National Research University Higher School of Economics

as a manuscript

Maksim Kaledin

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

Summary of the PhD thesis
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow - 2023

The PhD thesis was prepared at National Research University Higher School of Economics.

Academic supervisors:

Denis Belomestny, Candidate of Sciences, professsor, HSE University (Moscow, Russia), University of Duisburg-Essen (Essen, Germany)

Eric Moulines, PhD, professor, École Polytechnique, Institut Polytechnique de Paris (Paris, France).

Introduction

Stochastic optimal control problems are very often encountered in various practical areas: from finance [28, 64] to engineering [9]. 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 [58].

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 [58]) 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, [64, 28]) Assume $T > 0$ and let process X_t be set with an Ito SDE for $t \in [0, T)$

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

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)[44] and Tsitsiklis-Van Roy [67] 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 [18] and [75] general convergence properties of the method were established.

Problem 2. (Markov Decision Process, MDP, [58]) 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 [69, 8, 55] and going to energy management systems [41, 26], manufacturing and robotics [2] to name a few. Naturally, RL gives the practitioners new sets of control tools for any kind of automatization [27].

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 [51]. SA itself currently became a well-studied technique [7, 39, 12], 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 [57] for which finite time analysis has been reported in [56, 40, 10, 20].

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 1.1, we are aiming at presenting the complexity analysis of Weighted Stochastic Mesh(WSM) algorithm similar to the method of [13] 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 1.2 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 [57]) and gradient temporal difference algorithms (GTD[59],GTD2 and TDC [60]). 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 1.3 we set up to propose a new method for variance reduction based on empirical variance minimization of [5] 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 [61] 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 [13] 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 [44] and QTM [4].

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 [33, 19, 74, 22]. Unlike the prior works [42, 19, 74], 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 [5], 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 [68]: 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.

1 Contents

1.1 Semitractability of Optimal Stopping Problem via Weighted Stochastic Mesh Algorithm

The results of this section are published in [6].

1.1.1 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 [28]. 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 (1) 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 [49] 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 [75, 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 [67]. In [24] 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 [29] 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* [49], if there is an algorithm to solve it with complexity $\mathcal{C}(\varepsilon, d)$ satisfying

$$\lim_{d+\varepsilon^{-1}\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = 0. \tag{2}$$

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 [65] 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 [13], 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 [52] 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 Lipshitz continuity of transition densities with Lipschitz constant independent on the dimension d.

1.1.2 Complexity Metrics

It turns out that the criterion (2) 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*

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

is called semitractability index.

Definition 4. *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.

1.1.3 WSM Algorithm

Let us present a Weighted Stochastic Mesh (WSM) algorithm for a discrete-time optimal stopping problem. The algorithm is inspired by [13] 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:

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

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

$$\widetilde{U}_L(Z_L) = g_L(Z_L), \tag{3}$$
$$\widetilde{U}_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1})\ \mid\ Z_l\right]\right\}\cdot \mathbb{1}_{B_R}(Z_l),\quad l = 0,..,L-1, \tag{4}$$

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 transistion density $p(y|x)$. To estimate the conditional expectations, we use the following approximation:

$$\mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1})\ \mid\ Z_l = x\right] \approx \sum_{n=1}^{N} \widetilde{U}_{l+1}\left(Z_{l+1}^{(n)}\right)\frac{p\left(Z_{l+1}^{(n)}\mid x\right)}{\sum_{m=1}^{N} p\left(Z_{l+1}^{(n)}\mid Z_l^{(m)}\right)}. \tag{5}$$

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 (4) and (5) 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 (4) and (5) 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 $c_f^{(d)} N^2 L$.

1.1.4 Main Results

Using the bounds from the literature we have computed $\Gamma_{\mathcal{C}}$ for two popular in practice methods (Longstaff-Schwarz[44] and Quantization Tree [4], 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 [6]) 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.

