta_j^2 \prod_{\ell=j+1}^{k}(1 - a_\Delta\beta_\ell)^2 \leq \mathrm{C}_4^{\mathsf{exp}}\operatorname{Tr}(\Sigma)\beta_{k+1} \tag{B.124}$$

where

$$\mathrm{C}_4^{\mathsf{exp}} := p_\Delta \varrho^{a_\Delta} \tag{B.125}$$

Inequalities (B.122) and (B.124) together imply (2.40).

Remainder term in (B.121) The second term in (B.121), which we denote by R_{k+1} may be estimated as follows

$$|R_{k+1}| \leq p_\Delta d_\theta \widetilde{m}_{VW}^{\mathsf{exp}}(1 + \|A_{12}A_{22}^{-1}\|)^2 \sum_{j=0}^{k}\beta_j^2 \prod_{\ell=j+1}^{k}(1 - a_\Delta\beta_\ell)^2(\mathrm{M}_j^{\tilde{\theta}} + \mathrm{M}_j^{\tilde{w}})$$

Applying (B.117) and (B.118) and Lemma 30

$$\boxed{|R_{k+1}| \leq \mathrm{C}_{3,0}^{\mathsf{exp}} \prod_{\ell=0}^{k}(1 - (a_\Delta/4)\beta_\ell)\mathrm{V}_0\beta_{k+1} + \mathrm{C}_{3,1}^{\mathsf{exp}}\beta_{k+1}\gamma_{k+1},} \tag{B.126}$$

where

$$\begin{aligned}
\mathrm{C}_{3,0}^{\mathsf{exp}} &:= p_\Delta d_\theta \widetilde{m}_{VW}^{\mathsf{exp}}(1 + \|A_{12}A_{22}^{-1}\|)^2 \varrho^{a_\Delta}(\mathrm{C}_{0,\theta}^{\mathsf{exp}} + \mathrm{C}_{0,w}^{\mathsf{exp}})/(1 - a_\Delta\beta_\infty^{\mathsf{exp}}),\\
\mathrm{C}_{3,1}^{\mathsf{exp}} &:= p_\Delta d_\theta \widetilde{m}_{VW}^{\mathsf{exp}}(1 + \|A_{12}A_{22}^{-1}\|)^2 \varrho^{a_\Delta}(\kappa_\infty^{\mathsf{exp}}\mathrm{C}_{1,\theta}^{\mathsf{exp}} + \mathrm{C}_{1,w}^{\mathsf{exp}})
\end{aligned}$$

Estimation of J'_{k+1} To finish the proof of the theorem it remains to estimate J'_{k+1}. Applying (B.116) it is easy to check that

$$\mathrm{Tr}(\mathbb{E}\,[[]\, S^{(0)}_{k+1}(S^{(0)}_{k+1})^\top]) = \mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{0:k}\mathbb{E}\left[\tilde{\theta}_0\tilde{\theta}_0^\top\right][\widetilde{\Gamma}^{(1)}_{0:k}]^\top) \le p_\Delta \prod_{\ell=0}^{k}(1-a_\Delta\beta_\ell)^2\mathbb{E}\left[\left\|\tilde{\theta}_0\right\|^2\right]$$

Similarly, recalling that $\kappa^{\mathsf{exp}}_\infty \le (1/4)a_{22}/a_\Delta$ and using $\prod_{s=0}^{j}(1-a_{22}\gamma_s)(1-a_\Delta\beta_s)^{-1} \le \prod_{s=0}^{j}(1-(a_{22}/2)\gamma_s)$ we obtain

$$\begin{aligned}\mathrm{Tr}(\mathbb{E}\left[S^{(1)}_{k+1}(S^{(1)}_{k+1})^\top\right]) &= \sum_{j=0}^{k}\sum_{\ell=0}^{k}\beta_j\beta_\ell\,\mathrm{Tr}\left(\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}\widetilde{\Gamma}^{(2)}_{0:j}\mathbb{E}\left[\tilde{w}_0\tilde{w}_0^\top\right][\widetilde{\Gamma}^{(2)}_{0:\ell}\right]^\top A_{12}^\top[\widetilde{\Gamma}^{(1)}_{\ell+1:k}]^\top\right)\\ &\le \mathrm{C}^{\mathsf{exp}}_1\prod_{\ell=0}^{k}(1-a_\Delta\beta_\ell)^2\mathbb{E}\left[\|\tilde{w}_0\|^2\right],\end{aligned}$$

where

$$\mathrm{C}^{\mathsf{exp}}_1 := (4/a_{22}^2)p_{22}p_\Delta\|A_{12}\|^2(\kappa^{\mathsf{exp}}_\infty)^2$$

Hence, we may conclude from the previous two inequalities that

$$\boxed{\mathbb{E}\left[\mathrm{Tr}(\{S^{(0)}_{k+1}+S^{(1)}_{k+1}\}\{S^{(0)}_{k+1}+S^{(1)}_{k+1}\}^\top)\right] \le \mathrm{C}^{\mathsf{exp}}_{0+1}\prod_{\ell=0}^{k}(1-a_\Delta\beta_\ell)^2\mathrm{V}_0}, \tag{B.127}$$

where

$$\mathrm{C}^{\mathsf{exp}}_{0+1} := 2p_\Delta + 4\mathrm{C}^{\mathsf{exp}}_1(1+\mathrm{C}^2_\infty)$$

For the next term in the expansion we have

$$\mathrm{Tr}(\mathbb{E}\left[S^{(2)}_{k+1}(S^{(2)}_{k+1})^\top\right]) \quad \le \sum_{j=0}^{k}\sum_{\ell=0}^{k}\beta_j\beta_\ell\mathbb{E}\left[\left|\,\mathrm{Tr}\left(\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}L_j\tilde{\theta}_j\tilde{\theta}_l^\top L_l^\top A_{12}^T[\Gamma^{(1)}_{\ell+1:k}]^\top\right)\right|\right]$$

We apply Cauchy-Schwartz inequality twice, first $|\,\mathrm{Tr}(AB^\top)| \le \mathrm{Tr}^{1/2}(AA^\top)\,\mathrm{Tr}^{1/2}(BB^\top)$ and then for expectation. We obtain

$$\mathrm{Tr}(\mathbb{E}\,[[]\, S^{(2)}_{k+1}(S^{(2)}_{k+1})^\top]) \le \left(\sum_{j=0}^{k}\beta_j\big(\,\mathrm{Tr}\left(\widetilde{\Gamma}^{(1)}_{j+1:k}A_{12}L_j\mathbb{E}\left[\tilde{\theta}_j\tilde{\theta}_j^\top\right]L_j^\top A_{12}^\top[\widetilde{\Gamma}^{(1)}_{j+1:k}]^\top\right)\big)^{1/2}\right)^2$$

From Lemma 35 we conclude that $\|L_j\|_{Q_\Delta,Q_{22}} \le \mathrm{C}^{\mathsf{exp}}_L\beta_j\gamma_j^{-1}$, where $\mathrm{C}^{\mathsf{exp}}_L := C_D(L_\infty)\varrho^{a_{22}}$. This inequality and Jensen's inequality imply

$$\begin{aligned}\mathrm{Tr}(\mathbb{E}\left[S^{(2)}_{k+1}(S^{(2)}_{k+1})^\top\right]) &\le (\mathrm{C}^{\mathsf{exp}}_L)^2p_\Delta\|A_{12}\|^2_{Q_{22},Q_\Delta}\left(\sum_{j=0}^{k}\beta_j\kappa_j\prod_{\ell=j+1}^{k}(1-a_\Delta\beta_\ell)\{\mathbb{E}\left[\left\|\tilde{\theta}_j\right\|^2\right]\}^{1/2}\right)^2\\ &\le d_\theta(\mathrm{C}^{\mathsf{exp}}_L)^2p_\Delta a_\Delta^{-1}\|A_{12}\|^2_{Q_{22},Q_\Delta}\sum_{j=0}^{k}\beta_j\kappa_j^2\prod_{\ell=j+1}^{k}(1-a_\Delta\beta_\ell)\,\mathrm{M}_j^{\tilde{\theta}}\end{aligned}$$

Applying (B.117) and Lemma 30 we get

$$\mathrm{Tr}(\mathbb{E}\,[[]\, S^{(2)}_{k+1}(S^{(2)}_{k+1})^\top]) \le \mathrm{C}^{\mathsf{exp}}_{2,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0\kappa_{k+1}^2 + \mathrm{C}^{\mathsf{exp}}_{2,1}\beta_{k+1}\kappa_{k+1}^2\}$$

where

$$\mathrm{C}_{2,0}^{\mathsf{exp}} := d_\theta a_\Delta^{-1}(\mathrm{C}_L^{\mathsf{exp}})^2 p_\Delta \|A_{12}\|^2_{Q_{22},Q_\Delta} \mathrm{C}_{0,\theta}^{\mathsf{exp}} \varrho^{a_\Delta/2}/(1-a_\Delta\beta_\infty^{\mathsf{exp}}),$$
$$\mathrm{C}_{2,1}^{\mathsf{exp}} := d_\theta a_\Delta^{-1}(\mathrm{C}_L^{\mathsf{exp}})^2 p_\Delta \|A_{12}\|^2_{Q_{22},Q_\Delta} \mathrm{C}_{1,\theta}^{\mathsf{exp}} \varrho^{a_\Delta}$$

To estimate the next term we rewrite it as follows

$$S_{k+1}^{(4)} = \sum_{\ell=0}^{k} \beta_\ell N_{\ell+1,k} \delta_\ell^{(2)},$$

where

$$N_{\ell+1,k} := \sum_{j=\ell+1}^{k} \beta_j \widetilde{\Gamma}^{(1)}_{j+1:k} A_{12} \widetilde{\Gamma}^{(2)}_{\ell+1:j-1}.$$

It is straightforward to check

$$\begin{aligned}\|N_{\ell+1,k}\| &\le \sqrt{p_{22}p_\Delta}\|A_{12}\|\kappa_\ell \prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s) \sum_{j=\ell+1}^{k}\gamma_j \prod_{s=\ell+1}^{j-1}(1-(a_{22}/2)\gamma_s) \\ &\le \mathrm{C}_N^{\mathsf{exp}}\kappa_\ell(1-a_\Delta\beta_\infty^{\mathsf{exp}})^{-1}\prod_{s=\ell+1}^{k-1}(1-a_\Delta\beta_s),\end{aligned} \tag{B.128}$$

where $\mathrm{C}_N^{\mathsf{exp}} := (2/a_{22})\sqrt{p_{22}p_\Delta}\|A_{12}\|(1-a_\Delta\beta_\infty^{\mathsf{exp}})^{-1}$ and we used (compare with Lemma 29)

$$\sum_{j=\ell+1}^{k}\gamma_j\prod_{s=\ell+1}^{j-1}(1-(a_{22}/2)\gamma_s) = \frac{1}{a_{22}}\biggl\{1-\prod_{s=\ell+1}^{k}(1-(a_{22}/2)\gamma_s)\biggr\} \le 2/a_{22}. \tag{B.129}$$

Applying (B.131), Jensen's inequality and observation

$$\mathbb{E}\left[\left\|\delta_\ell^{(1)}\right\|^2\right] \le \mathrm{C}_\Delta^{\mathsf{exp}}\mathbb{E}\left[\left\|\tilde\theta_\ell\right\|^2\right], \tag{B.130}$$

$$\mathbb{E}\left[\left\|\delta_\ell^{(2)}\right\|^2\right] \le \mathrm{C}_{22}^{\mathsf{exp}}\mathbb{E}\left[\|\tilde w_\ell\|^2\right], \tag{B.131}$$

where
$\mathrm{C}_{22}^{\mathsf{exp}} := \mathrm{C}_\infty^2\,\|A_{12}\|^2$, we obtain

$$\mathrm{Tr}(\mathbb{E}\left[S_{k+1}^{(4)}(S_{k+1}^{(4)})^\top\right]) \le d_w a_\Delta^{-1}\mathrm{C}_{22}^{\mathsf{exp}}(\mathrm{C}_N^{\mathsf{exp}})^2\sum_{\ell=0}^{k}\beta_\ell\kappa_\ell^2\prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s)\,\mathrm{M}_\ell^{\tilde w}.$$

Applying (B.118) and Lemma 30 we get

$$\boxed{\mathrm{Tr}(\mathbb{E}\left[S_{k+1}^{(4)}(S_{k+1}^{(4)})^\top\right]) \le \mathrm{C}_{4,0}^{\mathsf{exp}}\prod_{s=0}^{k}(1-(a_\Delta/4)\beta_s)\mathrm{V}_0 + \mathrm{C}_{4,0}^{\mathsf{exp}}\beta_{k+1}\kappa_{k+1},} \tag{B.132}$$

where

$$\mathrm{C}_{4,0}^{\mathsf{exp}} := d_w\mathrm{C}_{0,w}^{\mathsf{exp}}\mathrm{C}_{22}^{\mathsf{exp}}(\mathrm{C}_N^{\mathsf{exp}})^2(a_\Delta)^{-1}\varrho^{a_\Delta/2}(1-a_\Delta\beta_\infty^{\mathsf{exp}})^{-1}\kappa_\infty^2,$$
$$\mathrm{C}_{4,1}^{\mathsf{exp}} := d_w\mathrm{C}_{1,w}^{\mathsf{exp}}\mathrm{C}_{22}^{\mathsf{exp}}(\mathrm{C}_N^{\mathsf{exp}})^2 a_\Delta^{-1}\varrho^{a_\Delta}.$$

To estimate the next term we proceed similarly. Using martingale property we obtain

$$\mathrm{Tr}(\mathbb{E}\left[S^{(5)}_{k+1}(S^{(5)}_{k+1})^\top\right]) \le d_\theta \widetilde{m}^{\mathsf{exp}}_{VW}\, \mathrm{C}^2_\infty \,\|A_{12}\|^2 (\mathrm{C}^{\mathsf{exp}}_N)^2 \sum_{\ell=0}^{k} \beta^2_\ell \kappa^2_\ell \prod_{s=\ell+1}^{k} (1-a_\Delta\beta_s)^2(\beta_\ell\gamma_\ell^{-1})^2(1+\mathrm{M}^{\tilde\theta}_\ell+\mathrm{M}^{\tilde w}_\ell)$$

Hence, due to Corollary

$$\mathrm{Tr}(\mathbb{E}\left[S^{(5)}_{k+1}(S^{(5)}_{k+1})^\top\right]) \le \mathrm{C}^{\mathsf{exp}}_{5,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0\beta_{k+1}\kappa^2_{k+1} + \mathrm{C}^{\mathsf{exp}}_{5,1}\beta_{k+1}\kappa^2_{k+1}$$

where

$$\begin{aligned}
\mathrm{C}^{\mathsf{exp}}_{5,0} &:= d_\theta \widetilde{m}^{\mathsf{exp}}_{VW}\, \mathrm{C}^2_\infty \,\|A_{12}\|^2 (\mathrm{C}^{\mathsf{exp}}_N)^2 \varrho^{a_\Delta}(\mathrm{C}^{\mathsf{exp}}_{0,w}+\mathrm{C}^{\mathsf{exp}}_{0,\theta})/(1-a_\Delta\beta^{\mathsf{exp}}_\infty),\\
\mathrm{C}^{\mathsf{exp}}_{5,1} &:= d_\theta \widetilde{m}^{\mathsf{exp}}_{VW}\, \mathrm{C}^2_\infty \,\|A_{12}\|^2 (\mathrm{C}^{\mathsf{exp}}_N)^2 (1+\mathrm{C}^{\mathsf{exp}}_{1,w}\gamma^{\mathsf{mtg}}_\infty+\mathrm{C}^{\mathsf{exp}}_{1,\theta}\beta^{\mathsf{exp}}_\infty)
\end{aligned}$$

It follows from the previous inequalities that

$$\boxed{\mathbb{E}\left[\mathrm{Tr}(\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top)\right] \le \mathrm{C}^{\mathsf{exp}}_{2+5,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0\kappa^2_{k+1} + \mathrm{C}^{\mathsf{exp}}_{2+5,1}\beta_{k+1}\kappa^2_{k+1}} \tag{B.133}$$

where

$$\begin{aligned}
\mathrm{C}^{\mathsf{exp}}_{2+5,0} &:= 2\mathrm{C}^{\mathsf{exp}}_{5,0}\beta^{\mathsf{exp}}_\infty + 2\mathrm{C}^{\mathsf{exp}}_{2,0},\\
\mathrm{C}^{\mathsf{exp}}_{2+5,1} &:= 2\mathrm{C}^{\mathsf{exp}}_{5,1} + 2\mathrm{C}^{\mathsf{exp}}_{2,1}
\end{aligned}$$

For the term $\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top)\right]$ we write

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top)\right] &\le \mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}(S^{(3)}_{k+1})^\top\right]^{1/2}\mathbb{E}\left[\mathrm{Tr}(\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top\right]^{1/2}\\
&\le \Big\{\mathrm{C}^{\mathsf{exp}}_{3,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0\beta_{k+1} + \mathrm{C}^{\mathsf{exp}}_{3,1}\beta_{k+1}\gamma_{k+1} + p_\Delta\varrho^{a_\Delta}\,\mathrm{Tr}(\Sigma)\beta_{k+1}\Big\}^{1/2}\\
&\times\Big\{\mathrm{C}^{\mathsf{exp}}_{2+5,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0\kappa^2_{k+1} + \mathrm{C}^{\mathsf{exp}}_{2+5,1}\beta_{k+1}\kappa^2_{k+1}\Big\}^{1/2}
\end{aligned}$$

We obtain

$$\boxed{\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}\{S^{(2)}_{k+1}+S^{(5)}_{k+1}\}^\top)\right] \le \mathrm{C}^{\mathsf{exp}}_{3/2+5,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}^{\mathsf{exp}}_{3/2+5,1}\beta_{k+1}\kappa_{k+1},} \tag{B.134}$$

where

$$\begin{aligned}
\mathrm{C}^{\mathsf{exp}}_{3/2+5,0} &:= (\mathrm{C}^{\mathsf{exp}}_{2+5,0}+\mathrm{C}^{\mathsf{exp}}_{3,0}+\mathrm{C}^{\mathsf{exp}}_{3,0}\mathrm{C}^{\mathsf{exp}}_{2+5,0})/2,\\
\mathrm{C}^{\mathsf{exp}}_{3/2+5,1} &:= \kappa^{\mathsf{exp}}_\infty/2 + \mathrm{C}^{\mathsf{exp}}_{2+5,1}\beta^{\mathsf{exp}}_\infty\kappa^{\mathsf{exp}}_\infty/2 + (\mathrm{C}^{\mathsf{exp}}_{3,1}\gamma^{\mathsf{mtg}}_\infty + p_\Delta\varrho^{a_\Delta}\,\mathrm{Tr}(\Sigma))\kappa^{\mathsf{exp}}_\infty/2\\
&\quad + (\mathrm{C}^{\mathsf{exp}}_{2+5,1}(\mathrm{C}^{\mathsf{exp}}_{3,1}\gamma^{\mathsf{mtg}}_\infty + p_\Delta\varrho^{a_\Delta}\,\mathrm{Tr}(\Sigma)))^{1/2}
\end{aligned}$$

Let us consider the term

$$\mathbb{E}\left[\mathrm{Tr}(S^{(3)}_{k+1}S^{(4)}_{k+1})^\top\right] = \sum_{j=0}^{k}\beta_j\sum_{\ell=j+1}^{k}\beta_\ell\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}Z_{j+1}\tilde{w}_\ell^\top A_{12}^\top C_\ell^\top N_{\ell+1,k}^\top), \tag{B.135}$$

where $Z_{j+1} := V_{j+1} + A_{12}A_{22}^{-1}W_{j+1}$. For $\tilde{w}_\ell$ we can use the following expansion

$$\tilde{w}_\ell = \widetilde{\Gamma}^{(2)}_{0:\ell-1}\tilde{w}_0 + \sum_{i=0}^{\ell-1}\gamma_i\widetilde{\Gamma}^{(2)}_{i+1:\ell-1}\widetilde{Z}_{i+1} + \sum_{i=0}^{\ell-1}\beta_i\widetilde{\Gamma}^{(2)}_{i+1:\ell-1}C_iA_{12}\tilde{w}_i,$$

where $\widetilde{Z}_{i+1} := W_{i+1} + \kappa_iC_iV_{i+1}$ Substituting this expansion into r.h.s of (B.135) and repeating this procedure until $\mathbb{E}\left[Z_{j+1}\tilde{w}_l^\top\right] = \gamma_j\mathbb{E}\left[Z_{j+1}\widetilde{Z}_{j+1}^\top(\widetilde{\Gamma}^{(2)}_{j+1:\ell_1-1})^\top\right]$ we come to the following expansion of (B.135)

$$\begin{aligned}
&\sum_{\ell=j+1}^{k}\beta_\ell\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}Z_{j+1}\tilde{w}_\ell^\top A_{12}^\top C_\ell^\top N_{\ell+1,k}^\top)\\
&= \sum_{\ell_1=j+1}^{k}\beta_{\ell_1}\sum_{\ell_2=j+1}^{\ell_1-1}\beta_{\ell_2}\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}Z_{j+1}\tilde{w}_{\ell_2}^\top A_{12}^\top C_{\ell_2}^\top(\widetilde{\Gamma}^{(2)}_{\ell_2+1:\ell_1-1})^\top A_{12}^\top C_{\ell_1}^\top N_{\ell_1+1,k}^\top)\\
&+ \gamma_j\sum_{\ell_1=j+1}^{k}\beta_{\ell_1}\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}Z_{j+1}\widetilde{Z}_{j+1}^\top(\widetilde{\Gamma}^{(2)}_{j+1:\ell_1-1})^\top A_{12}^\top C_{\ell_1}^\top N_{\ell_1+1,k}^\top)\\
&= \gamma_j\sum_{s=1}^{k-j}\sum_{\ell_1=j+1}^{k}\beta_{\ell_1}\sum_{\ell_2=1}^{\ell_1-1}\beta_{\ell_2}\ldots\sum_{\ell_s=j+1}^{l_{s-1}-1}\beta_{\ell_s}\\
&\qquad\times\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}Z_{j+1}\widetilde{Z}_{j+1}^\top(\widetilde{\Gamma}^{(2)}_{j+1:\ell_s-1})^\top A_{12}^\top C_{\ell_s}^\top(\widetilde{\Gamma}^{(2)}_{\ell_s+1:\ell_{s-1}-1})^\top A_{12}^\top C_{\ell_s}^\top\ldots(\widetilde{\Gamma}^{(2)}_{\ell_2+1:\ell_1-1})^\top A_{12}^\top C_{\ell_1}^\top N_{\ell_1+1,k}^\top)
\end{aligned}$$

where $\ell_0 := k+1$. Using iteratively Corollary and estimate (B.128) for $N_{\ell_1+1,k}$ we obtain the following bound

$$\begin{aligned}
&\left\|\sum_{\ell=j+1}^{k}\beta_\ell\,\mathrm{Tr}(\Gamma^{(1)}_{j:k-1}Z_{j+1}\tilde{w}_\ell^\top A_{12}^\top C_\ell^\top N_{\ell+1,k}^\top)\right\|\\
&\qquad\le \mathrm{C}_\infty\,\varrho^{a_{22}}\mathrm{C}_N^{\mathsf{exp}}\beta_j\kappa_j\prod_{\ell=j+1}^{k}(1-a_\Delta\beta_\ell)\mathbb{E}\left[\|Z_{j+1}\|^2\right]^{1/2}\mathbb{E}\left[\left\|\widetilde{Z}_{j+1}\right\|^2\right]^{1/2}\sum_{s=1}^{k-j}(\kappa_j\|A_{12}\|\,\mathrm{C}_\infty\,\varrho^{a_{22}})^{s-1}
\end{aligned}$$

