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

Alexey Naumov

HDI Lab

HSE University

NATIONAL RESEARCH UNIVERSITY

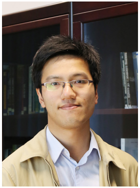
September 30, 2020, Optimization Seminar

Joint Work with

Maxim Kaledin
(HSE)

Eric Moulines
(Ecole Polytechnique)

Vladislav Tadic
(Bristol)



Hoi-To Wai
(CUHK)

Paper available at
https://www.colt2020.org/virtual/papers/paper_124.html

Summary

Motivation

- Many reinforcement learning algorithms are *stochastic approximation (SA)* schemes to a fixed point equation, e.g., finding $\theta^\star$ such that

$$f(\theta^\star) = 0 \quad \text{where } f(\theta) \text{ is the TD error.}$$

- Only stochastic samples of $f(\theta)$ are revealed, e.g., $F(\theta; X_k)$,

$$\theta_{k+1} = \theta_k + \gamma_k F(\theta_k; X_k).$$

- Random 'seeds' X_k are Markovian such that for a given θ,

$$\mathbb{E}[F(\theta; X_k)] \neq f(\theta) \quad \text{but} \quad \lim_{k\to\infty} \mathbb{E}[F(\theta; X_k)] = f(\theta).$$

- Understanding the performance of SA is the focus of many old and new works, e.g., Jaakkola et al. [1994], Kushner and Yin [2003], Benveniste et al. [2012], Bhandari et al. [2018], Srikant and Ying [2019]

The above only study **one-timescale SA** *for a fixed point equation.*

Fixed Point to System of Two Equations

Goal: find the unique fixed point $(\theta^\star, w^\star)$ to the system of 2 equations:

$$f_1(\theta, w) = 0, \quad f_2(\theta, w) = 0. \tag{FP}$$

- For min-max problems, (e.g., GTD2 learning)

$$\min_{\theta} \max_{w} \; L(\theta, w),$$

$$\implies \quad f_1(\theta, w) = -\nabla_\theta L(\theta, w), \; f_2(\theta, w) = \nabla_w L(\theta, w).$$

- For bilevel problems [Ghadimi and Wang, 2018], (e.g., Actor-critic)

$$\min_{\theta, w} L_1(\theta, w) \text{ s.t. } w \in \arg\min_{w} L_2(\theta, w),$$

$$\implies \quad \begin{aligned} & f_1(\theta, w) = -\nabla_\theta L_1(\theta, w) + \nabla^2_{\theta, w} L_2(\theta, w) \nabla^2_{w, w} L_2(\theta, w)^{-1} \nabla_w L_1(\theta, w), \\ & f_2(\theta, w) = -\nabla_w L_2(\theta, w) \end{aligned}$$

Finding Fixed Points with Stochastic Samples

- We only have **stochastic samples** and the system is **coupled**.
- Let X_{k+1} denotes the random 'seed' at iteration k, and $F_1(\cdot; X_{k+1})$, $F_2(\cdot; X_{k+1})$ denote the stochastic samples of f_1, f_2, respectively.
- If θ **is fixed** and under suitable conditions, the recursion

$$w_{k+1} = w_k + \gamma_k F_2(\theta, w_k; X_{k+1}) \stackrel{k\to\infty}{\longrightarrow} w^\star(\theta) \text{ s.t. } f_2(\theta, w^\star(\theta)) = 0.$$

- Furthermore, the recursion

$$\theta_{k+1} = \theta_k + \beta_k F_1(\theta_k, w^\star(\theta_k); X_{k+1}) \stackrel{k\to\infty}{\longrightarrow} \theta^\star \text{ s.t. } f_1(\theta^\star, w^\star(\theta^\star)) = 0.$$

- If one could run the two recursions, then (FP) is solved, but the w_k recursion requires θ to be fixed; and θ_k recursion requires $w^\star(\theta_k)$.

thus suggesting a double-loop algorithm...