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

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 [6]) 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

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

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 [6]). 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 [6]) 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

$$\mathcal{C}(\varepsilon,d)=c_1\overline{\alpha}^2c_g^4\overline{\kappa}^2c_f^{(d)}c_2^d\frac{T^{d+7}}{h^{d+5}}\varepsilon^{-4}\times\ln^{d+2}\left[\frac{(T/h)\left(1+c_{\overline{X}}+c_{\overline{X}}\left\|x_0\right\|_2\right)e^{\frac{c_{\overline{X}}\sqrt{\overline{\alpha}T}}{1+c_{\overline{X}}+c_{\overline{X}}\|x_0\|_2}}2^{3/4}\left(c_g\overline{\kappa}\vee1\right)}{\varepsilon}\right]$$

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

$$\mathcal{C}^\star(\varepsilon,d)=c_1\overline{\alpha}^2c_g^4\overline{\kappa}^2c_f^{(d)}c_2^d\frac{T^{d+7}}{\varepsilon^{2d+14}}\times\ln^{d+2}\left[\frac{T\left(1+c_{\overline{X}}+c_{\overline{X}}\left\|x_0\right\|_2\right)e^{\frac{c_{\overline{X}}\sqrt{\alpha T}}{1+c_{\overline{X}}+c_{\overline{X}}\|x_0\|_2}}2^{3/4}\left(c_g\overline{\kappa}\vee1\right)}{\varepsilon}\right].$$

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 [6] 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.

1.1.5 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

$$X_t=X_0e^{\sigma W_t+(r-\sigma/2)t}$$

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 [36] to be 6.9320. In Fig. 1 we show the lower bounds obtained by WSM, LS and VF (value function regression method of [67]) 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.

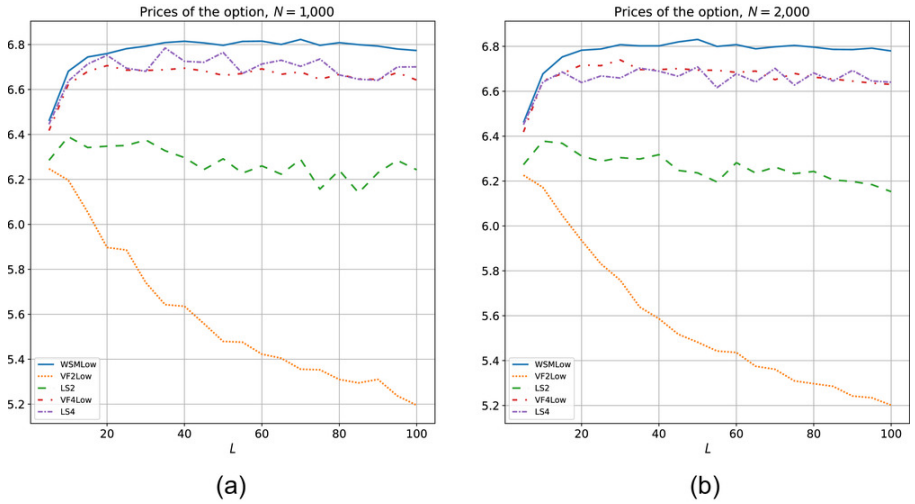

Figure 1: Lower bounds for the price of one-dimensional American put option approximated using different methods and uniform time discretization $t_k = kT/L, k = 0, .., L$ of exercise dates. The numbers of training paths are $N_{train} = 1000$(a) and $N_{train} = 2000$(b) and the number of test trajectories 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).

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

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

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. 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.