Since $\kappa_\infty^{\mathsf{exp}} \le (1/2)(\|A_{12}\|\,\mathrm{C}_\infty\,\varrho^{a_{22}})^{-1}$ and

$$\mathbb{E}\left[\|Z_{j+1}\|^2\right]^{1/2}\mathbb{E}\left[\left\|\widetilde{Z}_{j+1}\right\|^2\right]^{1/2} \le 2\widetilde{m}_{VW}^{\mathsf{exp}}(1+\|A_{12}A_{22}^{-1}\|)(1+\sqrt{\kappa_\infty^{\mathsf{exp}}\,\mathrm{C}_\infty})(1+\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}})$$

we obtain that

$$\begin{aligned}
&\left\|\sum_{\ell=j+1}^{k}\beta_\ell\,\mathrm{Tr}(\widetilde{\Gamma}^{(1)}_{j+1:k}Z_{j+1}\tilde{w}_\ell^\top A_{12}^\top C_\ell^\top N_{\ell+1,k-1}^\top)\right\|\\
&\qquad\le \mathrm{C}_\infty\,\mathrm{C}_N^{\mathsf{exp}}2\widetilde{m}_{VW}^{\mathsf{exp}}(1+\|A_{12}A_{22}^{-1}\|)(1+\sqrt{\kappa_\infty^{\mathsf{exp}}\,\mathrm{C}_\infty})\beta_j\kappa_j\prod_{s=j+1}^{k}(1-a_\Delta\beta_s)(1+\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}})
\end{aligned}$$

This inequality and (B.135) together imply

$$\begin{aligned}&\left|\mathbb{E}\left[\mathrm{Tr}(S_{k+1}^{(3)}S_{k+1}^{(4)})^\top\right]\right| \\ &\qquad\leq \mathrm{C}_\infty\,\mathrm{C}_N^{\mathsf{exp}}2\widetilde{m}_{VW}^{\mathsf{exp}}(1+\|A_{12}A_{22}^{-1}\|)(1+\sqrt{\kappa_\infty^{\mathsf{exp}}\,\mathrm{C}_\infty})\sum_{j=0}^{k}\beta_j^2\kappa_j\prod_{s=j+1}^{k}(1-a_\Delta\beta_s)(1+\mathrm{M}_j^{\tilde{\theta}}+\mathrm{M}_j^{\tilde{w}})\end{aligned}$$

Finally, the standard arguments will lead to

$$\boxed{\left|\mathbb{E}\left[\mathrm{Tr}(S_{k+1}^{(3)}S_{k+1}^{(4)})^\top\right]\right|\leq \mathrm{C}_{3/4,0}^{\mathsf{exp}}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)V_0+\mathrm{C}_{3/4,0}^{\mathsf{exp}}\beta_{k+1}\kappa_{k+1},}\tag{B.136}$$

where

$$\begin{aligned}\mathrm{C}_{3/4,0}^{\mathsf{exp}}&:=\mathrm{C}_\infty\,\mathrm{C}_N^{\mathsf{exp}}2\widetilde{m}_{VW}^{\mathsf{exp}}(1+\|A_{12}A_{22}^{-1}\|)(1+\sqrt{\kappa_\infty^{\mathsf{exp}}\,\mathrm{C}_\infty})\varrho^{a_\Delta/2}(\mathrm{C}_{0,w}^{\mathsf{exp}}+\mathrm{C}_{0,\theta}^{\mathsf{exp}})/(1-a_\Delta\beta_\infty^{\mathsf{exp}})\beta_\infty^{\mathsf{exp}}\kappa_\infty^{\mathsf{exp}},\\ \mathrm{C}_{3/4,1}^{\mathsf{exp}}&:=\mathrm{C}_\infty\,\mathrm{C}_N^{\mathsf{exp}}2\widetilde{m}_{VW}^{\mathsf{exp}}(1+\|A_{12}A_{22}^{-1}\|)(1+\sqrt{\kappa_\infty^{\mathsf{exp}}\,\mathrm{C}_\infty})\varrho^{a_\Delta}(1+\mathrm{C}_{1,w}^{\mathsf{exp}}\gamma_\infty^{\mathsf{mtg}}+\mathrm{C}_{1,\theta}^{\mathsf{exp}}\beta_\infty^{\mathsf{exp}})\end{aligned}$$

Finally, we estimate all terms involving $S_{k+1}^{(6)}$. We rewrite $S_{k+1}^{(6)}$ as follows

$$S_{k+1}^{(6)}=\sum_{\ell=0}^{k}\beta_l M_{\ell,k}W_{\ell+1}$$

where we defined

$$M_{\ell,k}:=\gamma_\ell\beta_\ell^{-1}\sum_{j=\ell+1}^{k}\beta_j\widetilde{\Gamma}_{j+1:k}^{(1)}A_{12}\widetilde{\Gamma}_{\ell+1:j-1}^{(2)}-\widetilde{\Gamma}_{\ell+1:k}^{(1)}A_{12}A_{22}^{-1}.$$

Using martingale property we obtain

$$\mathrm{Tr}(\mathbb{E}\left[S_{k+1}^{(6)}(S_{k+1}^{(6)})^\top\right])\leq\sum_{\ell=0}^{k}\beta_\ell^2\|M_{\ell,k}\|^2\mathbb{E}\left[\|W_{\ell+1}\|^2\right]$$

We rewrite $M_{\ell,k}$ as follows

$$M_{\ell,k}=\sum_{j=\ell+1}^{k}\gamma_j\left[\frac{\beta_j\gamma_l}{\beta_l\gamma_j}\,\mathrm{I}-\widetilde{\Gamma}_{\ell+1:j}^{(1)}\right]\widetilde{\Gamma}_{j+1:k}^{(1)}A_{12}\widetilde{\Gamma}_{\ell+1:j-1}^{(2)}+\widetilde{\Gamma}_{\ell+1:k}^{(1)}A_{12}\left(\sum_{j=l+1}^{k}\gamma_j\widetilde{\Gamma}_{\ell+1:j-1}^{(2)}-A_{22}^{-1}\right)$$

Since,

$$\sum_{j=l+1}^{k}\gamma_j\Gamma_{\ell+1:j-1}^{(2)}=A_{22}^{-1}\left(\mathrm{I}-\widetilde{\Gamma}_{\ell+1:k}^{(2)}\right)$$

this equation leads to

$$M_{\ell,k}=\sum_{j=\ell+1}^{k}\gamma_j\left[\frac{\beta_j\gamma_l}{\beta_l\gamma_j}\,\mathrm{I}-\widetilde{\Gamma}_{\ell+1:j}^{(1)}\right]\widetilde{\Gamma}_{j+1:k}^{(1)}A_{12}\widetilde{\Gamma}_{\ell+1:j-1}^{(2)}-\widetilde{\Gamma}_{\ell+1:k}^{(1)}A_{12}A_{22}^{-1}\widetilde{\Gamma}_{\ell+1:k}^{(2)}$$

We rewrite the term in the square brackets as follows

$$\frac{\beta_j\gamma_l}{\beta_l\gamma_j}\,\mathrm{I}-\widetilde{\Gamma}^{(1)}_{\ell+1:j}=\prod_{s=\ell+1}^{j}\frac{\kappa_s}{\kappa_{s-1}}\,\mathrm{I}-\prod_{s=\ell+1}^{j}(\mathrm{I}-\beta_s\Delta)=\sum_{t=\ell+1}^{j}\frac{\kappa_{t-1}}{\kappa_l}\,\mathrm{I}\left\{\beta_t\Delta+(\kappa_t/\kappa_{t-1}-1)\,\mathrm{I}\right\}\widetilde{\Gamma}^{(1)}_{t+1:j}$$

Using assumption 10 we may show that

$$|\kappa_t/\kappa_{t-1}-1|\le(a_\Delta/16)\beta_t$$

Taking norm of the both sides of the previous equation we obtain

$$\left\|\frac{\beta_j\gamma_l}{\beta_l\gamma_j}\,\mathrm{I}-\widetilde{\Gamma}^{(1)}_{\ell+1:j}\right\|\le\sqrt{p_\Delta}(\|\Delta\|+(a_\Delta/16))\kappa_l^{-1}\sum_{t=\ell+1}^{j}\beta_t\kappa_{t-1}\prod_{s=t+1}^{j}(1-a_\Delta\beta_s)$$

Finally we arrive at the following bound for $M_{\ell,k}$

$$\begin{aligned}\|M_{\ell,k}\|&\le\sqrt{p_{22}p_\Delta}\|A_{12}A_{22}^{-1}\|\prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s)\prod_{s=\ell+1}^{k}(1-a_{22}\gamma_s)\\&+\sqrt{p_\Delta p_{22}}\left(\|\Delta\|+\frac{a_\Delta}{16}\right)\kappa_l^{-1}\sum_{j=\ell+1}^{k}\gamma_j\sum_{t=\ell+1}^{j}\beta_t\kappa_{t-1}\prod_{s=t+1}^{k}(1-a_\Delta\beta_s)\prod_{s=\ell+1}^{j-1}(1-a_{22}\gamma_s)\end{aligned}\tag{B.137}$$

This bound will yield

$$\begin{aligned}\mathrm{Tr}(\mathbb{E}\left[S^{(6)}_{k+1}(S^{(6)}_{k+1})^\top\right])&\le 2p_{22}p_\Delta\|A_{12}A_{22}^{-1}\|^2\sum_{\ell=0}^{k}\beta_\ell^2\prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s)\prod_{s=\ell+1}^{k}(1-a_{22}\gamma_s)\mathbb{E}\left[\|W_{\ell+1}\|^2\right]\\&+2p_\Delta p_{22}(\|\Delta\|+(a_\Delta/16))^2\sum_{\ell=0}^{k}\gamma_\ell^2\mathbb{E}\left[\|W_{\ell+1}\|^2\right]\prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s)\\&\times\left\{\sum_{j=\ell+1}^{k}\gamma_j\prod_{s=\ell+1}^{j-1}(1-(a_{22}/2)\gamma_s)\sum_{t=\ell+1}^{j}\beta_t\kappa_{t-1}\prod_{s=t+1}^{j}(1-a_\Delta\beta_s)\right\}^2\\&=:A_1+A_2\end{aligned}\tag{B.138}$$

The estimate of A_1 follows from Lemma 30

$$A_1\le\mathrm{C}^{\mathsf{exp}}_{6,0,2}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0+\mathrm{C}^{\mathsf{exp}}_{6,1,1}\beta_{k+1}\kappa_{k+1}.\tag{B.139}$$

where

$$\begin{aligned}\mathrm{C}^{\mathsf{exp}}_{6,0,1}&:=2d_w\widetilde{m}^{\mathsf{exp}}_{VW}p_{22}p_\Delta\|A_{12}A_{22}^{-1}\|^2\varrho^{a_{22}}(\mathrm{C}^{\mathsf{exp}}_{0,w}+\mathrm{C}^{\mathsf{exp}}_{0,\theta})\beta^{\mathsf{exp}}_\infty\kappa^{\mathsf{exp}}_\infty/(1-a_\Delta\beta^{\mathsf{exp}}_\infty),\\\mathrm{C}^{\mathsf{exp}}_{6,1,1}&:=2d_w\widetilde{m}^{\mathsf{exp}}_{VW}p_{22}p_\Delta\|A_{12}A_{22}^{-1}\|^2(1+\mathrm{C}^{\mathsf{exp}}_{1,w}\gamma^{\mathsf{mtg}}_\infty+\mathrm{C}^{\mathsf{exp}}_{1,\theta}\beta^{\mathsf{exp}}_\infty)\varrho^{a_{22}}\end{aligned}$$

Inequality (B.129) and Jensen's inequality together imply

$$\begin{aligned}A_2&\le 2(2/a_{22})^2p_\Delta p_{22}(\|\Delta\|+(a_\Delta/16))^2\sum_{\ell=0}^{k}\gamma_\ell^2\mathbb{E}\left[\|W_{\ell+1}\|^2\right]\prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s)\\&\times\sum_{j=\ell+1}^{k}\gamma_j\left\{\sum_{t=\ell+1}^{j}\beta_t\kappa_{t-1}\prod_{s=t+1}^{j}(1-a_\Delta\beta_s)\right\}^2\prod_{s=\ell+1}^{j-1}(1-(a_{22}/2)\gamma_s)\end{aligned}$$

Similarly, applying Jensen's inequality for the second time we come to the following inequality

$$A_2 \le 2(2/a_{22})^2(1/a_\Delta)^2 p_\Delta p_{22}(\|\Delta\| + (a_\Delta/16))^2 \sum_{\ell=0}^{k} \gamma_\ell^2 \mathbb{E}\left[\|W_{\ell+1}\|^2\right] \prod_{s=\ell+1}^{k}(1 - a_\Delta\beta_s)$$
$$\times \sum_{j=\ell+1}^{k} \gamma_j \sum_{t=\ell+1}^{j} \beta_t \kappa_{t-1}^2 \prod_{s=t+1}^{j}(1 - a_\Delta\beta_s) \prod_{s=\ell+1}^{j-1}(1 - (a_{22}/2)\gamma_s)$$

Changing the order of summation we obtain

$$A_2 \le 2(2/a_{22})^2(1/a_\Delta)^2 p_\Delta p_{22}(\|\Delta\| + (a_\Delta/16))^2 \sum_{j=0}^{k} \gamma_j \prod_{s=j+1}^{k}(1 - a_\Delta\beta_s) \sum_{t=0}^{j} \beta_t \kappa_{t-1}^2 \prod_{s=t+1}^{j}(1 - (a_{22}/2)\gamma_s)$$
$$\times \sum_{\ell=0}^{t-1} \gamma_\ell^2 \prod_{s=\ell+1}^{t-1}(1 - (a_{22}/2)\gamma_s)\mathbb{E}\left[\|W_{\ell+1}\|^2\right]$$

Finally, estimating $\mathbb{E}\left[\|W_{\ell+1}\|^2\right]$ by (B.117) and (B.118) and applying Lemma 30 we obtain

$$A_2 \le \mathrm{C}_{6,0,2}^{\mathsf{exp}} \prod_{\ell=0}^{k}(1 - (a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}_{6,1,2}^{\mathsf{exp}} \beta_{k+1}\kappa_{k+1}, \tag{B.140}$$

where

$$\mathrm{C}_{6,0,2}^{\mathsf{exp}} := d_w \widetilde{m}_{VW}^{\mathsf{exp}} (\varrho^{a_{22}/4})^2 \varrho^{a_\Delta/2} 2(2/a_{22})^2 (1/a_\Delta)^2 p_\Delta p_{22}(\|\Delta\| + (a_\Delta/16))^2$$
$$\times (\mathrm{C}_{0,w}^{\mathsf{exp}} + \mathrm{C}_{0,\theta}^{\mathsf{exp}})(\kappa_\infty^{\mathsf{exp}})^2 \gamma_\infty^{\mathsf{mtg}}/(1 - a_\Delta\beta_\infty^{\mathsf{exp}})^2,$$
$$\mathrm{C}_{6,1,2}^{\mathsf{exp}} := d_w \widetilde{m}_{VW}^{\mathsf{exp}} (\varrho^{a_{22}/2})^2 \varrho^{a_\Delta} 2(2/a_{22})^2 (1/a_\Delta)^2 p_\Delta p_{22}(\|\Delta\| + (a_\Delta/16))^2 (1 + \mathrm{C}_{1,w}^{\mathsf{exp}} \gamma_\infty^{\mathsf{mtg}} + \mathrm{C}_{1,\theta}^{\mathsf{exp}} \beta_\infty^{\mathsf{exp}})$$

We conclude from (B.139) and (B.140) that

$$\boxed{\mathrm{Tr}(\mathbb{E}\left[S_{k+1}^{(6)}(S_{k+1}^{(6)})^\top\right]) \le \mathrm{C}_{6,0}^{\mathsf{exp}} \prod_{\ell=0}^{k}(1 - (a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}_{6,1}^{\mathsf{exp}} \beta_{k+1}\kappa_{k+1},} \tag{B.141}$$

where

$$\mathrm{C}_{6,0}^{\mathsf{exp}} := \mathrm{C}_{6,0,1}^{\mathsf{exp}} + \mathrm{C}_{6,0,2}^{\mathsf{exp}},$$
$$\mathrm{C}_{6,1}^{\mathsf{exp}} := \mathrm{C}_{6,1,1}^{\mathsf{exp}} + \mathrm{C}_{6,1,2}^{\mathsf{exp}} \tag{B.142}$$

It remains to consider $\mathbb{E}\left[\mathrm{Tr}(S_{k+1}^{(3)}(S_{k+1}^{(6)})^\top)\right]$. We proceed similarly and use (B.137) to get

$$\left|\mathbb{E}\left[\mathrm{Tr}(S_{k+1}^{(3)}(S_{k+1}^{(6)})^\top)\right]\right| \le \sum_{\ell=0}^{k} \beta_\ell^2 \|M_{\ell,k}\| \prod_{s=\ell+1}^{k}(1 - a_\Delta\beta_s)\mathbb{E}\left[\|Z_{\ell+1}\|^2\right]^{1/2} \mathbb{E}\left[\|W_{\ell+1}\|^2\right]^{1/2}$$
$$= A_1' + A_2'$$

The following estimate holds for A_1'

$$|A_1'| \le \sqrt{p_{22}p_\Delta}\|A_{12}A_{22}^{-1}\| \max(d_\theta, d_w)(2 + 2\|A_{12}A_{22}^{-1}\|)^{1/2} \sum_{\ell=0}^{k} \beta_\ell^2 \prod_{s=\ell+1}^{k}(1 - a_\Delta\beta_s)$$
$$\times \prod_{s=\ell+1}^{k}(1 - a_{22}\gamma_s)(1 + \mathrm{M}_\ell^{\tilde{\theta}} + \mathrm{M}_\ell^{\tilde{w}})$$

Applying standard arguments we get

$$A_1' \le \mathrm{C}^{\mathsf{exp}}_{3/6,0,1} \prod_{\ell=0}^{k} (1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}^{\mathsf{exp}}_{3/6,1,1}\beta_{k+1}\kappa_{k+1}, \tag{B.143}$$

where

$$\begin{aligned}
\mathrm{C}^{\mathsf{exp}}_{3/6,0,1} &:= \widetilde{m}^{\mathsf{exp}}_{VW}\sqrt{p_{22}p_\Delta}\|A_{12}A_{22}^{-1}\|\max(d_\theta,d_w)(2+2\|A_{12}A_{22}^{-1}\|)^{1/2} \\
&\qquad \times (\mathrm{C}^{\mathsf{exp}}_{0,w}+\mathrm{C}^{\mathsf{exp}}_{0,\theta})\varrho^{a_{22}}\beta^{\mathsf{exp}}_\infty\kappa^{\mathsf{exp}}_\infty/(1-a_\Delta\beta^{\mathsf{exp}}_\infty), \\
\mathrm{C}^{\mathsf{exp}}_{3/6,1,1} &:= 2\widetilde{m}^{\mathsf{exp}}_{VW}\sqrt{p_{22}p_\Delta}\|A_{12}A_{22}^{-1}\|\max(d_\theta,d_w)(2+2\|A_{12}A_{22}^{-1}\|)^{1/2}(1+\mathrm{C}^{\mathsf{exp}}_{1,w}\gamma^{\mathsf{mtg}}_\infty+\mathrm{C}^{\mathsf{exp}}_{1,\theta}\beta^{\mathsf{exp}}_\infty)\varrho^{a_{22}}
\end{aligned}$$

For A_2' we write the following bound

$$\begin{aligned}
A_2' &\le \sqrt{p_\Delta p_{22}}(\|\Delta\|+(a_\Delta/16))\max(d_\theta,d_w)(2+2\|A_{12}A_{22}^{-1}\|)^{1/2}\sum_{\ell=0}^{k}\gamma_\ell^2(1+\mathrm{M}^{\tilde\theta}_\ell+\mathrm{M}^{\tilde w}_\ell)\prod_{s=\ell+1}^{k}(1-a_\Delta\beta_s) \\
&\times\Big\{\sum_{j=\ell+1}^{k}\gamma_j\prod_{s=\ell+1}^{j-1}(1-a_{22}\gamma_s)\sum_{t=\ell+1}^{j}\beta_t\kappa_{t-1}\prod_{s=t+1}^{j}(1-a_\Delta\beta_s)\Big\}
\end{aligned}$$