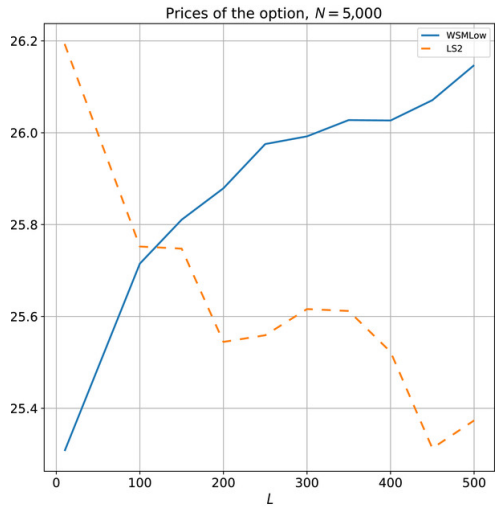

Figure 2: Lower bounds for the price of a five-dimensional American put option approximated using a uniform grid $t_k = kT/L, k = 0, .., L$ of exercise dates. The number of training paths is $N_{train} = 2000$ and the number of test trajectories is $N_{test} = 5000$.

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

The results of this section are published in [35].

1.2.1 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 [3, 66]). To circumvent this problem, [59, 60] 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 [11]:

$$\theta_{k+1} = \theta_k + \beta_k\{\widetilde{b}_1(X_{k+1}) - \widetilde{A}_{11}(X_{k+1})\theta_k - \widetilde{A}_{12}(X_{k+1})w_k\}, \tag{6}$$

$$w_{k+1} = w_k + \gamma_k\{\widetilde{b}_2(X_{k+1}) - \widetilde{A}_{21}(X_{k+1})\theta_k - \widetilde{A}_{22}(X_{k+1})w_k\}. \tag{7}$$

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 (6),(7). The almost-sure convergence of two timescale SA has been established in [11, 62, 63, 12], among others and [38, 48] characterized the asymptotic convergence rates. However, finite-time risk bounds for two timescale SA have not been analyzed until recently. With martingale samples, [42] provided the first finite time analysis of GTD method, [21, 19] 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, [33] studied the finite time expected error bound with constant step sizes; [74] and [22] provided similar analysis for general step sizes. It is important to notice that with homogeneous martingale noise, the asymptotic rate of (6), (7) without a projection step, as shown in [38, 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 [19]. 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.

1.2.2 Main Results

We investigate the linear two timescale SA given by the following equivalent form of (6), (7):

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{11}\theta_k - A_{12}w_k + V_{k+1}), \tag{8}$$

$$w_{k+1} = w_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}w_k + W_{k+1}), \tag{9}$$

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:

$$\begin{aligned} V_{k+1} &:= \widetilde{b}_1(X_{k+1}) - b_1 - (\widetilde{A}_{11}(X_{k+1}) - A_{11})\theta_k - (\widetilde{A}_{12}(X_{k+1}) - A_{12})w_k, \\ W_{k+1} &:= \widetilde{b}_2(X_{k+1}) - b_2 - (\widetilde{A}_{21}(X_{k+1}) - A_{21})\theta_k - (\widetilde{A}_{22}(X_{k+1}) - A_{22})w_k. \end{aligned} \tag{10}$$

The goal of the recursion (8), (9) is to find a stationary solution pair $(\theta^\star, w^\star)$ that solves the system of linear equations:

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

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

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

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

To analyze the convergence of $(\theta_k, w_k)_{k\geq 0}$ in (8), (9) to $(\theta^\star, w^\star)$, we require several assumptions which are common for linear two time-scale SA, see [38].

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

$$\gamma_k/\gamma_{k+1} \leq 1+(a_{22}/8)\gamma_{k+1}, \;\; \beta_k/\beta_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}, \;\; \gamma_k/\gamma_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}. \tag{13}$$

Our conditions on step sizes are similar to [38, 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 A2 is

$$\beta_k = c^\beta/(k + k_0^\beta), \quad \gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3} \tag{14}$$

with some constants c^β, c^γ, k_0^γ, k_0^β, e.g., as suggested in [21, Remark 9]; or a constant step size of $\beta_k = \beta, \gamma_k = \gamma$; or a piecewise constant step size, e.g., [33].

We present new results on the convergence rate of (8), (9) 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:

$$\mathcal{F}_k := \sigma\{\theta_0, w_0, X_1, X_2, ..., X_k\}, \quad \mathrm{V}_0 := \mathbb{E}\left[\|\theta^0 - \theta^\star\|^2 + \|w^0 - w^\star\|^2\right]. \tag{15}$$

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

$$\begin{aligned}\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\| &\le m_V(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|),\\ \|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| &\le m_W(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|)\,.\end{aligned} \tag{16}$$

Theorem 5. *Assume A1–4 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*

$$\mathbb{E}\left[\|\theta_k - \theta^*\|^2\right] \le d_\theta\left\{\mathrm{C}_0^{\tilde{\theta},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\tilde{\theta},\mathsf{mtg}}\beta_k\right\} \tag{17}$$

$$\mathbb{E}\left[\left\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\right\|^2\right] \le d_w\Big\{\mathrm{C}_0^{\widehat{w},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\widehat{w},\mathsf{mtg}}\gamma_k\Big\} \tag{18}$$

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 (6), (7) converges to a unique fixed point defined by the above mean field vectors/matrices, see (12). 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

$$\widetilde{b}_i(x) - b_i = \widehat{b}_i(x) - \mathrm{P}\,\widehat{b}_i(x), \quad \widetilde{A}_{ij}(x) - A_{ij} = \widehat{A}_{ij}(x) - \mathrm{P}\,\widehat{A}_{ij}(x) \tag{19}$$

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 B1 together with some regularity conditions, see [23, Section 21.2]. Moreover,

B 3. Under B2, the vector/matrix valued functions $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ are uniformly bounded: for any $i, j = 1, 2$, $x \in \mathsf{X}$,

$$\|\widehat{b}_i(x)\| \leq \overline{\mathrm{b}}, \ \|\widehat{A}_{ij}(x)\| \leq \overline{\mathrm{A}}. \tag{20}$$

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 B3, 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, B3 implies A4. Meanwhile, B4 imposes further restriction on the step size. The latter can also be satisfied by (14). 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 A1–2, B1–4 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*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq d_\theta \left\{ \mathrm{C}_0^{\tilde{\theta},\mathsf{mark}} \prod_{\ell=0}^{k-1} \left(1 - \beta_\ell \frac{a_\Delta}{8}\right)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\tilde{\theta},\mathsf{mark}} \beta_k \right\}, \tag{21}$$

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

The exact constants are given in the paper.

While Theorem 6 relaxes the martingale difference assumption A4 in Theorem 5, we remark that the results here do not generalize that in Theorem 5 due to the additional B3, B4. 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 5 and 6 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 (14); 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 (14) 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 5), the closest work to ours is [19] which analyzed the linear two timescale SA with martingale samples and diminishing step sizes. The authors improved on [21] and obtained the same convergence rate (in high probability) as our Theorem 5, 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 6), the closest works to ours are [22, 33, 74]. In particular, [33] analyzed the linear two timescale SA with constant step sizes and showed that the steady-state error for both θ_k, w_k is $\mathcal{O}(\gamma^2/\beta)$. [74] 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 (14) is used. [22] 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 6, 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 5, 6 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 A5 implies A4 and therefore poses a stronger assumption. We have

Theorem 7. *Assume A1–3, A5 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*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = I_k + J_k. \tag{23}$$

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

$$\mathrm{C}_3^{\mathsf{exp}} \operatorname{Tr}(\Sigma) \leq \frac{I_k}{\beta_k} \leq \mathrm{C}_4^{\mathsf{exp}} \operatorname{Tr}(\Sigma), \tag{24}$$

and J_k is bounded by

$$|J_k| \leq \mathrm{C}_0^{\mathsf{exp}} \prod_{\ell=0}^{k-1} \left(1 - \frac{a_\Delta}{4}\beta_\ell\right) \mathrm{V}_0 + \mathrm{C}_1^{\mathsf{exp}} \beta_k \left(\gamma_k + \frac{\beta_k}{\gamma_k}\right), \tag{25}$$

where V_0 was defined in (15). 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 (25), the dominant term for J_k is given by $\mathcal{O}(\beta_k\gamma_k + \frac{\beta_k^2}{\gamma_k})$. As such, using (24), 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 (23), (24) 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 A1–3, A5 imposed by the theorem imply A1–A4 required by Theorem 5. Hence, together with (17) in Theorem 5, the above observations constitute a *matching* lower bound on the convergence rate of linear two timescale SA with martingale noise.