Changing the order of summation and applying arguments from the estimation of A_2 we come to the following bound

$$A_2' \le \mathrm{C}^{\mathsf{exp}}_{3/6,0,2}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}^{\mathsf{exp}}_{3/6,1,2}\beta_{k+1}\kappa_{k+1}, \tag{B.144}$$

where

$$\begin{aligned}
\mathrm{C}^{\mathsf{exp}}_{3/6,0,2} &:= \sqrt{p_\Delta p_{22}}(\|\Delta\|+(a_\Delta/16))\max(d_\theta,d_w)(2+2\|A_{12}A_{22}^{-1}\|)^{1/2} \\
&\qquad \times(\mathrm{C}^{\mathsf{exp}}_{0,w}+\mathrm{C}^{\mathsf{exp}}_{0,\theta})(\varrho^{a_{22}/4})^2\varrho^{a_\Delta/2}\gamma^{\mathsf{mtg}}_\infty(\kappa^{\mathsf{exp}}_\infty)^2/(1-a_\Delta\beta^{\mathsf{exp}}_\infty)^2, \\
\mathrm{C}^{\mathsf{exp}}_{3/6,1,2} &:= \sqrt{p_\Delta p_{22}}(\|\Delta\|+(a_\Delta/16))\max(d_\theta,d_w)(2+2\|A_{12}A_{22}^{-1}\|)^{1/2}(1+\mathrm{C}^{\mathsf{exp}}_{1,w}\gamma^{\mathsf{mtg}}_\infty+ \\
&\quad +\mathrm{C}^{\mathsf{exp}}_{1,\theta}\beta^{\mathsf{exp}}_\infty)(\varrho^{a_{22}/2})^2\varrho^{a_\Delta}
\end{aligned} \tag{B.145}$$

We conclude from (B.143) and (B.144)

$$\boxed{\big|\operatorname{Tr}(\mathbb{E}\Big[S^{(3)}_{k+1}(S^{(6)}_{k+1})^\top\big)\big|\Big] \le \mathrm{C}^{\mathsf{exp}}_{3/6,0}\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}^{\mathsf{exp}}_{3/6,1}\beta_{k+1}\kappa_{k+1},} \tag{B.146}$$

where

$$\begin{aligned}
\mathrm{C}^{\mathsf{exp}}_{3/6,0} &:= \mathrm{C}^{\mathsf{exp}}_{3/6,0,1}+\mathrm{C}^{\mathsf{exp}}_{3/6,0,2}, \\
\mathrm{C}^{\mathsf{exp}}_{3/6,1} &:= \mathrm{C}^{\mathsf{exp}}_{3/6,1,1}+\mathrm{C}^{\mathsf{exp}}_{3/6,1,2}
\end{aligned}$$

Final estimate of the remainder term J_{k+1} Collecting (B.126), (B.127),(B.132), (B.133), (B.134), (B.136), (B.141), (**??**) we obtain the estimate of $J_{k+1} := R_{k+1} + J'_{k+1}$

$$\boxed{|J_{k+1}| \le \mathrm{C}^{\mathsf{exp}}_0\prod_{\ell=0}^{k}(1-(a_\Delta/4)\beta_\ell)\mathrm{V}_0 + \mathrm{C}^{\mathsf{exp}}_0\beta_{k+1}(\gamma_{k+1}+\kappa_{k+1}),}$$

where

$$\mathrm{C}_0^{\mathsf{exp}} := \mathrm{C}_{3,0}^{\mathsf{exp}}\beta_\infty^{\mathsf{exp}} + 3\mathrm{C}_{0/1}^{\mathsf{exp}} + 5\mathrm{C}_{4,0}^{\mathsf{exp}} + 5\mathrm{C}_{2+5,0}^{\mathsf{exp}}(\kappa_\infty^{\mathsf{exp}})^2 + 2\mathrm{C}_{3/2+5,0}^{\mathsf{exp}} + 2\mathrm{C}_{3/4,0}^{\mathsf{exp}} \tag{B.147}$$

$$+ 5\mathrm{C}_{6,0}^{\mathsf{exp}} + 2\mathrm{C}_{3/6,0}^{\mathsf{exp}}, \tag{B.148}$$

$$\mathrm{C}_1^{\mathsf{exp}} := \mathrm{C}_{3,1}^{\mathsf{exp}} + 5\mathrm{C}_{4,1}^{\mathsf{exp}} + 5\mathrm{C}_{2+5,1}^{\mathsf{exp}} + 2\mathrm{C}_{3/2+5,1}^{\mathsf{exp}} + 2\mathrm{C}_{3/4,1}^{\mathsf{exp}} + 5\mathrm{C}_{6,1}^{\mathsf{exp}} + 2\mathrm{C}_{3/6,1}^{\mathsf{exp}} \tag{B.149}$$

Hence, we obtained (2.41).

B.4 Auxiliary Lemmas

Lemma 29. *Let $a > 0$ and $(\gamma_k)_{k\geq 0}$ be a nonincreasing sequence such that $\gamma_0 < 1/a$. Then, for any integer $k \geq 1$,*

$$\sum_{j=0}^{k-1} \gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) = \frac{1}{a}\left\{1 - \prod_{l=0}^{k-1}(1-\gamma_l a)\right\}$$

Remark 1. If k_0 is such that $\sum_{l=0}^{k_0-1}\gamma_l \geq \log(2)/a$ then the r.h.s. of the previous equation is lower bounded by $1/(2a)$ for any $k \geq k_0$.

Proof. Let us denote $u_{j:k-1} = \prod_{l=j}^{k-1}(1-\gamma_l a)$. Then, for $j \in \{0,\dots,k-1\}$, $u_{j+1:k-1} - u_{j:k-1} = a\gamma_j u_{j+1}$. Hence,

$$\sum_{j=0}^{k-1} \gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) = \frac{1}{a}\sum_{j=0}^{k-1}(u_{j+1:k-1} - u_{j:k-1}) = a^{-1}(1-u_{0:k-1})\,.$$

□

Lemma 30. *Assume A10 and set*

$$\varrho^a = \frac{2}{a}\varsigma \max\{1, a_{22}/(4a_\Delta)\} \vee \frac{4}{a}\varsigma^3. \tag{B.150}$$

The following holds

1. *For any $a \in [a_{22}/4, \gamma_0^{-1}]$ and $k \in \mathbb{N}$, if in addition, we have $\kappa \leq a_{22}/(4a_\Delta)$, then*

$$\sum_{j=0}^{k-1}\gamma_j^2 \prod_{l=j+1}^{k-1}(1-\gamma_l a) \leq \varrho^a\gamma_k,\ \sum_{j=0}^{k-1}\beta_j\gamma_j \prod_{l=j+1}^{k-1}(1-\gamma_l a) \leq \varrho^a\beta_k,\ \sum_{j=0}^{k-1}\beta_j^2 \prod_{l=j+1}^{k-1}(1-\gamma_l a) \leq \varrho^a\beta_k$$

2. *For any $a \in [a_\Delta/8, \beta_0^{-1}]$ and $k \in \mathbb{N}$,*

$$\sum_{j=0}^{k-1}\beta_j\gamma_j \prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varrho^a\gamma_k, \quad \sum_{j=0}^{k-1}\beta_j^2 \prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varrho^a\beta_k$$

3. *For any $a \in [a_\Delta/4, \beta_0^{-1}]$ and $k \in \mathbb{N}$,*

$$\sum_{j=0}^{k-1}\beta_j^3/\gamma_j \prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varrho^a\beta_k^2/\gamma_k, \quad \sum_{j=0}^{k-1}\beta_j^4/\gamma_j^2 \prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varrho^a\beta_k^3/\gamma_k^2,$$

$$\sum_{j=0}^{k-1}\beta_j^2\gamma_j \prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varrho^a\beta_k\gamma_k$$

4. For any $a \in [a_{22}/4, \gamma_0^{-1}]$ and $k \in \mathbb{N}$,

$$\sum_{j=0}^{k-1} \beta_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varrho^a \beta_k/\gamma_k, \quad \sum_{j=0}^{k-1} \beta_j^2 \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varrho^a \beta_k^2/\gamma_k,$$
$$\sum_{j=0}^{k-1} \beta_j^3/\gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varrho^a \beta_k^3/\gamma_k^2$$

Proof. Part i) of the corollary, consider the first inequality and observe that

$$\sum_{j=0}^{k-1} \gamma_j^2 \prod_{l=j+1}^{k-1} (1-\gamma_l a) = \gamma_k \sum_{j=0}^{k-1} \frac{\gamma_{k-1}}{\gamma_k} \frac{\gamma_j}{\gamma_{k-1}} \gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varsigma \gamma_k \sum_{j=0}^{k-1} \gamma_j \prod_{l=j+1}^{k-1} \frac{\gamma_{l-1}}{\gamma_l} (1-\gamma_l a)$$

Note that as $a \geq a_{22}/4$, we have

$$\frac{\gamma_{l-1}}{\gamma_l}(1-\gamma_l a) \leq (1+\frac{a_{22}}{8}\gamma_l)(1-\gamma_l a) \leq 1 - a\gamma_l/2$$

Substituting into the above inequality yields

$$\sum_{j=0}^{k-1} \gamma_j^2 \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varsigma \gamma_k \sum_{j=0}^{k-1} \gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a/2) \leq \varsigma \frac{2}{a} \gamma_k$$

where we have applied Lemma 29 in the last inequality. Next and applying similar steps as before, we observe that

$$\sum_{j=0}^{k-1} \beta_j \gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varsigma \beta_k \sum_{j=0}^{k-1} \gamma_j \prod_{l=j+1}^{k-1} \frac{\beta_{l-1}}{\beta_l} (1-\gamma_l a)$$

As we have

$$\frac{\beta_{l-1}}{\beta_l}(1-\gamma_l a) \leq 1 - \gamma_l(a - \kappa a_\Delta/16) \leq 1 - \gamma_l(a - a_{22}/64) \leq 1 - \gamma_l a/2$$

we obtain

$$\sum_{j=0}^{k-1} \beta_j \gamma_j \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varsigma \frac{2}{a} \beta_k$$

Similarly, using $\beta_j \leq \kappa \gamma_j \leq a_{22}/(4a_\Delta)\gamma_j$, we obtain

$$\sum_{j=0}^{k-1} \beta_j^2 \prod_{l=j+1}^{k-1} (1-\gamma_l a) \leq \varsigma \frac{a_{22}}{2a\, a_\Delta} \beta_k$$

For part ii) of the corollary, we observe that the first inequality can be proven by:

$$\sum_{j=0}^{k-1} \beta_j \gamma_j \prod_{\ell=j+1}^{k-1} (1-a\beta_\ell) = \gamma_k \sum_{j=0}^{k-1} \beta_j \frac{\gamma_{k-1}}{\gamma_k} \frac{\gamma_j}{\gamma_{k-1}} \prod_{\ell=j+1}^{k-1} (1-a\beta_\ell)$$
$$\leq \varsigma \gamma_k \sum_{j=0}^{k-1} \beta_j \frac{\gamma_j}{\gamma_{k-1}} \prod_{\ell=j+1}^{k-1} (1-a\beta_\ell) \leq \varsigma \gamma_k \sum_{j=0}^{k-1} \beta_j \prod_{\ell=j+1}^{k-1} \frac{\gamma_{\ell-1}}{\gamma_\ell} (1-a\beta_\ell)$$

Note that as $a \geq a_\Delta/8$, we have

$$\frac{\gamma_{\ell-1}}{\gamma_\ell}(1-a\beta_\ell) \leq (1+\epsilon_\beta\beta_\ell)(1-a\beta_\ell) \leq 1-a\beta_\ell/2$$

Using Lemma 29, this yields

$$\sum_{j=0}^{k-1}\beta_j\gamma_j\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varsigma\gamma_k\sum_{j=0}^{k-1}\beta_j\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell/2) \leq \frac{2}{a}\varsigma\,\gamma_k$$

Similarly, the second inequality is:

$$\begin{aligned}&\sum_{j=0}^{k-1}\beta_j^2\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varsigma\beta_k\sum_{j=0}^{k-1}\beta_j\frac{\beta_j}{\beta_{k-1}}\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell)\\ &\leq \varsigma\beta_k\sum_{j=0}^{k-1}\beta_j\prod_{\ell=j+1}^{k-1}\frac{\beta_{\ell-1}}{\beta_\ell}(1-a\beta_\ell) \overset{(a)}{\leq} \varsigma\beta_k\sum_{j=0}^{k-1}\beta_j\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell/2) \overset{(b)}{\leq} \frac{2}{a}\varsigma\beta_k\end{aligned}$$

where (a) is due to the fact that we have $\frac{\beta_{\ell-1}}{\beta_\ell}(1-a\beta_\ell) \leq 1-a\beta_\ell/2$, and (b) is obtained by applying Lemma 29.

For the proof of part iii) we proceed similarly. We prove the second inequality only. The proof of the remaining follows the same lines. Denote $\kappa_\ell := \beta_\ell/\gamma_\ell$. Clearly, $\kappa_{\ell-1}/\kappa_\ell \leq \beta_{\ell-1}/\beta_\ell \leq \varsigma$. Then

$$\begin{aligned}&\sum_{j=0}^{k-1}\beta_j^2\kappa_j^2\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell) \leq \varsigma^3\beta_k\kappa_k^2\sum_{j=0}^{k-1}\beta_j\frac{\beta_j}{\beta_{k-1}}\left(\frac{\kappa_j}{\kappa_{k-1}}\right)^2\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell)\\ &\leq \varsigma^3\beta_k\kappa_k^2\sum_{j=0}^{k-1}\beta_j\prod_{\ell=j+1}^{k-1}\left(\frac{\beta_{\ell-1}}{\beta_\ell}\right)^3(1-a\beta_\ell) \overset{(a)}{\leq} \varsigma^3\beta_k\kappa_k^2\sum_{j=0}^{k-1}\beta_j\prod_{\ell=j+1}^{k-1}(1-a\beta_\ell/2) \overset{(b)}{\leq} \frac{4}{a}\varsigma^3\beta_k\kappa_k^2\end{aligned}$$

where (a) is due to the fact that we have $(\frac{\beta_{\ell-1}}{\beta_\ell})^3(1-a\beta_\ell) \leq 1-a\beta_\ell/4$, and (b) is obtained by applying Lemma 29. Part iv) may be proved in the similar way. □

Lemma 31. *Let $a > 0$, $p \geq 0$, $(\gamma_j)_{j\geq 0}$, $(\kappa_j)_{j\geq 0}$ and $(u_j)_{j\geq 0}$ be nonnegative sequences. Then, for all integers k,*

$$\sum_{j=0}^{k-1}\kappa_j\prod_{\ell=j}^{k-1}(1-a\gamma_\ell)\sum_{i=0}^{j-1}\gamma_i^p\prod_{n=i+1}^{j-1}(1-a\gamma_n)u_i = \sum_{i=0}^{k-1}\gamma_i^p u_i\left(\sum_{j=i+1}^{k-1}\kappa_j\right)\prod_{\ell=i+1}^{k-1}(1-a\gamma_\ell)$$

Lemma 32 (Lyapunov Lemma)**.** *A matrix A is Hurwitz if and only if for any positive symmetric matrix $P = P^\top \succ 0$ there is $Q = Q^\top \succ 0$ that satisfies the Lyapunov equation*

$$A^\top Q + QA = -P\,.$$

In addition, Q is unique.

Proof. See [61, Lemma 9.1, p. 140]. □

Lemma 33. *Assume that $-A$ is a Hurwitz matrix. Let Q be the unique solution of the Lyapunov equation*

$$A^\top Q + QA = \mathrm{I}\,.$$

Then, for any $\zeta \in [0, \zeta_A]$, where

$$\zeta_A := (1/2)\|A\|_Q^{-2}\|Q\|^{-1}, \tag{B.151}$$

we get

$$\|\mathrm{I} - \zeta A\|_Q^2 \le (1 - a\zeta) \quad \textit{with} \quad a = (1/2)\|Q\|^{-1}\,.$$

If in addition $\zeta \le \|Q\|$ then

$$1 - a\zeta \ge 1/2.$$

Proof. For any $x \in \mathbb{R}^d$, we get

$$\frac{x^\top(\mathrm{I} - \gamma A)^\top Q(\mathrm{I} - \gamma A)x}{x^\top Q x} = 1 - \gamma\frac{\|x\|^2}{x^\top Q x} + \gamma^2\frac{x^\top A^\top QAx}{x^\top Q x}$$

Hence, we get that for all $\gamma \in [0, (1/2)\|A\|_Q^{-2}\|Q\|^{-1}]$,

$$\begin{aligned} 1 - \gamma\frac{\|x\|^2}{x^\top Q x} + \gamma^2\frac{x^\top A^\top QAx}{x^\top Q x} &\le 1 - \gamma\|Q\|^{-1} + \gamma^2\|A\|_Q^2 \\ &\le 1 - (1/2)\|Q\|^{-1}\gamma\,. \end{aligned}$$

The proof follows. □

Lemma 34. *Assume that $\|L\|_{Q_\Delta, Q_{22}} \le \varepsilon$ for some $\varepsilon > 0$ and*

$$0 \le \beta \le (1/2)\{\|\Delta\|_{Q_\Delta} + \varepsilon\|A_{12}\|_{Q_{22}, Q_\Delta}\}^{-1} \tag{B.152}$$

$$0 \le \gamma \le (1/2)\|Q_{22}\|^{-1}\|A\|_{Q_{22}}^{-2}\,. \tag{B.153}$$

Set $B_{11}(L) = \Delta - A_{12}L$. Then, the equation

$$L'\{\mathrm{I} - \beta B_{11}(L)\} = (\mathrm{I} - \gamma A_{22})L + \beta A_{22}^{-1}A_{21}B_{11}(L) \tag{B.154}$$

has a unique solution satisfying

$$\|L'\|_{Q_\Delta, Q_{22}} \le (1 - \gamma a_{22})\|L\|_{Q_\Delta, Q_{22}} + \beta\,\mathrm{C}_D(\varepsilon)$$

where

$$\mathrm{C}_D(\varepsilon) = 2\{\|A_{22}^{-1}A_{21}\|_{Q_\Delta, Q_{22}} + \varepsilon\}\{\|\Delta\|_{Q_\Delta} + \varepsilon\|A_{12}\|_{Q_{22}, Q_\Delta}\}\,. \tag{B.155}$$

If $\beta/\gamma \le \varepsilon a_{22}/\,\mathrm{C}_D(\varepsilon)$, then $\|L'\|_{Q_\Delta, Q_{22}} \le \varepsilon$.

Proof. Since $\|L\|_{Q_\Delta, Q_{22}} \le \varepsilon$, we get that $\|B_{11}(L)\|_{Q_\Delta} \le \|\Delta\|_{Q_\Delta} + \varepsilon\|A_{12}\|_{Q_{22}, Q_\Delta}$. Hence, using (B.152) and the triangular inequality, we get that $\beta\|B_{11}(L)\|_{Q_\Delta} \le 1/2$ and thus

$$\|\mathrm{I} - \beta B_{11}(L)\|_{Q_\Delta} \ge 1/2\,. \tag{B.156}$$

Hence, $\mathrm{I} - \beta B_{11}(L)$ is invertible and (B.154) has a unique solution given by

$$\begin{aligned} L' &= \left\{(\mathrm{I} - \gamma A_{22})L + \beta A_{22}^{-1}A_{21}B_{11}(L)\right\}\{\mathrm{I} - \beta B_{11}(L)\}^{-1} \\ &= (\mathrm{I} - \gamma A_{22})L + \beta D(L) \end{aligned}$$

where

$$D(L) = \{A_{22}^{-1}A_{21} + (\mathrm{I} - \gamma A_{22})L\}B_{11}(L)\{\mathrm{I} - \beta B_{11}(L)\}^{-1}\,.$$

Using (B.156) and $\|L\|_{Q_\Delta,Q_{22}} \le \varepsilon$, we get that $\|D(L)\|_{Q_\Delta,Q_{22}} \le \mathrm{C}_D(\varepsilon)$. Hence, for γ satisfying (B.153), we get that

$$\|L'\|_{Q_\Delta,Q_{22}} \le (1-\gamma a_{22})\|L\|_{Q_\Delta,Q_{22}} + \beta\,\mathrm{C}_D(\varepsilon) \le \varepsilon + \gamma\big(\frac{\beta}{\gamma}\,\mathrm{C}_D(\varepsilon) - a_{22}\varepsilon\big) \le \varepsilon\,,$$

where the last inequality is due to $\frac{\beta}{\gamma} \le \varepsilon a_{22}/\,\mathrm{C}_D(\varepsilon)$. □

Lemma 35. *Let $L_0 = 0$. Assume that $\|L_k\|_{Q_\Delta,Q_{22}} \le L_\infty$ and*

$$0 \le \beta_0 \le (1/2)\{\|\Delta\|_{Q_\Delta} + L_\infty\|A_{12}\|_{Q_{22},Q_\Delta}\}^{-1}$$
$$0 \le \gamma_0 \le (1/2)\|Q_{22}\|^{-1}\|A\|_{Q_{22}}^{-2}$$

Then for any $k \in \mathbb{N}$

$$\|L_k\|_{Q_\Delta,Q_{22}} \le \mathrm{C}_D(L_\infty)\varrho^{a_{22}}\beta_k/\gamma_k,$$

where

$$\mathrm{C}_D(L_\infty) := 2\{\|A_{22}^{-1}A_{21}\|_{Q_\Delta,Q_{22}} + L_\infty\}\{\|\Delta\|_{Q_\Delta} + L_\infty\|A_{12}\|_{Q_{22},Q_\Delta}\}$$

Proof. Similarly to Lemma 34 we may show that

$$L_{k+1} = (\mathrm{I} - \gamma A_{22})L_k + \beta_k D(L_k)$$

where $\|D(L_k)\|_{Q_\Delta,Q_{22}} \le \mathrm{C}_D(L_\infty)$. Hence,

$$\|L_k\|_{Q_\Delta,Q_{22}} \le \mathrm{C}_D(L_\infty)\sum_{j=0}^{k}\beta_j\prod_{s=j+1}^{k}(1-a_{22}\gamma_s)$$

Application of Lemma 30 to the right hand side of the above completes the proof. □

Lemma 36. *Let $L_1 := L_0 := 0$. Assume that $\|L_k\|_{Q_\Delta,Q_{22}} \le L_\infty$ and*

$$0 \le \beta_0 \le (1/2)\{\|\Delta\|_{Q_\Delta} + L_\infty\|A_{12}\|_{Q_{22},Q_\Delta}\}^{-1}$$
$$0 \le \gamma_0 \le (1/2)\|Q_{22}\|^{-1}\|A\|_{Q_{22}}^{-2}$$
$$\beta_{k-1} - \beta_k \le \rho_\beta\beta_k^2, \gamma_{k-1} - \gamma_k \le \rho_\gamma\gamma_k^2$$
$$\beta_k/\gamma_k \le (1/(2C_1^U))a_{22}$$

with

$$C_1^U := 2(\|\Delta\|_{Q_\Delta} + \|A_{22}^{-1}A_{21}\|_{Q_\Delta,Q_{22}}\|A_{12}\|_{Q_{22},Q_\Delta} + 2L_\infty\|A_{12}\|_{Q_{22},Q_\Delta}).$$