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

The results of this section are published in [34].

1.4 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 [71, 61]. The straightforward way to tackle gradient estimation is Monte Carlo scheme resulting in the algorithm called REINFORCE [71]. 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:

$$\theta_{n+1} = \theta_n + \eta_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla} J|_{\theta_n}(X_n^{(k)}), \quad n = 1, 2, \ldots \tag{26}$$

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 [71] 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 [70]. 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 [37] and advantage actor-critic: A2C [61] and asynchronous version of it, A3C [46]. 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:

$$\theta_{n+1} = \theta_n + \alpha_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla}^{b_\phi} J(X_n^{(k)}), \tag{27}$$

$$\phi_{n+1} = \phi_n - \beta_n \nabla_\phi V_{K,n}^{A2C}(\phi)|_{\phi_n}, \tag{28}$$

where

$$V_{K,n}^{A2C}(\phi) := \frac{1}{K} \sum_{k=1}^{K} \sum_{t=0}^{T-1} (G_t(X_n^{(k)}) - b_\phi(S_t^{(k)}))^2 \tag{29}$$

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 [46].

Recently a new interest in such methods has emerged due to the introduction of deep reinforcement learning [47], a very comprehensive review is done in [27]. During several decades a large number of new variance reduction methods were proposed, including sub-sampling methods like SVRPG [50, 73] and various control variate approaches of [53], [32], [43], [68], [72]. There are also approaches of a bit different nature: trajectory-wise control variates [15] using the control variate based on future rewards and variance reduction in input-driven environments [45]. Apart from that, in ergodic case there were both theoretic [31] and also some practical advancements [17]. The importance of the criteria for variance reduction is well-known in Monte-Carlo and MCMC [54] and recently was also addressed in RL by [25], 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 [68]. 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 [5]. 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 [43]. Despite that, the implementation and theoretical studies of this approach are still missing in the literature.

1.4.1 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:

$$V_{n,K}^{EVv}(\theta,\phi) := \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2 - \frac{1}{K^2}\left\|\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2, \tag{30}$$

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

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

$$\theta_{n+1} = \theta_n + \alpha_n\frac{1}{K}\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi}J(X_n^{(k)}), \tag{32}$$

$$\phi_{n+1} = \phi_n - \beta_n\nabla_\phi V_K^{EV}(\phi,\theta)|_{\phi_n,\theta_n}. \tag{33}$$

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 [34]). 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:

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

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 $\alpha \geq 2$ 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 8. *Under Assumptions 6-8 it holds with probability at least* $1 - 4e^{-t}$,

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

with

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

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 8 firstly gives some reliability guarantee. Secondly, it also shows that if we have K large enough, we can reduce the variance even more.

1.4.2 Numerical Experiments

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

- Gym Minigrid [16] (`Unlock-v0, GoToDoor-5x5-v0`);
- Gym Classic Control [14] (`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 [30].