Then

$$\|L_{k+1} - L_k\|_{Q_\Delta,Q_{22}} \le C_2^U\varrho^{a_{22}/2}\gamma_{k+1},$$

where

$$C_2^U := 2\rho_\gamma L_\infty\|A_{22}\|_{Q_{22}} + 2\rho_\beta(L_\infty + \|A_{22}^{-1}A_{21}\|_{Q_\Delta,Q_{22}})(\|\Delta\|_{Q_\Delta} + L_\infty\|A_{12}\|_{Q_{22},Q_\Delta})$$

Proof. Recall that $B_{11}(L) = \Delta - A_{12}L$. It follows from Lemma 34 that $\mathrm{I} - \beta_k B_{11}(L_k)$ is invertible matrix with bounded norm. Equation

$$L_k(\mathrm{I} - \beta_{k-1}B_{11}(L_{k-1})) = \left\{(\mathrm{I} - \gamma_{k-1}A_{22})L_{k-1} + \beta_{k-1}A_{22}^{-1}A_{21}B_{11}(L_{k-1})\right\}$$

may be rewritten as follows

$$L_k(\mathrm{I} - \beta_k B_{11}(L_k)) = (\mathrm{I} - \gamma_k A_{22})L_{k-1} + \beta_k B_{11}(L_k) + E_k, \tag{B.157}$$

where $E_k := (\gamma_k - \gamma_{k-1})A_{22}L_{k-1} + (L_k + A_{22}^{-1}A_{21})D_k$, $D_k := -\beta_k A_{12}(L_k - L_{k-1}) + (\beta_k - \beta_{k-1})B_{11}(L_{k-1})$. Let $U_k = L_k - L_{k-1}$. Then

$$U_{k+1}(\mathrm{I} - \beta_k B_{11}(L_k)) = (\mathrm{I} - \gamma_k A_{22})U_k - E_k.$$

Then

$$U_{k+1} = (\mathrm{I} - \gamma_k A_{22})U_k + \beta_k(\mathrm{I} - \gamma_k A_{22})U_k B_{11}(L_k)(\mathrm{I} - \beta_k B_{11}(L_k))^{-1} - E_k(\mathrm{I} - \beta_k B_{11}(L_k))^{-1}$$

It is easy to check that

$$\|(\mathrm{I} - \gamma_k A_{22})U_k B_{11}(L_k)(\mathrm{I} - \beta_k B_{11}(L_k))^{-1}\|_{Q_\Delta, Q_{22}} \le 2\|U_k\|_{Q_\Delta, Q_{22}}\{\|\Delta\|_{Q_\Delta} + L_\infty \|A_{12}\|_{Q_{22}, Q_\Delta}\}$$

Moreover,

$$\|E_k(\mathrm{I} - \beta_k B_{11}(L_k))^{-1}\|_{Q_\Delta, Q_{22}} \le 2\rho_\gamma \gamma_k^2 L_\infty \|A_{22}\|_{Q_{22}}$$
$$+ 2(L_\infty + \|A_{22}^{-1}A_{21}\|_{Q_\Delta, Q_{22}})\{\rho_\beta \beta_k^2(\|\Delta\|_{Q_\Delta} + L_\infty \|A_{12}\|_{Q_{22}, Q_\Delta}) + \beta_k \|A_{12}\|_{Q_{22}, Q_\Delta}\|U_k\|_{Q_\Delta, Q_{22}}\}$$

Applying previous inequalities we obtain

$$\|U_{k+1}\|_{Q_\Delta, Q_{22}} \le (1 - \gamma_k a_{22} + C_1^U \beta_k)\|U_k\|_{Q_\Delta, Q_{22}} + C_2^U \gamma_k^2$$

Since $\beta_k/\gamma_k \le (1/(2C_1^U))a_{22}$ we obtain

$$\|U_{k+1}\|_{Q_\Delta, Q_{22}} \le C_2^U \varrho^{a_{22}/2}\gamma_{k+1}$$

□

Lemma 37. *Let Q be a symmetric definite positive $n \times n$ matrix and Σ be a $n \times n$ matrix. Then*

$$\mathrm{Tr}(Q\Sigma) \le \|\Sigma\|_Q \,\mathrm{Tr}(Q)\,.$$

Proof. Denote by $(e_i)_{i=1}^n$ an orthonomal basis of eigenvectors of Q, $Qe_i = \lambda_i(Q)e_i$, $i = 1, \ldots, n$, $\langle e_i, e_j\rangle = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker symbol. We get that

$$\mathrm{Tr}(Q\Sigma) = \sum_{i=1}^n \langle e_i, Q\Sigma e_i\rangle = \sum_{i=1}^n \langle e_i, \Sigma e_i\rangle_Q$$
$$\le \|\Sigma\|_Q \sum_{i=1}^n \|e_i\|_{[}Q]^2 = \|\Sigma\|_Q \,\mathrm{Tr}\, Q$$

where we have used $\|e_i\|_{[}Q] = \lambda_i$ and $\mathrm{Tr}\, Q = \sum_{i=1}^n \lambda_i(Q)$. □

Corollary 38. *If X is a $n \times 1$ random vector such that $\mathbb{E}[\|X\|_{[} 2]] < \infty$. Then,*

$$\mathbb{E}\left[[]\, \|X\|_{[} Q]^2\right] \leq \mathrm{Tr}(Q) \|\mathbb{E}\left[[]\, XX^\top\right]\|_Q \,.$$

Proof. Note that $\mathbb{E}\left[[]\, \|X\|_{[} Q]^2\right] = \mathrm{Tr}(Q\mathbb{E}\left[[]\, XX^\top\right]) \leq \|\mathbb{E}\left[[]\, XX^\top\right]\|_Q \,\mathrm{Tr}\, Q$ □

Lemma 39. *Let m and n be two integers, P and Q be $m \times m$ and $n \times n$ symmetric positive definite matrices. Let X and Y be $m \times 1$ and $n \times 1$ random vectors such that $\mathbb{E}[\|X\|^2] < \infty$ and $\mathbb{E}[\|Y\|^2] < \infty$. Then,*

$$\|\mathbb{E}[XY^\top]\|_{Q,P} \leq \lambda_{\min}(Q)^{-1}\{\mathrm{Tr}(Q)\}^{1/2}\{\mathrm{Tr}(P)\}^{1/2}\|\mathbb{E}[XX^\top]\|_P^{1/2}\|\mathbb{E}[YY^\top]\|_Q^{1/2}$$

Proof. Note that $\|\mathbb{E}[XY^\top]\|_{Q,P} \leq \mathbb{E}[\|XY^\top\|_{Q,P}]$ and

$$\begin{aligned}\|XY^\top\|_{Q,P} &= \sup_{\|y\|_Q=1} \|X\langle Y, y\rangle_Q\| = \|X\|_{[} P] \sup_{\|y\|_Q=1} \langle Q^{-1}Y, y\rangle_Q \\ &= \|X\|_P \left\|Q^{-1}Y\right\|_Q = \|X\|_P \|Y\|_{Q^{-1}} \leq \lambda_{\min}^{-1}(Q)\, \|X\|_P \|Y\|_Q \,.\end{aligned}$$

By applying the Cauchy-Schwarz inequality, we obtain

$$\|\mathbb{E}[XY^\top]\|_{Q,P} \leq \lambda_{\min}^{-1}(Q) \left\{\mathbb{E}[\|X\|_P^2]\right\}^{1/2} \left\{\mathbb{E}[\|Y\|_Q^2]\right\}^{1/2} .$$

The proof follows from Corollary 38. □

B.5 Details on Numerical Experiments

This section provides details about the numerical experiments and verification that the convergence conditions are satisfied.

B.5.1 Toy Example

In this toy example, we consider randomly generated instances of linear two timescale SA in the form (2.1), (2.2) with i.i.d. samples (and thus the martingale noise setting). In particular, we let the iterates $\theta_k, w_k \in \mathbb{R}^d$ be d-dimensional and construct a problem instance as follows:

1. Sample a random matrix T whose entries are drawn i.i.d. from the uniform distribution $U[-1, 1]$; Compute the QR-decomposition as $T = QR$.
2. Set $A_{12} = Q$ and $A_{22} = Q^\top \Lambda_0 Q$, where Λ_0 is a diagonal matrix with i.i.d. entries from $U[-1, 1]$.
3. Sample a random matrix R whose entries are drawn i.i.d. from the uniform distribution $U[-1, 1]$.
4. Set $A_{11} = RR^\top + I$ and $A_{21} = Q^\top \Lambda_1$, where Λ_1 is a diagonal matrix with i.i.d. entries from $U[-1, 1]$.
5. Sample a stationary solution pair θ^*, w^* with i.i.d. entries from $U[-1, 1]$.

6. Compute b_1, b_2 using the generated matrices and stationary points, i.e.,

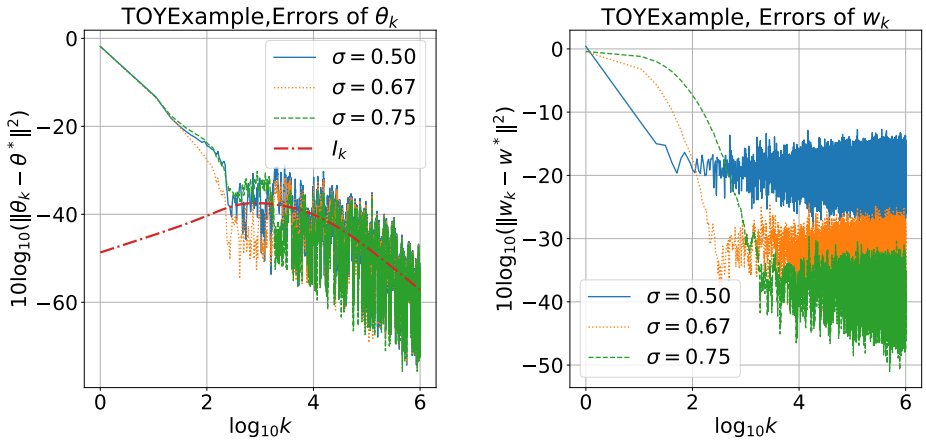
$$b_1 = A_{11}\theta^\star + A_{12}w^\star, \quad b_2 = A_{21}\theta^\star + A_{22}w^\star.$$

During the linear two timescale SA iteration, the noise terms are generated as

$$V_{k+1} = F_V^k + A_{V,\theta}^k\theta_k + A_{V,w}^k w_k, \quad W_{k+1} = F_W^k + A_{W,\theta}^k\theta_k + A_{W,w}^k w_k$$

where $F_V^k, A_{V,\theta}^k, A_{V,w}^k$ are vectors/matrices with entries drawn i.i.d. from the standard normal distribution $\mathcal{N}(0, 0.1)$, and $F_W^k, A_{W,\theta}^k, A_{W,w}^k$ are vectors/matrices with entries drawn i.i.d. from the standard normal distribution $\mathcal{N}(0, 0.5)$. With the above constructions, it can be verified that the required assumptions A9, A11, A12 of the martingale noise setting hold. It remains to verify that the step sizes chosen satisfy A10.

Below, we show the plots of deviations in θ_k and w_k without normalization by the step sizes (see Fig. B.1).



(a) (b)

Figure B.1: Unnormalized deviations from stationary point $(\theta^\star, w^\star)$ and term I_k : the toy example.

B.5.2 Garnet Problems

GTD Algorithm and Policy Evaluation Problem The specific form of linear two timescale SA used in this example follows from that of the classical GTD algorithm [72, 73], which is described below for completeness. Let $\mathcal{S}, \mathcal{A}$ be some discrete state and action spaces (for clarity we bound ourselves by discrete setting, but one could formulate it in more general way), $\rho \in (0, 1)$ and $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ be a *stochastic policy*, i.e. mapping from states to probability measures over actions. When in state s the agent performs action a (distributed according to its policy π), it transitions randomly to state s' with probability $p(s'|s, a)$ and obtains reward $r(s, a)$. This induces a Markov chain with transition probabilities $p_\pi(s'|s) := \sum_{a\in\mathcal{A}} \pi(a|s)p(s'|s, a)$.

The goal of policy evaluation is to estimate the average discounted cumulative reward obtained with the policy π. In detail, we evaluate the value function $V_\pi(s) := \mathbb{E}\left[r(s, a) + \sum_{k=1} \rho^k r(s_k, a_k)\right]$ with ρ being the *discounting factor*. As the state space $|\mathcal{S}|$ is often large, we use the linear approximation $V_\pi(s) \approx V_\theta(s) := \langle\theta, \phi(s)\rangle$, where $\phi : \mathcal{S} \to \mathbb{R}^d$ is a pre-defined feature map. Define also *temporal difference* at iteration

$k \in \mathbb{Z}_+$ for transition $s_k \to s_{k+1}$ as $\delta_k := r(s_k, a_k) + \rho V_{\theta_k}(s_{k+1}) - V_{\theta_k}(s_k)$. For brevity, denote the observation at iteration $k \in \mathbb{Z}_+$, namely, $\phi(s_k)$, $\phi(s_{k+1})$, $r(s_k, a_k)$ as ϕ_k, ϕ_{k+1}, r_k respectively. The GTD algorithm iterations are described as:

$$\theta_{k+1} = \theta_k + \beta_k \left[\phi_k - \rho\phi_{k+1}\right] \langle \phi_k, w_k \rangle, \quad w_{k+1} = w_k + \gamma_k \left[\phi_k \delta_k - w_k\right]. \tag{B.158}$$

The above is a special case of our linear two timescale SA in (2.3), (2.4) with the notations:

$$b_1 = 0,\ A_{11} = 0,\ A_{12} = -\mathbb{E}\left[(\phi_k - \rho\phi_{k+1})\phi_k^\top\right], \tag{B.159}$$

$$b_2 = \mathbb{E}\left[\phi_k r_k\right],\ A_{21} = -\mathbb{E}\left[\phi_k(\rho\phi_{k+1} - \phi_k)^\top\right],\ A_{22} = \mathrm{I}_d, \tag{B.160}$$

$$V_{k+1} = \left((\phi_k - \rho\phi_{k+1})\phi_k^\top - \mathbb{E}\left[(\phi_k - \rho\phi_{k+1})\phi_k^\top\right]\right) w_k, \tag{B.161}$$

$$W_{k+1} = \phi_k r_k - \mathbb{E}\left[\phi_k r_k\right] + \left((\phi_k - \rho\phi_{k+1})\phi_k^\top - \mathbb{E}\left[(\phi_k - \rho\phi_{k+1})\phi_k^\top\right]\right) \theta_k, \tag{B.162}$$

where the expectations above are taken with respect to the stationary distribution of the MDP under policy π. Particularly, the noise terms V_{k+1}, W_{k+1} follow the Markovian noise setting.

Garnet Problem The Garnet problem refers to a set of policy evaluation problems with randomly generated problem instances, originally proposed in [3]. Here, we consider a simpler version of Garnet problems described in [34]. Particularly, we consider a finite-state MDP with the parameters n_S as the number of states, n_A as the number of possible actions in each state, b as the branching factor, i.e., the number of transitions from each state-action pair to a new state, p as the number of features in the linear function approximation applied. For any pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ we choose b states $\mathcal{S}' \subset \mathcal{S}$ out of $|\mathcal{S}|$ at random and then draw the probabilities of going from (s, a) to $s' \in \mathcal{S}'$. For the features, for each state $s \in \mathcal{S}$ the corresponding feature vector $\phi(s)$ is generated from $(U[0, 1])^p$. In our numerical example, we consider a particular problem from the family $n_S = 30, n_A = 8, b = 2, p = 8$.

By the above constructions, we observe that the assumptions A9, B5–B7 are all satisfied. It remains to verify that the step sizes chosen satisfy A10, B8.

B.5.3 Step Size Parameters

We consider the family of step size schedules:

$$\beta_k = c^\beta/(k_0^\beta + k), \quad \gamma_k = c^\gamma/(k_0^\gamma + k)^\sigma, \tag{B.163}$$

with $\sigma \in [0.5, 1]$ and the parameters $c^\beta, c^\gamma, k_0^\beta, k_0^\gamma$. Note that

$$\frac{\beta_k}{\gamma_k} = \frac{c^\beta (k_0^\gamma + k)^\sigma}{c^\gamma (k_0^\beta + k)} \le \frac{c^\beta}{c^\gamma} \left(\frac{k_0^\gamma}{k_0^\beta}\right)^\sigma =: \kappa$$

since we have $\sigma \le 1$. This ensures A10-1. Furthermore, we observe that

$$\frac{\gamma_{k-1}}{\gamma_k} = \left(1 + \frac{1}{k_0^\gamma + k - 1}\right)^\sigma \le 1 + \frac{\sigma}{k_0^\gamma + k - 1} \le 1 + \frac{\sigma k_0^\gamma}{c^\gamma (k_0^\gamma - 1)} \frac{c^\gamma}{(k_0^\gamma + k)^\sigma} = 1 + \frac{\sigma k_0^\gamma}{c^\gamma (k_0^\gamma - 1)} \gamma_k,$$

On the other hand, we also have

$$\frac{\gamma_{k-1}}{\gamma_k} \le 1 + \frac{\sigma k_0^\beta}{c^\beta (k_0^\gamma - 1)} \frac{c^\beta}{k_0^\beta + k} = 1 + \frac{\sigma k_0^\beta}{c^\beta (k_0^\gamma - 1)} \beta_k$$

Similar upper bound can be derived for β_{k-1}/β_k. Setting c^γ, c^β large enough ensures A10-2. Lastly, B8 can be guaranteed by observing that $\sigma \geq 0.5$.

The above discussions illustrate that the satisfaction of A10 hinge on setting a large c^γ, c^β. However, this requirement can be hard to satisfy since we also have requirements such as $\gamma_k \leq \gamma_\infty^{\mathsf{mark}}$, $\beta_k \leq \beta_\infty^{\mathsf{mark}}$. To this end, we have to set a large k_0^β, k_0^γ. As a result, there are four inter-related hyper parameters to be tuned in order to ensure the desired convergence of linear two timescale SA. We remark that tuning the step size parameters for SA scheme is generally difficult.

Appendix C

The code of EV-based variance reduction which was used is available on GitHub [37].

C.1 Proofs

C.1.1 Verification of the Assumptions

Proof of Proposition 25

Proposition 40. *(Proposition 25) If there exist constants $C_L > 0$ and $C_R > 0$ such that*

$$\begin{aligned} \forall \theta \in \Theta,\ a \in \mathcal{A}, s \in \mathcal{S}, b_\phi \in \mathcal{B}_\Phi \quad & \|\nabla_\theta \log \pi(a|s)\| \leq C_L, \\ & |R(s,a)| \leq C_R, \\ & |b_\phi(s,a)| \leq C_R, \end{aligned}$$

then Assumption 1 is satisfied.

▷ Note that the class of estimators $\mathcal{H}$ in the gradient scheme consists of the maps

$$\widetilde{\nabla}_\theta^{b_\phi} J : X \mapsto \sum_{t=0}^{T-1} \gamma^t (G_t - b_\phi(S_t, A_t)) \nabla_\theta \log \pi(A_t|S_t).$$

Therefore,

$$\|\widetilde{\nabla}_\theta^{b_\phi} J(X)\| \leq \sum_{t=0}^{T-1} \gamma^t |G_t(X) - b_\phi(S_t, A_t)| \|\nabla_\theta \log \pi(A_t|S_t)\| \leq \frac{2C_R C_L}{1-\gamma}$$

and in the case $\gamma = 1$

$$\|\widetilde{\nabla}_\theta^{b_\phi} J(X)\| \leq 2C_R C_L T.$$

□

Proof of Proposition 26

Proposition 41. *(Proposition 26) Suppose that Assumption 3 holds for $\mathcal{B}_\phi$, i.e.*

$$\mathcal{N}(\epsilon, \mathcal{B}_\Phi, \|\cdot\|_{L^2(P_K)}) \leq \left(\frac{c}{\epsilon}\right)^\alpha$$

for some $c, \alpha > 0$. If there exist constant $C_L > 0$ such that

$$\forall \theta \in \Theta,\ a \in \mathcal{A}, s \in \mathcal{S} \quad \|\nabla_\theta \log \pi(a|s)\| \leq C_L,$$

then Assumption 3 holds also for $\mathcal{H}$ with the same constant $\alpha' = \alpha$ and constant $c' = cC_L\sqrt{2/(1-\gamma^2)}$.

▷ Let us fix $\epsilon' > 0$ and consider two estimators from $\mathcal{H}$: $\widetilde{\nabla}^{b_\phi} J$ and $\widetilde{\nabla}^{b_{\phi'}} J$ which is a member of the ϵ'-net of $\mathcal{B}_\Phi$, in other words, such that

$$\|b_\phi - b_{\phi'}\|_{L^2(P_K)} := \sqrt{P_K(b_\phi - b_{\phi'})^2} \leq \epsilon'.$$

Recall that

$$\left\|\widetilde{\nabla}^{b_\phi} J - \widetilde{\nabla}^{b_{\phi'}} J\right\|_{L^2(P_K)} = \sqrt{P_K \left\|\widetilde{\nabla}^{b_\phi} J - \widetilde{\nabla}^{b_{\phi'}} J\right\|_2^2}$$

and let us bound $\left\|\widetilde{\nabla}^{b_\phi} J(X_i) - \widetilde{\nabla}^{b_{\phi'}} J(X_i)\right\|_2$ for some arbitrary $i = 1, .., K$. We could derive

$$\left\|\widetilde{\nabla}^{b_\phi} J(X_i) - \widetilde{\nabla}^{b_{\phi'}} J(X_i)\right\|_2 \leq C_L \sum_{t=0}^{T-1} \gamma^t \left|b_\phi(S_t^{(i)}, A_t^{(i)}) - b_{\phi'}(S_t^{(i)}, A_t^{(i)})\right|$$

and, thus, it leads to

$$P_K \left\|\widetilde{\nabla}^{b_\phi} J - \widetilde{\nabla}^{b_{\phi'}} J\right\|_2^2 \leq 2C_L^2 \frac{1}{K} \sum_{i=1}^{K} \sum_{t=0}^{T-1} \gamma^{2t} \left|b_\phi(S_t^{(i)}, A_t^{(i)}) - b_{\phi'}(S_t^{(i)}, A_t^{(i)})\right|^2 \leq$$

$$\leq 2C_L^2 \sum_{t=0}^{T-1} \gamma^{2t} \frac{1}{K} \sum_{i=1}^{K} \left|b_\phi(S_t^{(i)}, A_t^{(i)}) - b_{\phi'}(S_t^{(i)}, A_t^{(i)})\right|^2 \leq \frac{2C_L^2 \epsilon'^2}{1-\gamma^2}.$$

This allows us to use the ϵ'-net for $\mathcal{B}_\Phi$ to construct ϵ-net for $\mathcal{H}$. Hence, Assumption 16 is satisfied with $\alpha' = \alpha$ and $c' = cC_L\sqrt{2/(1-\gamma^2)}$. □

Let us briefly remark that the Proposition allows transferring any covering assumption for baselines to the vector setting and so one could use the assumptions for baselines which are much easier to check in practice.

C.1.2 Proof of Proposition 27: A2C as an Upper Bound for EV

Proposition 42. *(Proposition 27) If there exist constant $C_L > 0$ such that*

$$\forall \theta \in \Theta,\ a \in \mathcal{A}, s \in \mathcal{S} \quad \|\nabla_\theta \log \pi(a|s)\| \leq C_L,$$

then for all $K \geq 2$ A2C goal function $V_K^{A2C}(\phi)$ is an upper bound (up to a constant) for EV goal functions:

$$V_K^{EVm}(\phi) \leq 2C_L^2 V_K^{A2C}(\phi), \quad V_K^{EVv}(\phi) \leq 2C_L^2 V_K^{A2C}(\phi).$$

▷ First, note that for all ϕ, by Jensen's inequality, $V_K^{EVv}(\phi) \leq V_K^{EVm}(\phi)$, so we could work with the bound for EVm. Secondly, $K = 1$ simply does not allow using EVv-criterion, but the bound for EVm remains valid. Via Young's inequality we get

$$V_K^{EVm}(\phi) \leq \frac{2}{K} \sum_{k=1}^{K} \sum_{t=0}^{T-1} \gamma^{2t} \left(G_t(X^{(k)}) - b_\phi(S_t^{(k)}, A_t^{(k)})\right)^2 \left\|\nabla_\theta \log \pi\left(A_t^{(k)}|S_t^{(k)}\right)\right\|_2^2 \leq 2C_L^2 V_K^{A2C}(\phi).$$

□

C.1.3 Proof of the Main Theorem

Suppose we are given sample $X, X_1, ..., X_K$ of random vectors taking values in $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{H} := \left\{h : \mathcal{X} \to \mathbb{R}^D \;\; s.t. \;\; \mathbb{E}[h(X)] = \mathcal{E}\right\}$. Also denote $\|\cdot\| := \|\cdot\|_2$ for shorter notation, when applied to function $h : \mathcal{X} \to \mathbb{R}^D$, $\|h\| := \sup_{x\in\mathcal{X}} \|h(x)\|$ by default. The brackets $(\cdot, \cdot)$ denote the standard inner product.

Our goal is to find

$$h_* \in \text{argmin}_{h\in\mathcal{H}} V(h)$$

with variance functional defined as

$$V(h) := \mathbb{E}\left[\|h(X) - \mathcal{E}\|_2^2\right].$$

In order to tackle this problem we consider the simpler one called *Empirical Variance(EV)* and calculate

$$h_K \in \text{argmin}_{h\in\mathcal{H}} V_K(h)$$

with

$$V_K(h) := \frac{1}{K-1} \sum_{k=1}^{K} \|h(X_k) - P_K h\|^2,$$

where P_K is the empirical measure based on $X_1, .., X_K$, so $P_K h = \frac{1}{K}\sum_{k=1}^{K} h(X_k)$. In what follows we will operate with several key assumptions about the problem at hand.

A 17. (Assumption 14) Class $\mathcal{H}$ consists of bounded functions:

$$\forall h \in \mathcal{H} \quad \sup_{x\in\mathcal{X}} \|h(x)\| \leq b.$$

A 18. (Assumption 15) The solution h_* is unique and $\mathcal{H}$ is star-shaped around h_*:

$$\forall h \in \mathcal{H},\ \alpha \in [0, 1] \quad \alpha h + (1-\alpha)h_* \in \mathcal{H}.$$

Star-shape assumption replaces the assumption of the convexity of $\mathcal{H}$ which is stronger and yet this replacement does not change much in the analysis.

A 19. (Assumption 16) Class $\mathcal{H}$ has covering of polynomial size: there are $\alpha \geq 2$ and $c > 0$ such that for all $u \in (0, b]$

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(P_K)}, u) \leq \left(\frac{c}{u}\right)^{\alpha} \;\; a.s.$$

where the norm is defined as

$$\|h\|_{L^2(P_K)} = \|h\|_{(K)} := \sqrt{P_K \|h\|_2^2}$$

The basis of the analysis lies in usage of

Lemma 43. *(Lemma 4.1 in [8]) Let $\{\phi(\delta) : \delta \geq 0\}$ be non-negative r.v. indexed by $\delta \geq 0$ such that a.s. $\phi(\delta) \leq \phi(\delta')$ if $\delta \leq \delta'$. Define $\{\beta(\delta, t) : \delta \geq 0, t \geq 0\}$, deterministic real numbers such that*

$$\mathbb{P}(\phi(\delta) \geq \beta(\delta, t)) \leq e^{-t}.$$

Set for all non-negative t

$$\beta(t) := \inf\left\{\tau > 0 \ : \ \sup_{\delta \geq \tau} \frac{\beta(\delta, t\delta/\tau)}{\delta} \leq \frac{1}{2}\right\}.$$

If $\hat{\delta}$ is a non-negative random variable which is a priori bounded and such that almost surely $\hat{\delta} \leq \phi(\hat{\delta})$, then for all $t \geq 0$

$$\mathbb{P}\left(\hat{\delta} \geq \beta(t)\right) \leq 2e^{-t}.$$

We would like to stress out that the main idea of the proof remains the same but on the way there must be done some changes to fit it into the setting of vector estimation we consider.

Bound for Functions with δ-Optimal Variance

The idea is to construct an upper bound with high probability for excess risk $V(h_K) - V(h_*)$ under assumptions $V(h) - V(h_*) \leq \delta$ and use that as ϕ in Lemma 43. This will give us the desired w.h.p. bound for excess risk in general. Let us start with the basic bound from which we obtain all further results. Essentially, the sequence $\phi(\delta)$ from the Lemma appears in the left part.

Theorem 44. *Assume A14, A15. If $h \in \mathcal{H}(\delta) := \{h \in \mathcal{H} \mid V(h) - V(h_*) < \delta\}$, then with probability at least $1 - e^{-t}$*

$$V(h_K) - V(h_*) \leq 2\mathbb{E}\phi_K^{(1)}(\delta) + 4\left(\mathbb{E}\sup_{h \in \mathcal{H}(\delta)} \|(P - P_K)h\|\right)^2 + \frac{40b^2t + 24b^2}{3K} + 12b\sqrt{\frac{\delta t}{K}}$$

with

$$\phi_K^{(1)}(\delta) := \sup_{h \in \mathcal{H}(\delta)} (P - P_K)l(h).$$

▷ To begin with, add and subtract $V_K(h_K), V_K(h_*)$ to get

$$V(h_K) - V(h_*) \pm V_K(h_K) \pm V_K(h_*) \leq V(h_K) - V(h_*) - (V_K(h_K) - V_K(h_*)).$$

the last terms can be represented as

$$V_K(h) = P_K\|h - \mathcal{E}\|_2^2 - \frac{1}{K(K-1)}\sum_{i \neq j = 1}^{K} (h(X_i) - \mathcal{E}, h(X_j) - \mathcal{E})$$

giving us

$$\begin{aligned}
&V(h_K) - V(h_*) - (V_K(h_K) - V_K(h_*)) = \\
&= (P - P_K)\left(\|h_K - \mathcal{E}\|_2^2 - \|h_* - \mathcal{E}\|_2^2\right) + \\
&+ \frac{1}{K(K-1)}\sum_{i \neq j = 1}^{K} (h(X_i) - \mathcal{E}, h(X_j) - \mathcal{E}) - (h_*(X_i) - \mathcal{E}, h_*(X_j) - \mathcal{E}) = \\
&= T_K(h_K) + W_K(h_K)
\end{aligned}$$

which introduces

$$T_K(h_K) := (P - P_K)\left(\|h_K - \mathcal{E}\|_2^2 - \|h_* - \mathcal{E}\|_2^2\right),$$
$$W_K(h_K) := w(h_K) - w(h_*),\ w(h) = \frac{1}{K(K-1)}\sum_{i\neq j=1}^{K}\left(h(X_i) - \mathcal{E}, h(X_j) - \mathcal{E}\right).$$

Since $h \in \mathcal{H}(\delta)$ it is true that

$$V(h_K) - V(h_*) \leq \sup_{h\in\mathcal{H}(\delta)} T_K(h) + W_K(h) \leq \sup_{h\in\mathcal{H}(\delta)} T_K(h) + \sup_{h\in\mathcal{H}(\delta)} W_K(h) = \phi_K^{(1)}(\delta) + \phi_K^{(2)}(\delta).$$

Bound for $\phi_K^{(1)}$. Firstly, let us introduce

$$l(h) = \|h - \mathcal{E}\|_2^2 - \|h_* - \mathcal{E}\|_2^2.$$

We can exploit the same Talagrand's inequality as in [8, p.12]. Recall that functions $h \in \mathcal{H}$ are bounded, therefore $|l(h)| \leq 4b^2$ and, hence, with probability at least $1 - e^{-t}$

$$\phi_K^{(1)}(\delta) \leq \mathbb{E}\phi_K^{(1)}(\delta) + \sqrt{\frac{2t}{K}\left(\sigma^2(\delta) + 8b^2\mathbb{E}\phi_K^{(1)}(\delta)\right)} + \frac{4b^2t}{3K},$$

where

$$\sigma^2(\delta) := \sup_{h\in\mathcal{H}(\delta)} Pl(h)^2.$$

Let us bound this quantity. In order to proceed, notice that for all $h_1, h_2 \in \mathcal{H}$

$$l(h_1) - l(h_2) = (h_1, h_1) - (h_2, h_2) + 2(h_2 - h_1, \mathcal{E}) = (h_2 - h_1, h_2 - h_1) + 2(h_1 - h_2, h_2 - \mathcal{E})$$

and so for all $x \in \mathcal{X}$

$$|l(h_1)(x) - l(h_2)(x)| \leq 6b\|h_2(x) - h_1(x)\| \tag{C.1}$$

is obtained with Cauchy-Schwarz inequality. This results particularly in

$$Pl(h)^2 \leq 36b^2P\|h - h_*\|_2^2.$$

Since $l(h)$ has very specific form involving square norms, we could state that

$$P\|h - h_*\|_2^2 = 2Pl(h) - 4Pl\left(\frac{h + h_*}{2}\right) \leq 2Pl(h)$$

implying

$$Pl(h)^2 \leq 72b^2Pl(h) \leq 72b^2\delta$$

by definition of $\mathcal{H}(\delta)$.

With this and $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v},\ 2\sqrt{uv} \leq u + v$ the bound can be simplified to

$$\phi_K^{(1)}(\delta) \leq 2\mathbb{E}\phi_K^{(1)}(\delta) + 12b\sqrt{\frac{\delta t}{K}} + \frac{16b^2t}{3K}.$$

Bound for $\phi_K^{(2)}$. This is much simpler, observe that

$$w_K(h) = \frac{1}{K(K-1)} \left\{ \sum_{i,j=1}^{K} (h(X_i) - \mathcal{E}, h(X_j) - \mathcal{E}) - \sum_{i=1}^{K} \|h(X_i) - \mathcal{E}\|^2 \right\} =$$

or,

$$= \frac{K}{K-1} (P_K(h-\mathcal{E}), P_K(h-\mathcal{E})) - \frac{1}{K-1} P_K \|h-\mathcal{E}\|_2^2.$$

So,

$$W_K(h_K) = w_K(h_K) - w_K(h_*) \leq \frac{K}{K-1} (P_K(h-\mathcal{E}), P_K(h-\mathcal{E})) + \frac{1}{K-1} P_K \|h_*-\mathcal{E}\|_2^2 \leq$$

where the first inequality is due to negative terms, applying bound for h now results in

$$\leq \frac{K}{K-1} (P_K(h-\mathcal{E}), P_K(h-\mathcal{E})) + \frac{4b^2}{K-1}.$$

Finally,

$$\phi_K^{(2)} \leq 2 \left(\sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \right)^2 + \frac{4b^2}{K-1}$$

and after adding and subtracting the expectation of the supremum and exploiting $2ab \leq a^2 + b^2$ together with $1/(K-1) \leq 2/K$ for $K \geq 2$ we arrive to

$$\leq 4 \left(\sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \; - \; \mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \right)^2 + 4 \left(\mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \right)^2 + \frac{8b^2}{K}.$$

Apply now probabilistic inequality for bounded differences [16, Th. 6.2] to estimate the first term, with probability $\geq 1 - e^{-t}$

$$\left(\sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \; - \; \mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \right)^2 \leq \frac{2b^2 t}{K}.$$

Therefore, with such probability

$$\phi_K^{(2)} \leq \frac{8b^2 t}{K} + 4 \left(\mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \right)^2 + \frac{8b^2}{K}.$$

The resulting bound is now

$$V(h_K) - V(h_*) \leq 2\mathbb{E}\phi_K^{(1)}(\delta) + 4 \left(\mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P-P_K)h\| \right)^2 + \frac{40b^2 t + 24b^2}{3K} + 12b\sqrt{\frac{\delta t}{K}}.$$

with probability $\geq 1 - e^{-t}$. $\square$

Bounding the Suprema

To proceed further we need now to bound the two suprema in Theorem 44. Lemma 5.3 in [8] gives us a tool, it is stated as follows.

Lemma 45. *Assume $X_1, ..., X_K$ to be i.i.d. sample and P_K be empirical measure. Let*

$$\mathcal{H} := \{h : \mathcal{X} \to [-b, b]\}$$

and suppose that for all $u \in (0, b]$

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(P_K)}, u) \leq \left(\frac{c}{u}\right)^{\alpha} \ a.s.,$$

then $\forall \sigma \in [\sigma_{\mathcal{H}}, b]$

$$\mathbb{E} \sup_{h \in \mathcal{H}} (P - P_K)h \leq A \left(\sqrt{\frac{\alpha \sigma^2}{K} \log \frac{c}{\sigma}} + \frac{\alpha b}{K} \log \frac{c}{\sigma} \right)$$

where constants are explicitly given.

Lemma 46. *Let A14, A16 hold. Then*

$$\mathbb{E}\phi_K^{(1)} \leq 2592 \left(\sqrt{\frac{72 b^2 \delta \alpha}{K} \log \frac{c}{6b\sqrt{2\delta}}} + \frac{\alpha b}{K} \log \frac{c}{6b\sqrt{2\delta}} \right).$$

▷ Define

$$L(\delta) := \{l(h) \mid h \in \mathcal{H}(\delta)\}$$

and note that in our case it also holds that

$$\mathcal{N}(L(\delta), \|\cdot\|_{L^2(P_K)}, u) \leq \mathcal{N}(\mathcal{H}(\delta), \|\cdot\|_{L^2(P_K)}, u),$$

therefore, we could apply Lemma 45 to $L(\delta)$ and get the result. □

The second supremum, fortunately, can be handled simpler.

Lemma 47. *If A14 is satisfied, it holds that*

$$\mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P - P_K)h\| \leq \frac{2b}{\sqrt{K}}.$$

▷ First note that by symmetrization

$$\mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \|(P - P_K)h\| \leq \frac{2}{K} \mathbb{E} \sup_{h \in \mathcal{H}(\delta)} \mathbb{E}_\xi \| \sum_{k=1}^{K} \xi_k h(X_k) \|$$

where ξ_k are i.i.d. Rademacher's random variables. Expand the norm, apply Jensen's inequality to the square root and get

$$\mathbb{E}_\xi \| \sum_{k=1}^{K} \xi_k h(X_k) \| = \mathbb{E}_\xi \sqrt{\sum_{k=1}^{K} \|h(X_k)\|^2 + 2 \sum_{d=1}^{D} \sum_{1 \leq i < j \leq K} \xi_i h(X_i) \xi_j h(X_j)} \leq \tag{C.2}$$

$$\leq \sqrt{\sum_{k=1}^{K} \|h(X_k)\|^2 + 2 \mathbb{E}_\xi \sum_{d=1}^{D} \sum_{1 \leq i < j \leq K} \xi_i h(X_i) \xi_j h(X_j)} = \sqrt{\sum_{k=1}^{K} \|h(X_k)\|^2} \leq b\sqrt{K}. \tag{C.3}$$

□

With Theorem 44, Lemma 46 and Lemma 47 we make a conclusion.

Theorem 48. *Let A14, A15, A16 hold. If $h \in \mathcal{H}(\delta)$ then*

$$V(h_K) - V(h_*) \leq 5184 \left(\sqrt{\frac{72b^2\delta\alpha}{K} \log \frac{c}{6b\sqrt{2\delta}}} + \frac{\alpha b}{K} \log \frac{c}{6b\sqrt{2\delta}} \right) + \\ + \frac{40b^2t + 72b^2}{3K} + 12b\sqrt{\frac{\delta t}{K}}$$

with probability $\geq 1 - e^{-t}$.

Proof of the Main Theorem

Finally, we apply Lemma 43. What remains is to carefully compute $\beta(t)$ and obtain

Theorem 49. *(Theorem 2 in the main text) It holds that*

$$V(h) - V(h_*) \leq \max_j \beta^{(j)}(t)$$

with probability at least $1 - 4e^{-t}$, $\beta^{(j)}(t)$ are defined in the proof.

▷ We have bounded with probability $\geq 1 - e^{-t}$ the excess risk of $\mathcal{H}(\delta)$, so that

$$\beta_K(\delta, t) = C_0 \left(\sqrt{\frac{b^2\delta\alpha}{K} \log \frac{c}{6b\sqrt{2\delta}}} + \frac{\alpha b}{K} \log \frac{c}{6b\sqrt{2\delta}} \right) + \frac{40b^2t + 72b^2}{3K} + 12b\sqrt{\frac{\delta t}{K}}.$$

Now compute for $\tau > 0$

$$\sup_{\delta \geq \tau} \frac{\beta_K(\delta, t\delta/\tau)}{\delta} = C_0 \left(\sqrt{\frac{b^2\alpha}{\tau K} \log \frac{c}{6b\sqrt{2\tau}}} + \frac{\alpha b}{\tau K} \log \frac{c}{6b\sqrt{2\tau}} \right) + \\ + \frac{40b^2t + 72b^2}{3K\tau} + 12b\sqrt{\frac{\delta t}{K\tau}}.$$

Finally, observe that

$$\beta_K(t) := \inf \left\{ \tau > 0 \ : \ \sup_{\delta \geq \tau} \frac{\beta_K(\delta, t\delta/\tau)}{\delta} \leq \frac{1}{2} \right\} \leq \max_j \beta^j(t)$$

where

$$\beta^1(t) = \inf \left\{ \tau > 0 : 72\sqrt{\frac{32b^2\alpha}{K\tau} \log \frac{c}{4b\sqrt{2\tau}}} \leq \frac{1}{8} \right\} \leq C_1 \frac{\log K}{K},$$

$$\beta^2(t) = \inf \left\{ \tau > 0 : 2592 \frac{\alpha b}{K\tau} \log \frac{c}{4b\sqrt{2\tau}} \leq \frac{1}{8} \right\} \leq C_2 \frac{\log K}{K},$$

$$\beta^3(t) = \inf \left\{ \tau > 0 : \frac{40b^2t + 72b^2}{3K\tau} \leq \frac{1}{8} \right\} = \frac{8(40b^2t + 72b^2)}{3K},$$

$$\beta^4(t) = \inf \left\{ \tau > 0 : 12b\sqrt{\frac{t}{K\tau}} \leq \frac{1}{8} \right\} = \frac{9216b^2t}{K}.$$

It holds with probability $1 - 4e^{-t}$ □

C.1.4 Proposition: Unbiasedness of S-Baseline

S-baseline is known to result in unbiased estimate, here for the sake of completeness we give a proof.

Proposition 50. *For all $b_\phi : \mathcal{S} \to \mathbb{R}$ the expected value*

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t b_\phi(S_t) \nabla_\theta \log \pi_\theta(A_t|S_t)\right] = 0.$$

▷ Let us consider one term of the sum and note that we can use tower property of conditional expectation:

$$\mathbb{E}\left[\gamma^t b_\phi(S_t) \nabla_\theta \log \pi_\theta(A_t|S_t)\right] = \gamma^t \mathbb{E}\left\{\mathbb{E}\left[b_\phi(S_t) \nabla_\theta \log \pi_\theta(A_t|S_t) \mid S_t\right]\right\}.$$

Now note that $b_\phi(S_t)$ is measurable in the inner expectation, so,

$$= \gamma^t \mathbb{E}\left\{b_\phi(S_t) \mathbb{E}\left[\nabla_\theta \log \pi_\theta(A_t|S_t) \mid S_t\right]\right\}.$$

Finally, with the help of the log derivative we show that

$$\mathbb{E}\left[\nabla_\theta \log \pi_\theta(A_t|S_t) \mid S_t\right] = 0$$

and the result follows. □

C.1.5 Why Variance Reduction Matters (for Local Convergence)

We base our proof on some techniques of [86] where SVRPG algorithm is considered but the proof we need has the same structure with $b = m = 1$ and some adjustments.

Let $\widetilde{\nabla} J : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T \to \mathbb{R}^D$ be an unbiased gradient estimate (with baseline or just REINFORCE). Our gradient algorithm reads as

$$\theta_{n+1} = \theta_n + \alpha_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla} J(X_n^{(k)}),$$

where $\theta_n \in \Theta \subset \mathbb{R}^D$ are policy parameters at iteration n and $X_n^{(k)} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R})$ is the trajectory data at iteration n of which there are K independent samples. Let us for shorter notation set $\widetilde{\nabla} J_n^K := \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla} J(X_n^{(k)})$. The Lemma below is the in the core of non-convex smooth optimization.