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 (31) 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) we conducted several experiments and present here two policy configurations: one with simpler neural network (config5, see Fig. 3(a,b,c)) and one with more complex network (config8, see Fig. 3(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

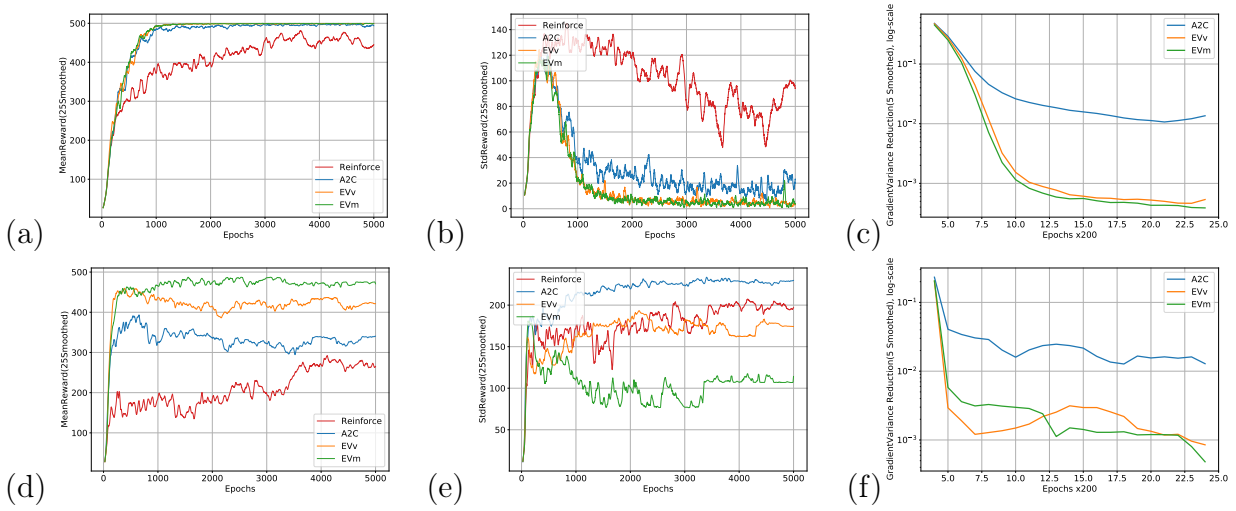

Figure 3: The charts representing the results for `CartPole` environment: (a,b,c) represent mean rewards, standard deviation of the rewards and gradient variance reduction ratio for config5 and (d,e,f) show the same information about config8.

phenomenon of better performance of EV in CartPole with more complex policies is observed often, more detailed discussion is placed in Supplementary. As to `Acrobot` (see Fig. 4(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. 5(a)) is the example of the environments where all algorithms work similarly: in terms of rewards we cannot see significant improvement even over Reinforce.

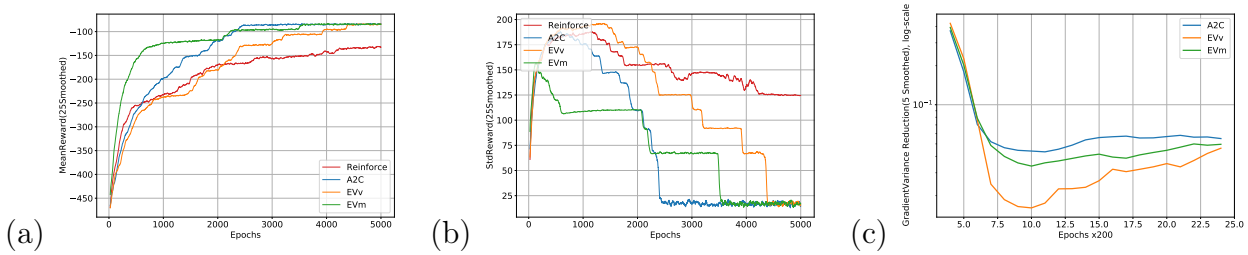

Figure 4: The charts representing the results for `Acrobot` environment: (a) depicts mean rewards, (b) shows the standard deviations of the rewards and (c) displays the gradient variance reduction ratios.

Stability of Training. When we study the charts for standard deviation of the rewards (Fig. 3(b,e),4(b),5(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(b,e) where the standard deviation is about 2 times less than in case of A2C. This holds for both configurations. Fig. 4 illustrating the experiments with `Acrobot` show that until the ceiling is reached EV methods still can have lower variance. In `Unlock` presented in Fig. 5(b) we have not observed a significant difference in reward variance.