Lemma 51. *If $\forall \theta \in \Theta \quad \|\nabla^2 J(\theta)\|_2 \leq L$, then for all $n \in \mathbb{Z}_{>0}$*

$$J(\theta_{n+1}) \geq J(\theta_n) - \frac{3\alpha_n}{4} \left\|\nabla J(\theta_n) - \widetilde{\nabla} J_n^K\right\|_2^2 + \left(\frac{1}{4\alpha_n} - \frac{L}{2}\right) \|\theta_{n+1} - \theta_n\|_2^2 + \frac{\alpha_n}{8} \|\nabla J(\theta_n)\|_2^2, \tag{C.4}$$

where $v_n = \alpha_n$

▷ It can be obtained by applying lower quadratic bound:

$$J(\theta_{n+1}) \geq J(\theta_n) + \langle \nabla J(\theta_n), \theta_{n+1} - \theta_n \rangle - \frac{L}{2} \|\theta_{n+1} - \theta_n\|_2^2 . \tag{C.5}$$

Next, notice that $\alpha_n \widetilde{\nabla} J_n^K = \theta_{n+1} - \theta_n$ and add and subtract $\widetilde{\nabla} J_n^K$ in the left entry of the second term:

$$J(\theta_{n+1}) \geq J(\theta_n) + \langle \nabla J(\theta_n) - \widetilde{\nabla} J_n^K, \alpha_n \widetilde{\nabla} J_n^K \rangle + \alpha_n \left\| \widetilde{\nabla} J_n^K \right\|_2^2 - \frac{L}{2} \|\theta_{n+1} - \theta_n\|_2^2 . \tag{C.6}$$

Now apply Young's polarization inequality ($ab \geq -(a^2 + b^2)/2$) to the same term and arrive to

$$J(\theta_{n+1}) \geq J(\theta_n) - \frac{\alpha_n}{2} \left\| \nabla J(\theta_n) - \widetilde{\nabla} J_n^K \right\|_2^2 + \frac{\alpha_n}{2} \left\| \widetilde{\nabla} J_n^K \right\|_2^2 - \frac{L}{2} \|\theta_{n+1} - \theta_n\|_2^2 . \tag{C.7}$$

Observe that the second and the third term can be bounded further using

$$\|\nabla J(\theta_{n+1})\|_2^2 \leq 2 \left\| \widetilde{\nabla} J_n^K \right\|_2^2 + 2 \left\| \nabla J(\theta_{n+1}) - \widetilde{\nabla} J_n^K \right\|_2^2 , \tag{C.8}$$

which results in

$$J(\theta_{n+1}) \geq J(\theta_n) - \frac{3\alpha_n}{4} \left\| \nabla J(\theta_n) - \widetilde{\nabla} J_n^K \right\|_2^2 + \left(\frac{1}{4\alpha_n} - \frac{L}{2} \right) \|\theta_{n+1} - \theta_n\|_2^2 + \frac{\alpha_n}{8} \|\nabla J(\theta_n)\|_2^2 . \tag{C.9}$$

□

With this Lemma we can prove a variety of different convergence results, we would rather refer here to [86, 90]. And yet, to illustrate the need for the variance reduction, consider the following theorem.

Theorem 52. *There is a constant $C_R > 0$ such that for all $k > 0$ and $N \leq k$ the following bound holds assuming non-increasing step sizes $\alpha_n \leq 2/L$:*

$$\frac{1}{k} \sum_{n=k-N}^{k} \mathbb{E} \|\nabla J(\theta_n)\|_2^2 \leq \frac{16 C_R}{k \alpha_k} + \frac{1}{k} \sum_{n=k-N}^{k} \mathbb{E} \left\| \nabla J(\theta_n) - \widetilde{\nabla} J_n^K \right\|_2^2 . \tag{C.10}$$

In particular, when $N = k - 1$, one gets

$$\frac{1}{k} \sum_{n=1}^{k} \mathbb{E} \|\nabla J(\theta_n)\|_2^2 \leq \frac{16 C_R}{k \alpha_k} + \frac{1}{k} \sum_{n=1}^{k} \mathbb{E} \left\| \nabla J(\theta_n) - \widetilde{\nabla} J_n^K \right\|_2^2 . \tag{C.11}$$

▷ Introduce quantity $U(\theta) := J(\theta^*) - J(\theta)$. Let us use Lemma 51, divide both parts by α_n and sum them from $n = k - N$ to k with k, N satisfying $N \leq k$, then take the expectation:

$$\sum_{n=k-N}^{k} \mathbb{E} \left[\|\nabla J(\theta_n)\|_2^2 \right] \leq 8 \sum_{n=k-N}^{k} \frac{1}{\alpha_n} \mathbb{E} \left[U(\theta_n) - U(\theta_{n+1}) \right] + 6 \sum_{n=k-N}^{k} \mathbb{E} \left[\left\| \nabla J(\theta_n) - \widetilde{\nabla} J_n^K \right\|_2^2 \right] . \tag{C.12}$$

Notice that we used $\alpha_n < 2/L$ to drop the term with $\|\theta_{n+1} - \theta_n\|_2^2$. One could rewrite the first sum on the right to get

$$\sum_{n=k-N}^{k} \mathbb{E}\left[\|\nabla J(\theta_n)\|_2^2\right] \leq 8 \sum_{n=k-N}^{k} \left(\frac{1}{\alpha_n} - \frac{1}{\alpha_{n-1}}\right) \mathbb{E}\left[U(\theta_n)\right] - \frac{8}{\alpha_k}\mathbb{E}\left[U(\theta_{k+1})\right] + \frac{8}{\alpha_{k-N-1}}\mathbb{E}\left[U(\theta_{k-N})\right] + \tag{C.13}$$

$$+ 6 \sum_{n=k-N}^{k} \mathbb{E}\left[\left\|\nabla J(\theta_n) - \widetilde{\nabla} J_n^K\right\|_2^2\right]. \tag{C.14}$$

Since the rewards are bounded, there is C_R such that for all θ the difference $U(\theta) \leq C_R$; secondly, the step sizes are non-increasing; finally, we can discard the second term which is non-positive. Thus,

$$\sum_{n=k-N}^{k} \mathbb{E}\left[\|\nabla J(\theta_n)\|_2^2\right] \leq 8C_R \sum_{n=k-N}^{k} \left(\frac{1}{\alpha_n} - \frac{1}{\alpha_{n-1}}\right) + \frac{8C_R}{\alpha_{k-N-1}} + 6 \sum_{n=k-N}^{k} \mathbb{E}\left[\left\|\nabla J(\theta_n) - \widetilde{\nabla} J_n^K\right\|_2^2\right] \leq \tag{C.15}$$

$$\leq \frac{8C_R}{\alpha_n} + \frac{8C_R}{\alpha_{k-N-1}} + 6 \sum_{n=k-N}^{k} \mathbb{E}\left[\left\|\nabla J(\theta_n) - \widetilde{\nabla} J_n^K\right\|_2^2\right]. \tag{C.16}$$

We could again use the fact that α_n are non-increasing to simplify the first two terms, then divide both parts by k:

$$\frac{1}{k} \sum_{n=k-N}^{k} \mathbb{E}\left[\|\nabla J(\theta_n)\|_2^2\right] \leq \frac{16C_R}{\alpha_n} + \frac{6}{k} \sum_{n=k-N}^{k} \mathbb{E}\left[\left\|\nabla J(\theta_n) - \widetilde{\nabla} J_n^K\right\|_2^2\right]. \tag{C.17}$$

□

This result shows, that the convergence of the gradient to zero is influenced by the variance of the gradient estimator. In practice, however, the variance reduction ratio is very low and therefore it slightly but not dramatically improves the algorithm. Theory of SVRPG [86], however, suggests that in terms of rates with the accurate design of the step sizes the rate can be slightly improved. Despite all this, variance reduction provably improves *local* convergence but as to global convergence (which is more tricky to specify), the variance also may play a good role in avoiding local optima as shown by [90]. This, we believe partially explains, why in practice the quality of the algorithms is not so strongly influenced by the variance reduction, as one might have thought.

C.2 Additional Experiments and Implementation Details

Here we present additional experimental results. The detailed config-files can be found on GitHub page [37].

C.2.1 Minigrid

Minimalistic Gridworld Environment (MiniGrid) provides gridworld Gym environments that were designed to be simple and lightweight, therefore, ideally fitting for making experiments. In particular, we considered GoToDoor and Unlock.

In both environments we have used 20 independent runs of the algorithms. All measurements of mean rewards and variance of the gradient estimator (measured each 250 epochs on a newly generated pool of 100 trajectories) are averaged over these runs. Standard deviations of the rewards are obtained as the sample standard deviation of the observed rewards reflecting the width of the confidence intervals of the mean reward curves. The exact config-files used for experiments can be found on attached GitHub.

Go-To-Door-5x5

This environment is a room with four doors, one on each wall. The agent receives a textual (mission) string as input, telling it which door to go to, (e.g: "go to the red door"). It receives a positive reward for performing the done action next to the correct door, as indicated in the mission string.

In GoToDoor environment we can clearly see that the EV-agent is at least as good as A2C at lower number of samples ($K = 5$), and the more samples are available during training, the better performance we can observe from EV-agents. We can see mean rewards in absolute values in Fig. C.1 and in relative scale (normalized by the results of REINFORCE) in Fig. C.2 to see improvements over the results of REINFORCE algorithm.

We also address the effect of gradient variance reduction and its effect on the performance of the algorithm. We can see that variance reduction depends on the number of samples. It's negligible, most of the time even an increase is presented, when number of samples is small ($K = 5$ and $K = 10$). We see reduction happening with larger $K = 15$ and $K = 20$. It seems that variance reduction might speed-up the training process, but it is clearly not a key contributor. Variance reduction also seems to be useless at the start since EV-agents' and A2C's seem to have even higher variance than REINFORCE and better performance. However, later reduction might allow to increase final rewards. With $K = 15, 20$ the algorithms are able to reduce REINFORCE gradient variance only by 30%. We give charts demonstrating gradient variance in absolute values in Fig. C.3 and in relative scale (normalized by the gradient variance of REINFORCE) in Fig. C.4.

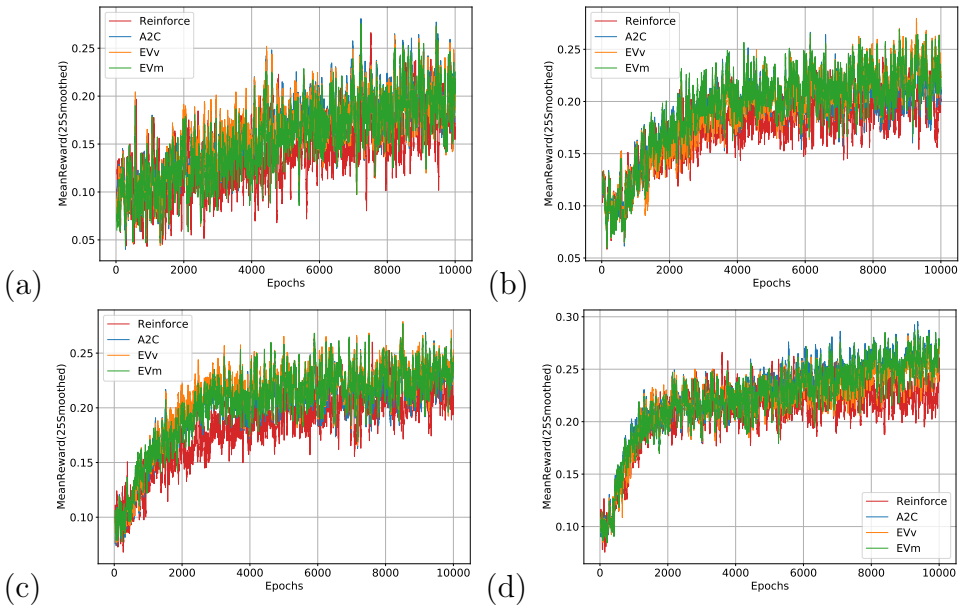


Figure C.1: The charts representing mean rewards in GoToDoor environment, standing for absolute values for cases $K = 5$(a), $K = 10$(b), $K = 15$(c), $K = 20$(d). The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25.

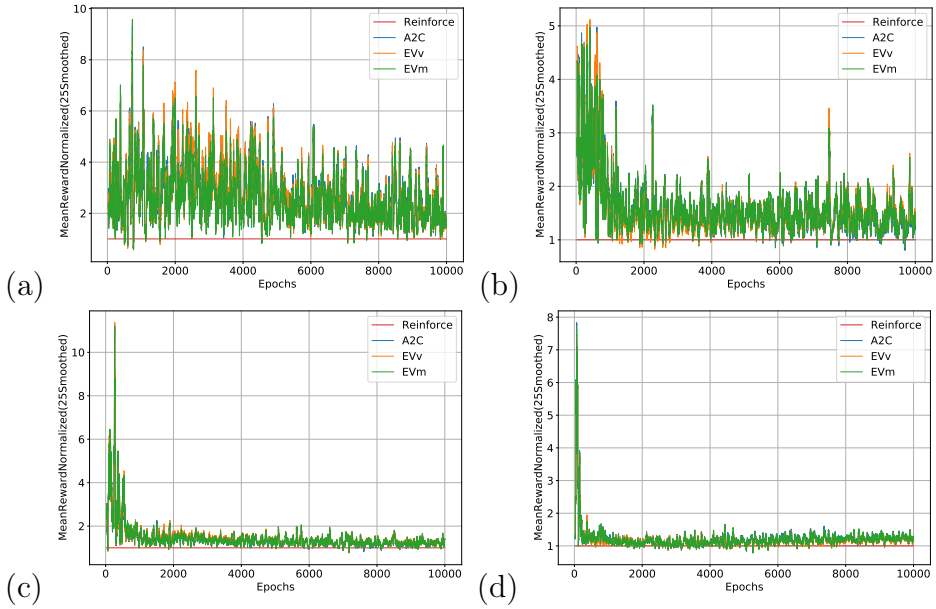


Figure C.2: The charts representing mean rewards in GoToDoor environment, the curves normalized by the mean reward of the REINFORCE. for cases K=5(a), K=10(b), K=15(c), K=20(d). The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25

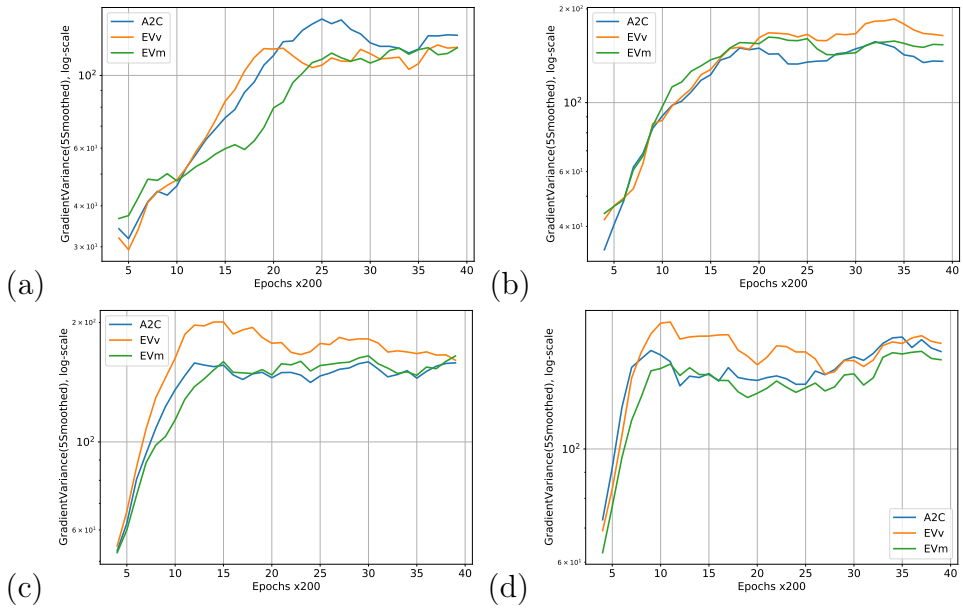


Figure C.3: The charts representing the variance of the gradient estimator in absolute values for cases $K = 5$(a), $K = 10$(b), $K = 15$(c), $K = 20$(d). The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 5.

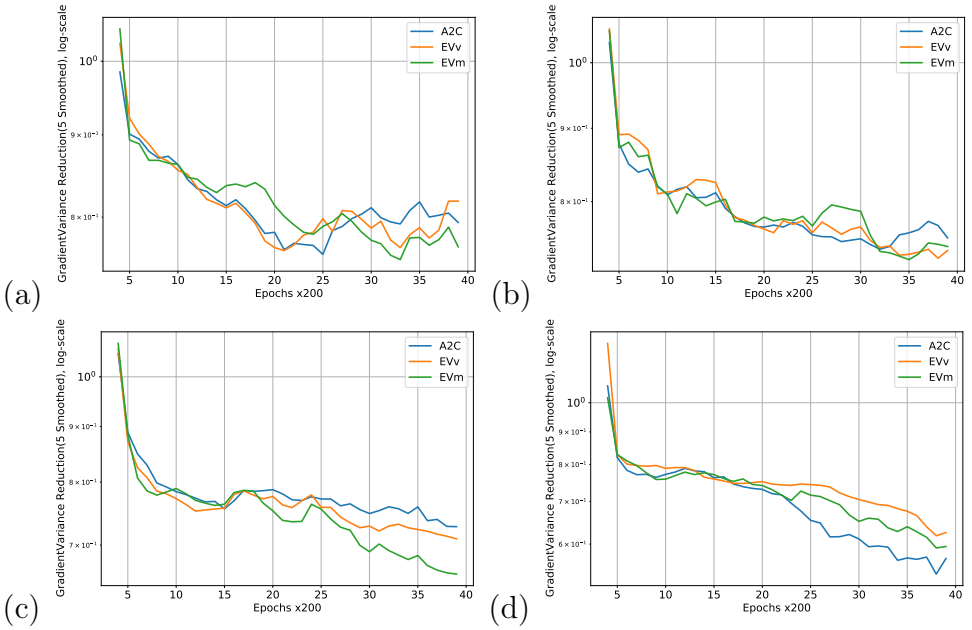


Figure C.4: The charts representing the variance of the gradient estimator normalized by the variance of REINFORCE for cases $K = 5$(a), $K = 10$(b), $K = 15$(c), $K = 20$(d). The numbers < 1 indicate the relative reduction. The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 5.

One can also evaluate the algorithms by looking at the standard deviation of the rewards. Between the methods no significant difference is observed when the sample size is small ($K = 5$ or $K = 10$). It becomes considerable though in cases of $K = 15$ and $K = 20$. EV-agents turn out to have the biggest reward standard deviation among the algorithms. The standard deviation of the rewards is demonstrated in absolute values in Fig. C.5 and in relative scale (normalized by the standard deviation of REINFORCE) in Fig. C.6. We note that this standard deviation does not at all reflect the variance reduction of the gradient estimator as follows from the comparison of the charts. In fact, REINFORCE with no variance reduced is slightly better in this regard than other methods.

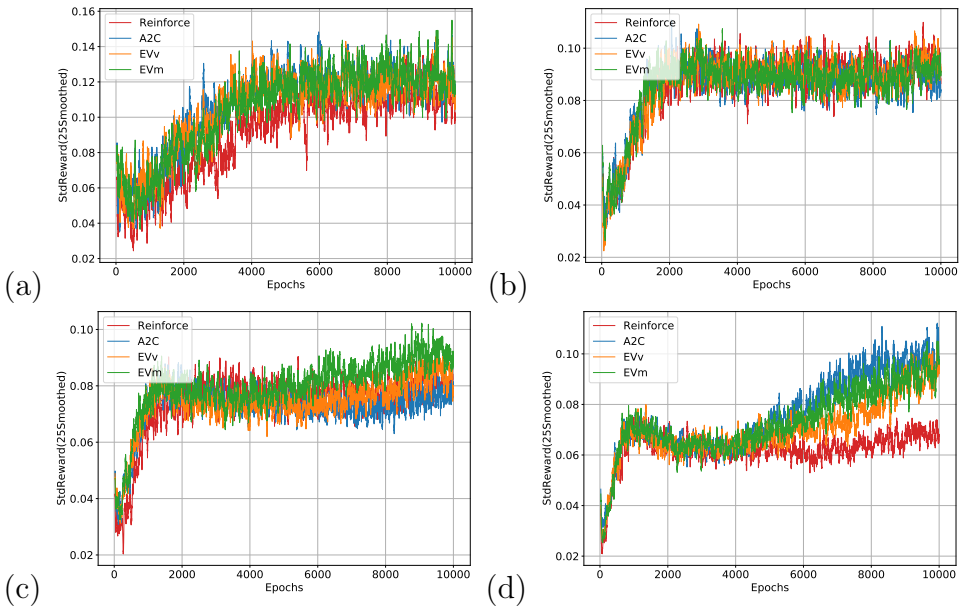


Figure C.5: The charts representing the standard deviation of the rewards for cases $K = 5$(a), $K = 10$(b), $K = 15$(c), $K = 20$(d). The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25.

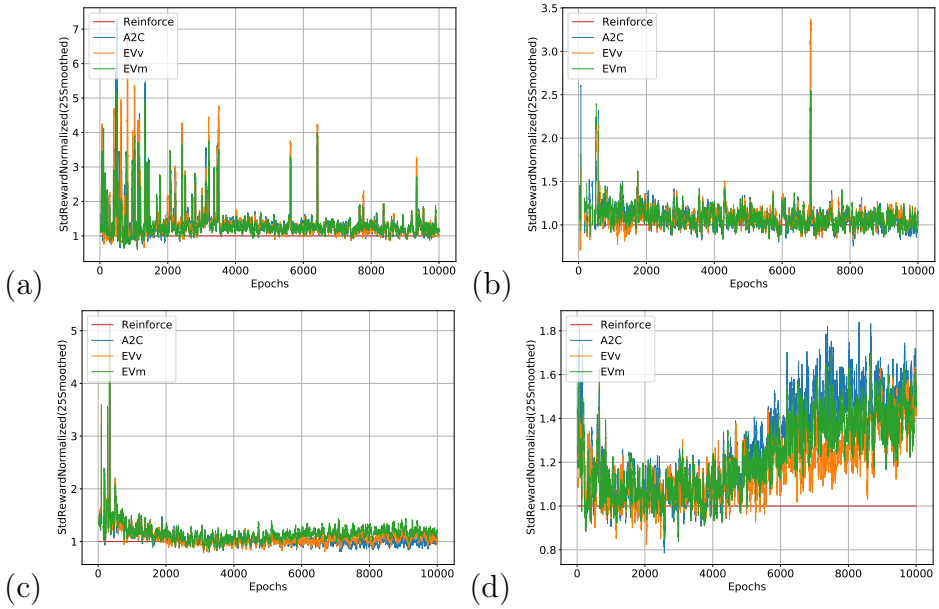
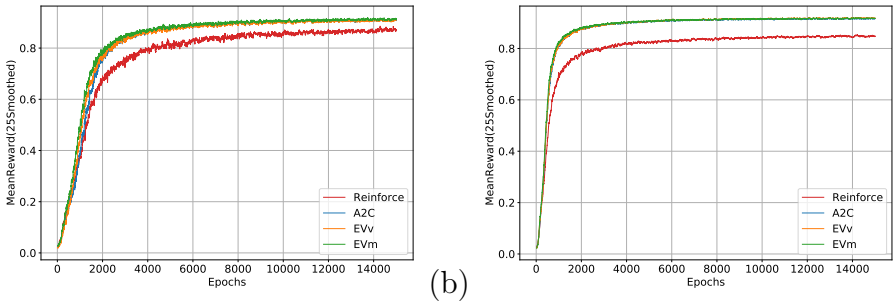


Figure C.6: The charts representing the standard deviation of the rewards normalized by the standard deviation of the REINFORCE for cases $K = 5(a)$, $K = 10$(b), $K = 15$(c), $K = 20$(d). The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 5.

Unlock

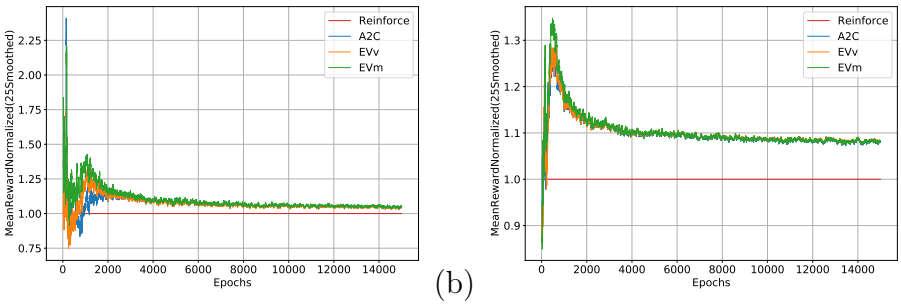
The agent has to open a locked door. First, it has to find a key and then go to the door.

In this environment we considered two different sample sizes: $K = 5$ and $K = 20$. Here EV agents and A2C seem to converge to the same policy (see Fig. C.7 and Fig. C.8). The charts on Fig. C.9 indicate that the variance is reduced 10-100 times similarly for A2C- and EV-algorithms. Considering mean rewards we clearly see that such reduction results in considerable gain of about 10-20% and, what is important, it considerably adds to the stability which is displayed on Fig. C.11 and C.12.



(a) (b)

Figure C.7: The charts representing mean rewards in Unlock environment, standing for absolute values. The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25.



(a) (b)

Figure C.8: The charts representing mean rewards in Unlock environment, normalized by the mean reward of the REINFORCE. The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25.

Important thing to notice is that in the beginning (before approximately 2000 Epochs passed) we observe small gain of EV over A2C (especially in case of smaller sample with $K = 5$) and it goes together with more stability which is indicated by the plots of standard deviation. Hence, a clever use of EV method instead of A2C sometimes can give an additional confident gain despite the fact that the gradient variance reduction is almost the same.

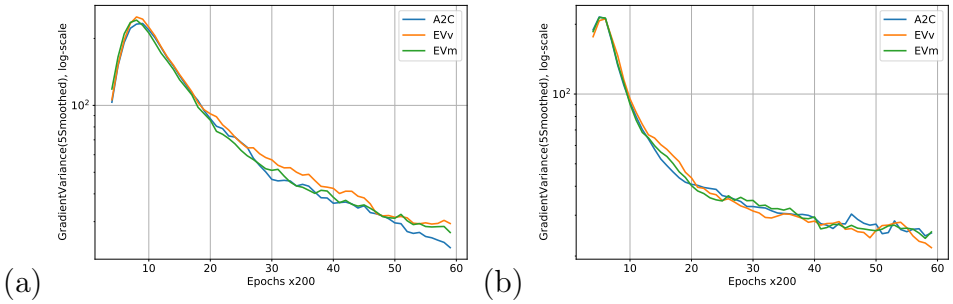


Figure C.9: The charts representing the variance of the gradient estimator in absolute values. The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 5.

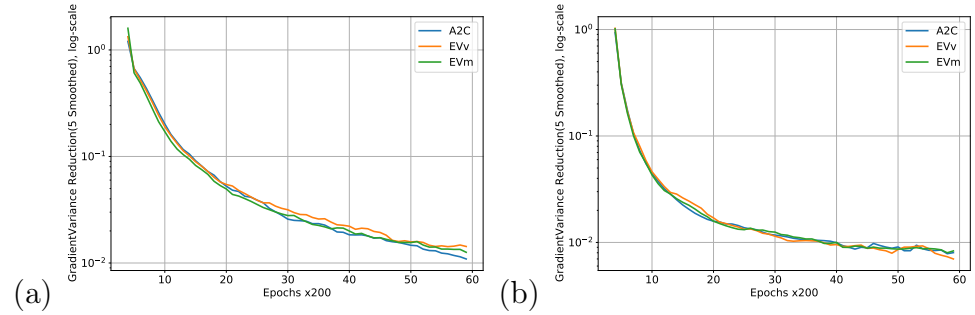


Figure C.10: The charts representing the variance of the gradient estimator normalized by the gradient variance of the REINFORCE (log-scale is set up along y-axis). The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 5.

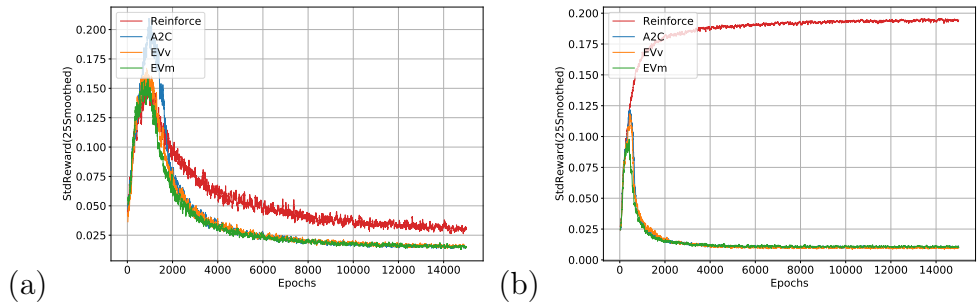


Figure C.11: The charts representing the standard deviation of the rewards. The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25.

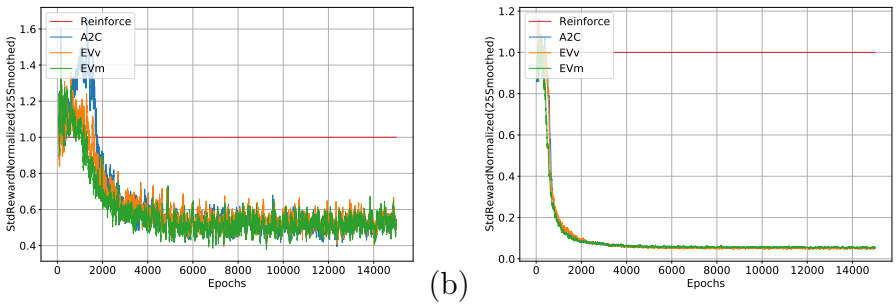


Figure C.12: The charts representing the standard deviation of the rewards normalized by the standard deviation of the REINFORCE. The results are averaged over 20 runs. The resulting curves are smoothed with sliding window of size 25.

C.2.2 OpenAI Gym: Cartpole-v1

CartPole is a Gym environment where a pole is attached by a joint to a cart, which moves along x-axis. Agent can apply a force +1 or -1 to the cart making it move right or left. The pole starts upright and the agent has to keep it as long as possible preventing from falling. The agent receives +1 reward every timestamp that the pole remains upright. The episode ends when the pole is more than 15 degrees from vertical, or the cart moves more than 2.4 units from the center.

In this environment we demonstrate 5 configurations with different policy and baseline architectures to look how algorithms behave with changing policy and baseline configurations. The exact config-files can be found on GitHub [37]. The measurements of mean rewards are averaged over 40 independent runs of the algorithms and reward variance is measured as the sample variance of the observed rewards in each epoch. We provide the charts relative to REINFORCE which are obtained by dividing the curves by the corresponding values of REINFORCE. These allow to see the improvements over REINFORCE more clearly.

Cartpole config1 (see Fig.C.13) has two hidden layers in policy network with 128 neurons each and 1 hidden layer in baseline network with 128 neurons. We assume, that is a medium complexity setting for this environment. Both networks have ReLU activations.

We can observe here that even with simple configuration EV agents have similar or slightly higher rewards, achieving about 500 points and decrease rewards variance significantly showing that EV methods are more stable than A2C and do not have many deep falls during the training as A2C or REINFORCE. It is clearly an effect of the gradient variance which is reduced drastically: almost 100-1000 times.

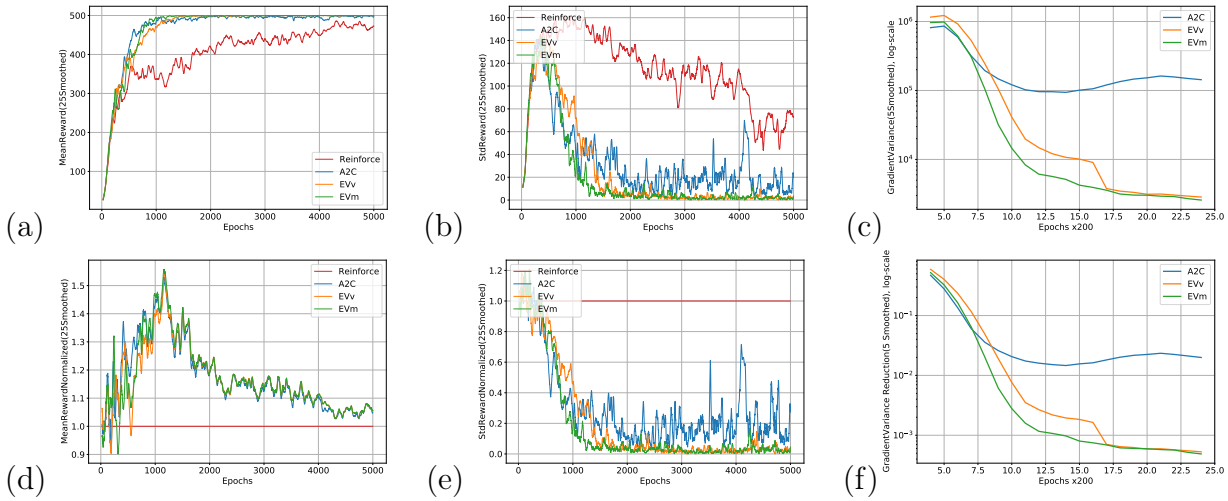


Figure C.13: The charts representing the results of the experiments in CartPole environment (config1): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d,e) the first two quantities are shown relative to REINFORCE and (f) shows gradient variance reduction ratio.

In config5 (see Fig.C.14) we keep the architecture from config1, but change the activation function with MISH. The results are almost the same: EV-agents show a little

predominance over A2C, preserving the least reward variance and gradient variance reduction among all the algorithms.

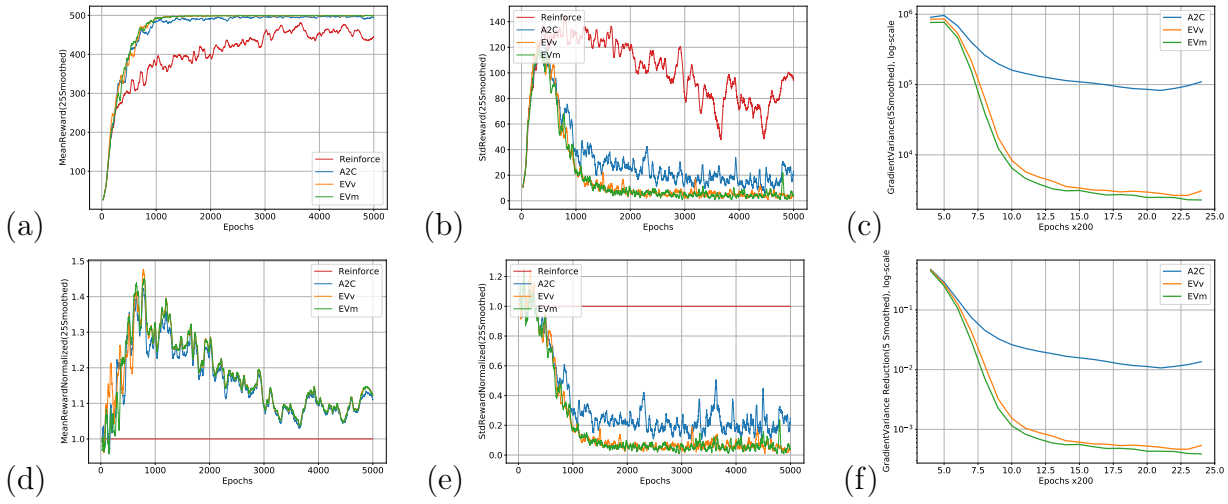


Figure C.14: The charts representing the results of the experiments in CartPole environment (config5): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d,e) the first two quantities are shown relative to REINFORCE and (f) shows gradient variance reduction ratio.

In config7 (see Fig.C.15) we move towards more complex architecture of baseline function: now it has 3 hidden layers of 128, 256, 128 neurons respectively. EV agents demonstrate better performance but this increment is rather small. Nevertheless, reward variance and gradient variance again remain the best in EV-methods.

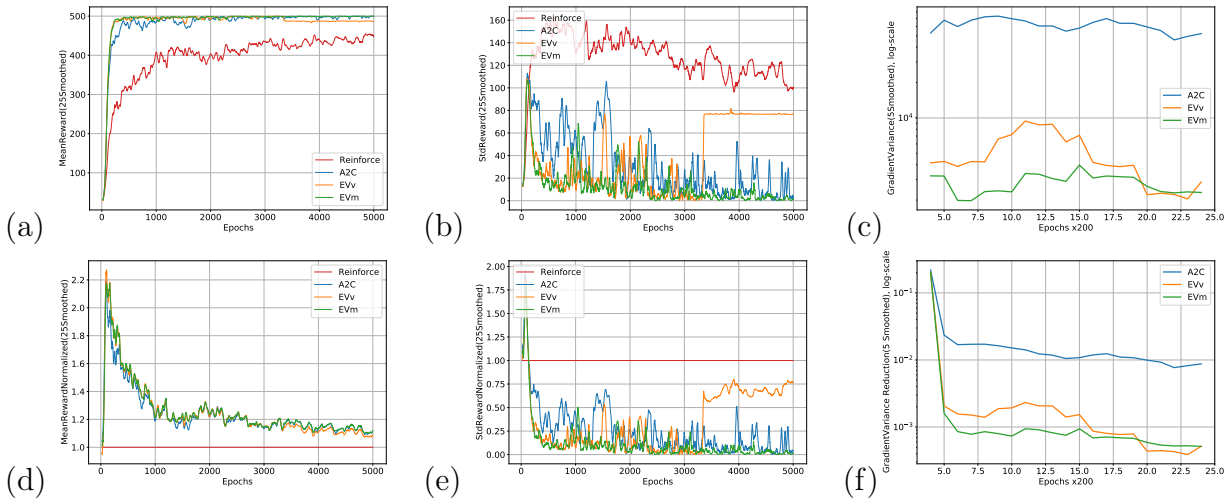


Figure C.15: The charts representing the results of the experiments in CartPole environment (config7): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d,e) the first two quantities are shown relative to REINFORCE and (f) shows gradient variance reduction ratio.

In config8 (see Fig.C.16) we address more complex setting of policy, adding two layers. The policy network has finally 3 hidden layers with MISH activation with 64, 128, 256 neurons respectively. This change greatly increases efficiency of EV algorithms enabling to achieve more than 400 points of reward and demonstrating a big dominance over A2C

which is unable to train such a complex policy to have similar performance. At the same time, reward variance of EVs remains at the level of REINFORCE while A2C level exceeds it by almost 30-50%.

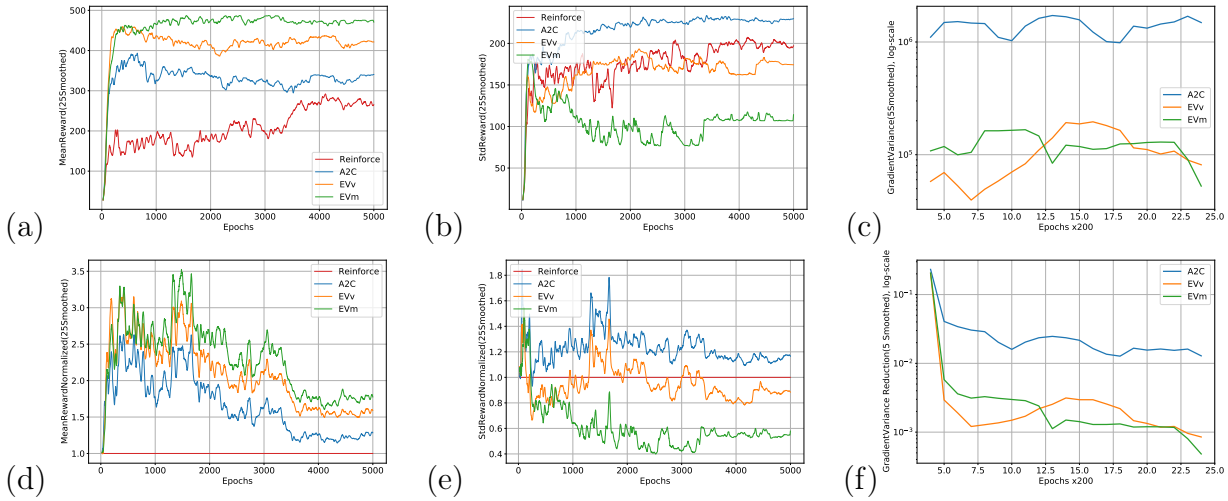


Figure C.16: The charts representing the results of the experiments in CartPole environment (config8): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d,e) the first two quantities are shown relative to REINFORCE and (f) shows gradient variance reduction ratio.

If config9 we again preserve architecture settings changing only activation from MISH to ReLU. We observe a small difference in mean rewards but another activation function clearly helped in A2C training. Regardless, EV-methods are still predominant: more stable, with less gradient variance and with higher rewards achieved.

In conclusion, our experiments show that EV methods are sometimes considerably better in terms of mean rewards than A2C, or work at least as A2C. Study of the reward variance shows that EV-methods in CartPole are considerably more stable and do not have deep falls as in A2C or Reinforce. This study allows us to judge about the stability of the training process in case of EV algorithms and claim that they are able to perform better than A2C if more complex policies are used.

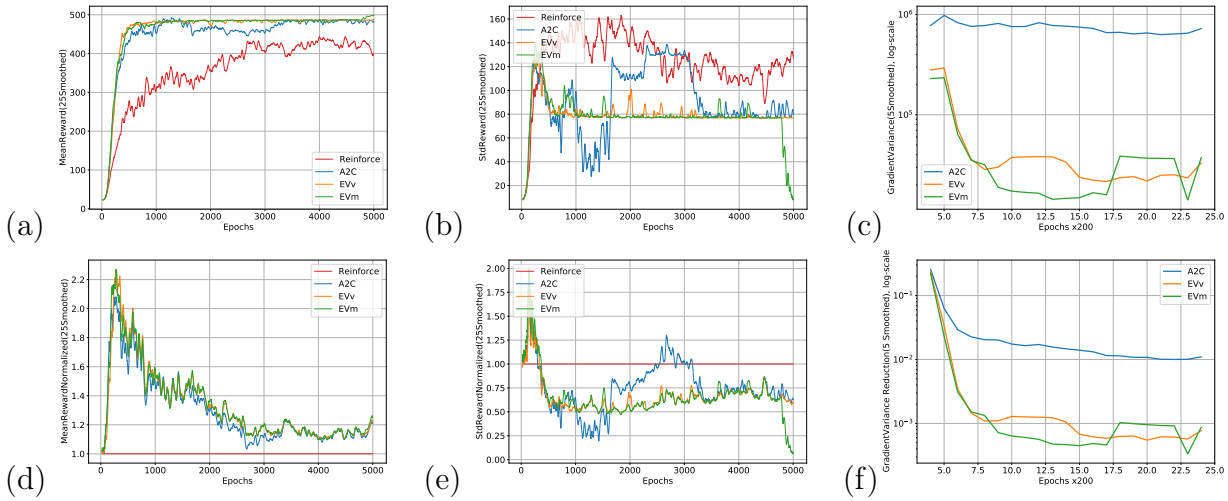


Figure C.17: The charts representing the results of the experiments in CartPole environment (config9): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d,e) the first two quantities are shown relative to REINFORCE and (f) shows gradient variance reduction ratio.

C.2.3 OpenAI Gym: LunarLander-v2

LunarLander is a console-like game where the agent can observe the physical state of the system and decide which engine to fire (the primary one at the bottom or one of the secondaries, the left or right). There are 8 state variables: two coordinates of the lander, its linear velocities, its angle and angular velocity, and two boolean values that show whether each leg is in contact with the ground.

LunarLander (see Fig. C.18) is the example of the case where all algorithms work in the same way and there is no significant difference between A2C and EV. It happens regardless to the policy type we choose; the final performances are different among the configs but inside one config A2C and EV gave the same result. We see that all algorithms behave similarly in variance reduction as well, showing that EV-methods are still good but sometimes A2C works with the same result.