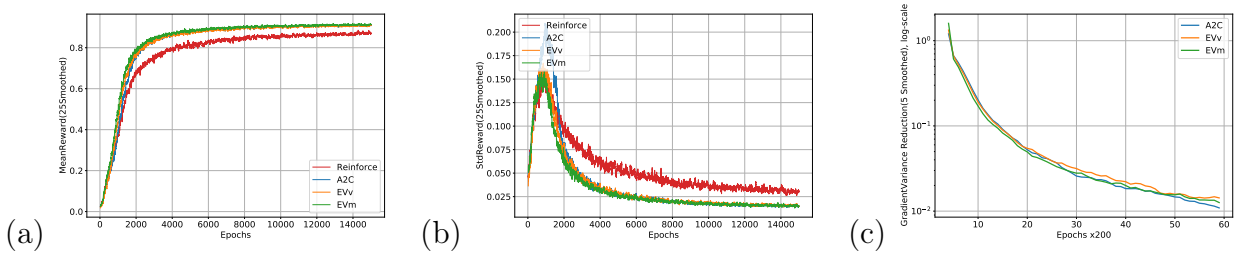

Figure 5: The charts representing the results for `Unlock` environment: (a) depicts mean rewards, (b) shows the standard deviations of the rewards and (c) displays the gradient variance reduction ratios.

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(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. 5 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] D. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific optimization and computation series. Athena Scientific, 2019.

[10] 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.

[11] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[12] Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[13] Mark Broadie and Paul Glasserman. A stochastic mesh method for pricing high-dimensional american options. *Journal of Computational Finance*, 7:35–72, 2004.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[15] 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.

[16] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[17] Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 21(52):1–51, 2020.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Thinh T Doan. Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation. *arXiv preprint arXiv:1912.10583*, 2019.

[23] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

[24] 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.

[25] 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.

[26] Vincent Francois, David Taralla, Damien Ernst, and Raphael Fonteneau. Deep reinforcement learning solutions for energy microgrids management. 12 2016.

[27] 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.

[28] Paul Glasserman. *Monte Carlo methods in financial engineering*. Springer, New York, 2004.

[29] David A. Goldberg and Yilun Chen. Beating the curse of dimensionality in options pricing and optimal stopping, 2018.

[30] Alexander Golubev and Maksim Kaledin. EVRLLib, a library implementing policy gradient algorithms in Reinforcement Learning. `https://github.com/DJAlexJ/EVRLLib`, June 2022.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Francis Longstaff and Eduardo Schwartz. Valuing american options by simulation: A simple least-squares approach. *Review of Financial Studies*, 14:113–47, 02 2001.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] Erich Novak and Henryk Wo'zniakowski. *Tractability of multivariate problems. Vol. 1: Linear information.* 01 2008.

[50] 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.

[51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[52] John Rust. Using randomization to break the curse of dimensionality. *Econometrica*, 65(3):487–516, 1997.

[53] 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.

[54] 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.

[55] 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.

[56] R. Srikant and Lei Ying. Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning. In *Conference on Learning Theory*, 2019.

[57] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[58] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] Vladislav Tadic. Asymptotic analysis of temporal-difference learning algorithms with constant step-sizes. *Machine Learning*, 63:107–133, 05 2006.

[64] Nizar Touzi. Optimal stochastic control, stochastic target problems, and backward sde. *Fields Institute Monographs*, 29, 01 2013.

[65] Lloyd Trefethen. Multivariate polynomial approximation in the hypercube. *Proceedings of the American Mathematical Society*, 145, 08 2016.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] L. Weaver and N. Tao. The optimal reward baseline for gradient·based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2001.

[71] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[72] 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.

[73] 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.

[74] 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.

[75] Daniel Zanger. Quantitative error estimates for a least-squares monte carlo algorithm for american option pricing. *Finance and Stochastics*, 17, 07 2013.