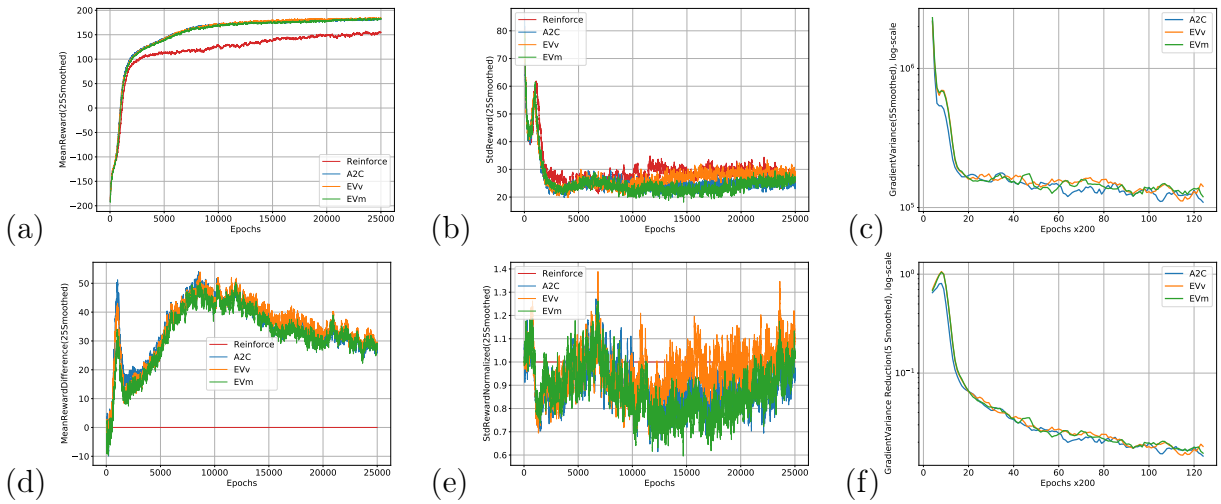


Figure C.18: The charts representing the results of the experiments in LunarLander environment (config1): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d) the difference between the algorithm and REINFORCE is shown, (e) shows the standard deviation of the rewards relative to REINFORCE and (f) shows gradient variance reduction ratio.

C.2.4 OpenAI Gym: Acrobot-v1

The system consists of two links forming a chain, with one end of the chain fixed. The joint between the two links is actuated. The goal is to apply torques on the actuated joint to swing the free end of the linear chain above a given height while starting from the initial state of hanging downwards. The actions can be to apply ± 1 or 0 torque to the joint and the goal is to have the free end reach a designated target height in as few steps as possible, and as such all steps that do not reach the goal incur a reward of -1.

The config we show here and in the main text (see Fig. C.19) is an example where EV can boost training sometimes and that a clever combination of EV and A2C may result in even better algorithms than these three. We can clearly see that until the agent reaches reward ceiling there is a clear predominance of EVm over EVv and A2C but in the end they result in the same policy. It can be seen that standard deviation of the rewards indicate positive effect in the same time. Still, it must be noted that variance reduction is the best in EVm and EVv until the ceiling is reached. Hence, the environment itself does not require so excessive variance reduction and there is still an open space for discussions about whether the variance reduction needed in such environment.

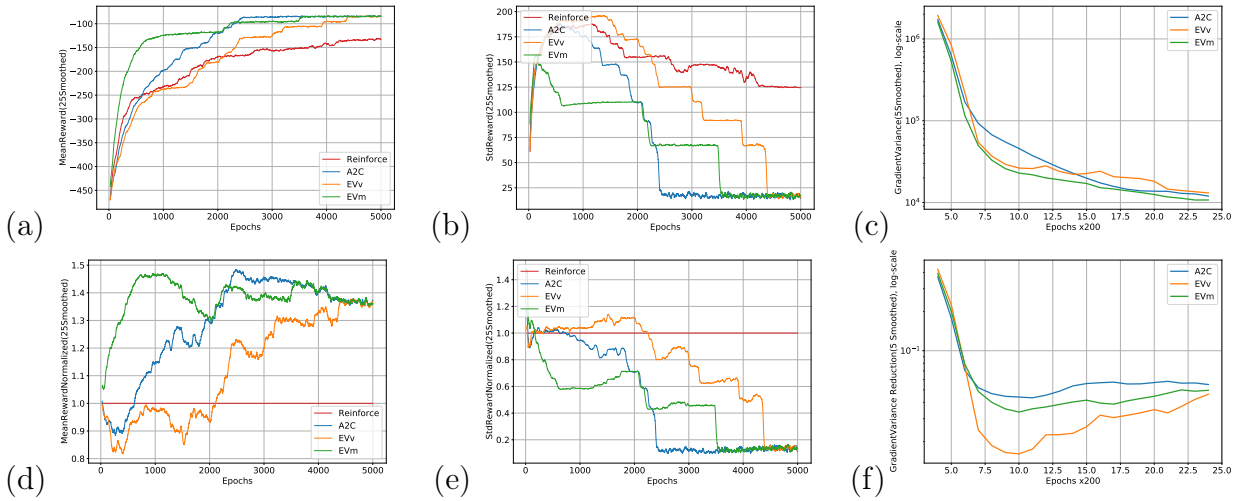


Figure C.19: The charts representing the results of the experiments in Acrobot environment (config1): (a) displays mean rewards, (b) shows standard deviation of the rewards, (c) depicts gradient variance, in (d) the difference between the algorithm and REINFORCE is shown, (e) shows the standard deviation of the rewards relative to REINFORCE and (f) shows gradient variance reduction ratio.

C.2.5 Time Complexity Discussion

In Figures C.20 - C.27 we demonstrate how training time depends on processed transitions for all environments. One can clearly see that EV algorithms are more time-consuming and sensitive to the growth of the sample size K but its excessive training time mostly can be explained by the implementation. PyTorch-compatibility requires that we have to make K extra backpropagations in order to compute empirical variance. We believe that this part of computations can be optimized and, therefore, accelerated in practice.

Scatter plots demonstrate how consumed time depends on the number of processed data. This allows better understanding of the processing cost of one transition from the simulated trajectory. We also provide the measured execution times per transition in box plots to see the difference between all considered algorithms regardless of the trajectory length. We used high-performance computing units with the same computation powers for each run inside one environment, so that these measurements were accurate and comparable.

Summing up, considering all the advantages of EV algorithms, they have higher time costs (see also Figures C.24 - C.27) and demand more specific implementation allowing faster computation of many gradients which currently cannot be easily developed in the framework of PyTorch. PyTorch allows great flexibility and very general models for the approximations of policy and baseline; if these are more specific, our algorithm can be implemented to be more effective.

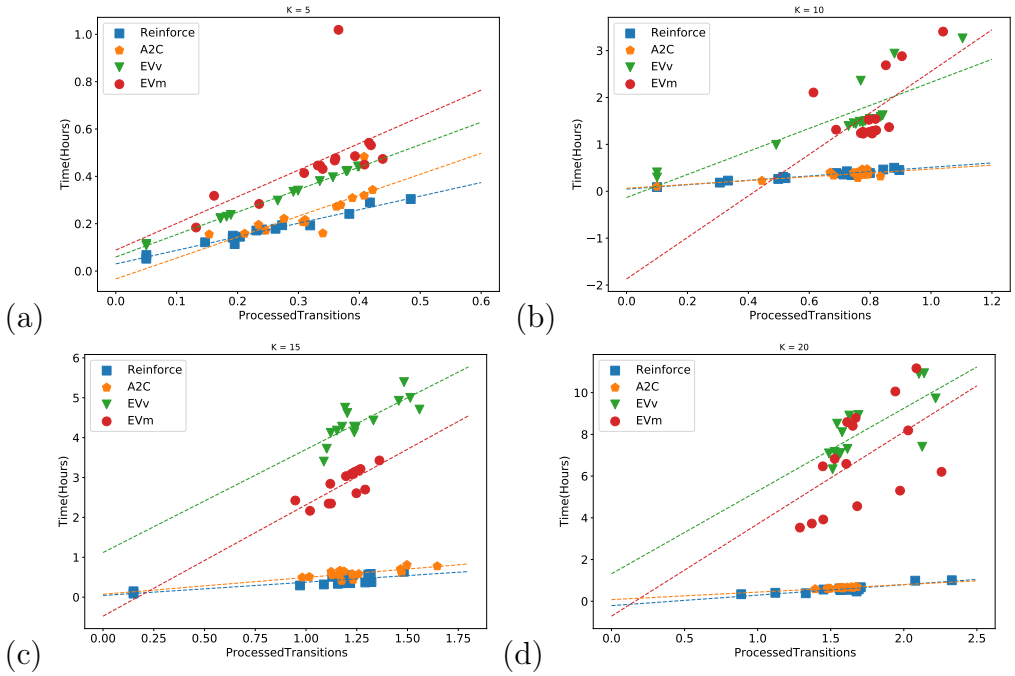
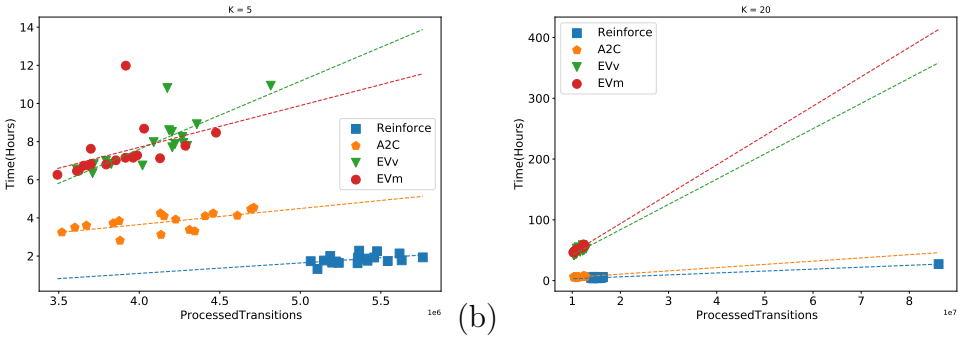
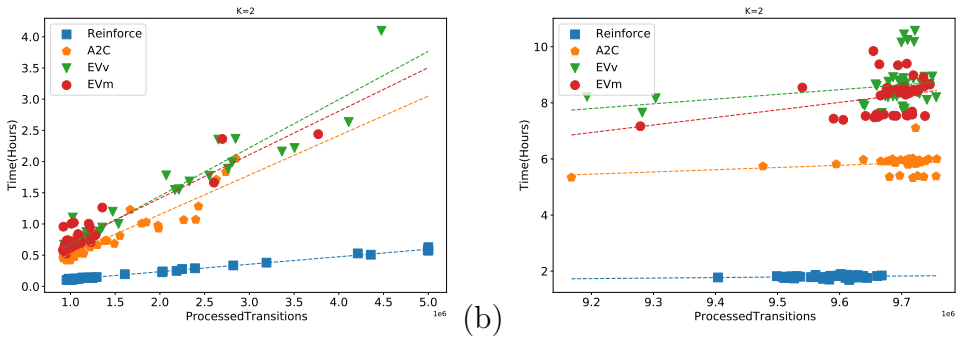


Figure C.20: The charts representing dependency of training time from number of the processed transitions(scale of millions) for GoToDoor



(a) (b)

Figure C.21: The charts representing dependency of training time from number of the processed transitions(scale of millions) for Unlock



(a) (b)

Figure C.22: The charts representing dependency of training time from number of the processed transitions(scale of millions) for Acrobot (a) and for LunarLander (b)

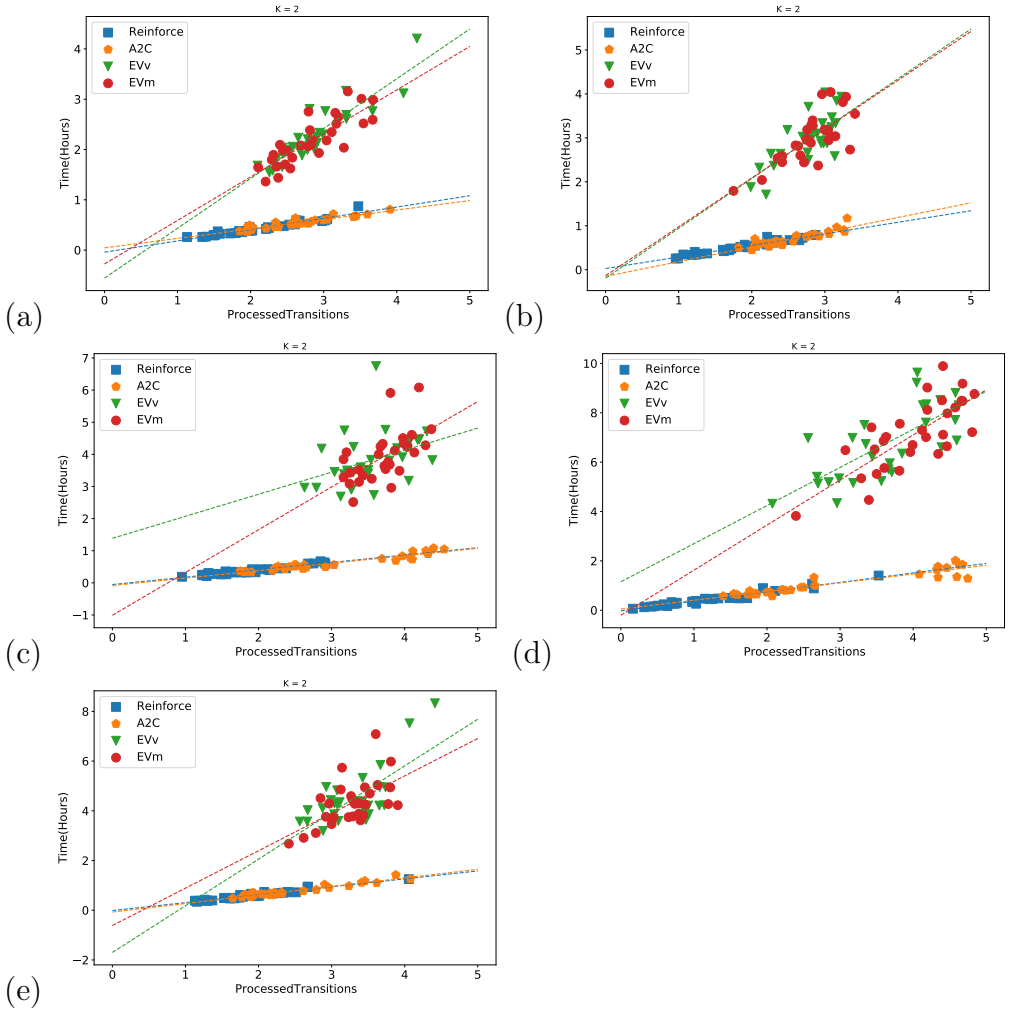


Figure C.23: The charts representing dependency of training time from number of the processed transitions(scale of millions) for CartPole environment: (a) config1, (b) config5, (c) config7, (d) config8, (e) config9

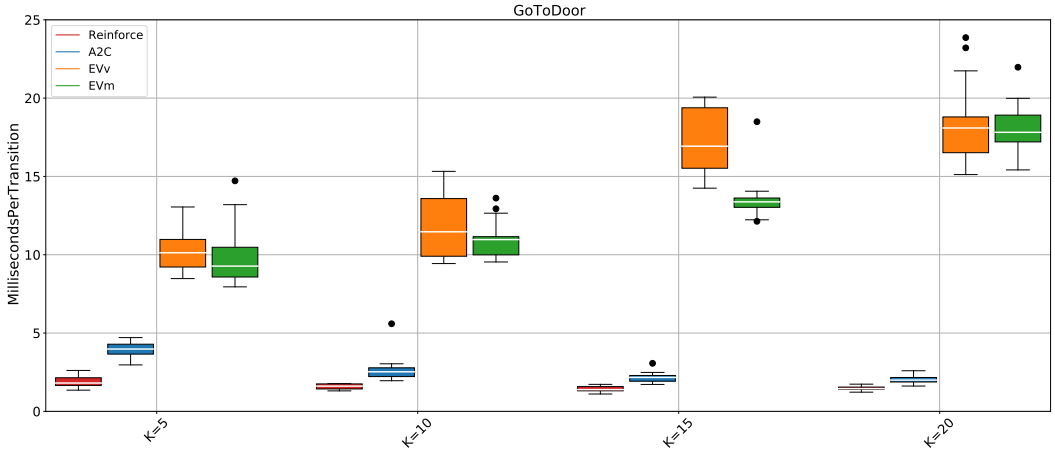


Figure C.24: The charts representing distribution of time per transition (scale of milliseconds) w.r.t. number of trajectories used for training in GoToDoor environment, $K = 5, 10, 15, 20$

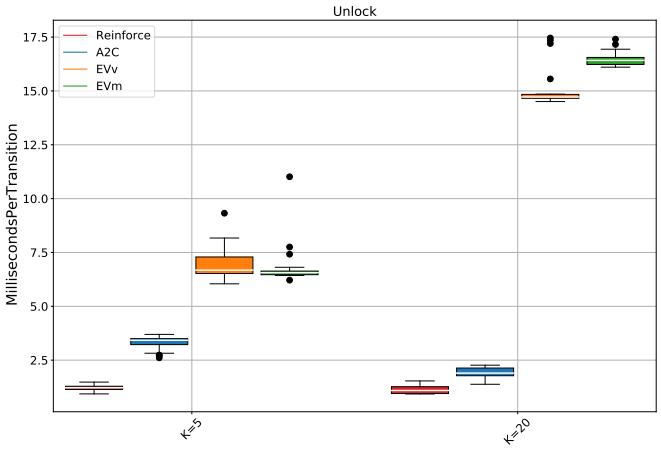
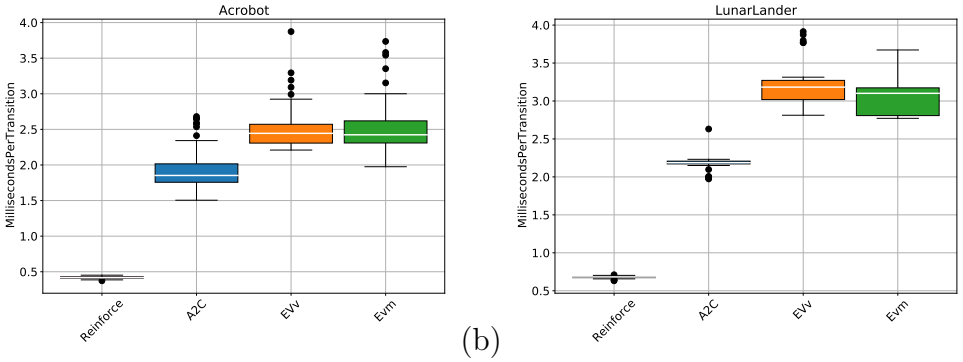


Figure C.25: The charts representing distribution of time per transition (scale of milliseconds) w.r.t. number of trajectories used for training in Unlock environment, $K = 5, 20$



(a) (b)

Figure C.26: The charts representing distribution of time per transition (scale of milliseconds) w.r.t. an algorithm for (a) Acrobot and for (b) LunarLander

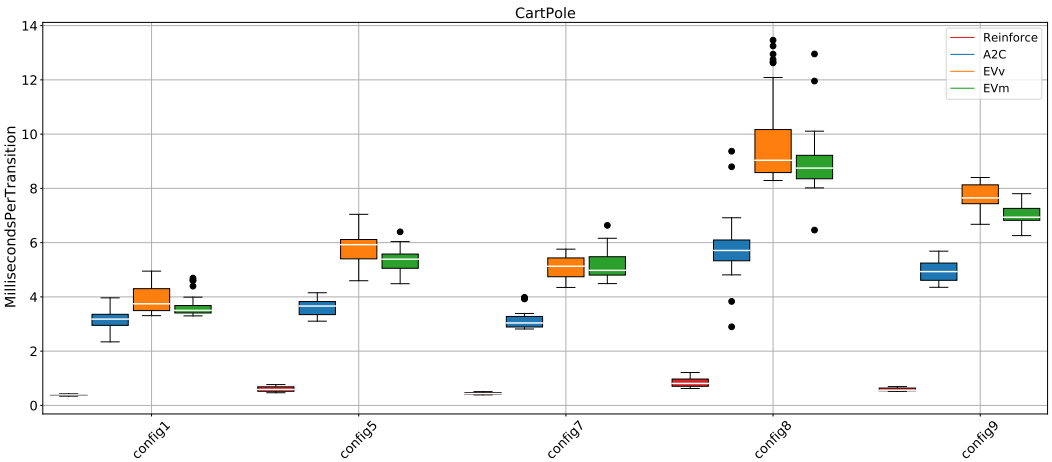


Figure C.27: The charts representing distribution of time per transition (scale of milliseconds) w.r.t. config number, CartPole environment.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Le développement et l'analyse théorique des algorithmes de contrôle optimal et d'apprentissage par renforcement

Mots clés : Apprentissage par renforcement, Contrôle optimal, Apprentissage automatique

Résumé : Dans la thèse, nous abordons les problèmes de l'arrêt optimal et l'apprentissage dans les processus décisionnel de Markov utilisés en apprentissage par renforcement (Reinforcement Learning, RL). Dans la première direction, nous dérivons des estimations de complexité pour l'algorithme appelé Weighted Stochastic Mesh (WSM) et donnons une nouvelle méthodologie pour le compairison de la complexité des algorithmes d'arrêt optimale avec l'indice de semi-tractabilité. Nous montrons que WSM est optimal par rapport à ce critère, quand les méthodes de régression couramment utilisées ne sont pas très bien.
Pour l'apprentissage par renforcement nous donnons une analyse de convergence non asymptotique d'un schéma d'approximation stochastique à deux échelles de temps – Gradient TD- sous des hypothèses de bruit ≪ incrément de martingale ≫ - buffer replay - et de Markov. Nous obtenons des bornes supérieures qui sont optimales en taux en construisant une méthode de développement de l'erreur, qui permet d'obtenir un contrôle précis des restes.
Nous présentons aussi un nouvel algorithme de réduction de variance dans les schémas de ≪ policy-gradient ≫. L'approche proposée basée sur la minimisation d'un estimateur de la variance empirique des récompeneses pondérées. Nous avons établi théoriquement et pratiquement un gain par rapport à la méthode classique d'acteur-critique (A2C).

Title : Development and Theoretical Analysis of the Algorithms for Optimal Control and Reinforcement Learning

Keywords : Reinforcement learning, Machine Learning, Optimal Control

Abstract : In this PhD dissertation, we address the problems of optimal stopping and learning in Markov decision processes used in reinforcement learning (RL). In the first direction, we derive complexity estimates for the algorithm called Weighted Stochastic Mesh (WSM) and give a new method for comparing the complexity of optimal stopping algorithms with the semi tractability index. We show that WSM is optimal with respect to this criterion when the commonly used regression methods are much less effective.
For reinforcement learning, we give a non-asymptotic convergence analysis of a stochastic approximation scheme with two time scales - gradient TD - under assumptions of "martingale increment" noise - buffer replay - and of "Markov noise" (when learning is done along a single run). We obtain upper bounds that are rate-optimal by constructing an error expansion method that provides accurate control of the remainders terms.
We also present a new algorithm for variance reduction in policy gradient schemes. The proposed approach is based on minimising an estimator for the empirical variance of the weighted rewards. We establish theoretical and practical gains over the classical actor-critic (A2C) method.

Institut Polytechnique de Paris
91120 Palaiseau, France