Two Timescale Stochastic Approximation (TTSA)

- Consider the **single-loop, two timescale** algorithm [Borkar, 1997]:

$$w_{k+1} = w_k + \gamma_k F_2(\theta_k, w_k; X_{k+1})$$
$$\theta_{k+1} = \theta_k + \beta_k F_1(\theta_k, w_k; X_{k+1})$$

- We require that

$$\lim_{k\to\infty} \frac{\beta_k}{\gamma_k} = 0$$

- **Intuition**: when updating w_k, as $\beta_k \ll \gamma_k$, then θ_k is *almost static*; when updating θ_k, the used w_k have *almost converged* to $w^\star(\theta_k)$.
- θ-update is at slow timescale; while w-update is at fast timescale.

This Talk

We focus on **linear TTSA** where F_1, F_2 are linear functions of θ, w.
Examples: policy evaluation with gradient TD learning.

Motivation: Policy Evaluation Problem

- $\mathcal{S}, \mathcal{A}$ – discrete state, action spaces, $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ – stationary *policy*.
- At step k, the agent performs action $a_k \sim \pi(\cdot|s_k)$ and transits to state $s_{k+1} \sim p(\cdot \mid s_k, a_k)$ to obtain a reward $r_{k+1} \sim r(\cdot \mid s_{k+1}, a_k)$.
- Let $\alpha \in [0, 1)$, the value function for discounted reward is

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{k=0}^{\infty} \alpha^k r_k \right], \quad s \in \mathcal{S}.$$

- The Markov chain $\{s_k\}_{k=1}^{\infty}$ induced by π is assumed to be ergodic with the stationary distribution μ.

Policy evaluation w/ linear function approximation

To approximate the value function as $\hat{V}^\pi(s) = \theta^\top \phi(s)$, where $\phi(s)$ is a feature map and θ is a parameter vector.

Motivation: Gradient TD Principle

- Let $\delta_k(\theta_k) = r_k + \alpha\theta_k^\top \phi_k' - \theta_k^\top \phi_k$ with $\phi_k = \phi(s_k)$, $\phi_k' = \phi(s_{k+1})$
- The **linear TD solution** $\theta^\star$ shall satisfy

$$0 = \mathbb{E}^\pi[\phi \cdot \delta(\theta^\star)] = \lim_{k\to\infty} \mathbb{E}^\pi\left[\phi(s_k) \cdot \delta_k(\theta^\star)\right] = -A\theta^\star + b$$

$$\text{where} \quad \begin{aligned} A &= \lim_{k\to\infty} \mathbb{E}^\pi[\phi(s_k)\{\phi(s_k) - \alpha\phi(s_{k+1})\}^\top] = \mathbb{E}^\pi[\phi\{\phi - \gamma\phi'\}^\top] \\ b &= \lim_{n\to\infty} \mathbb{E}^\pi[r_k\phi(s_k)] = \mathbb{E}^\pi[r\phi]. \end{aligned}$$

- In GTD0 [Sutton et al., 2009a], we consider the objective function given as the norm of expected TD update (NEU):

$$J(\theta) = (1/2)\left\|\mathbb{E}^\pi[\phi \cdot \delta(\theta)]\right\|^2 = (1/2)\|b - A\theta\|^2$$

†alternative formulations: MSPBE in [Sutton et al., 2009b] leads to GTD2.

Motivation: GTD0 algorithm

- The **gradient** of the objective function is

$$\nabla J(\theta) = A^\top (A\theta - b) = -\mathbb{E}^\pi[\{\phi - \alpha\phi'\}\phi^\top]\,\mathbb{E}^\pi[\phi \cdot \delta(\theta)]$$

- A naive gradient estimator as $\{(\phi_k - \alpha\phi_{k+1})\phi_k^\top\}\{\phi_k \cdot \delta_k(\theta)\}$ does not work as it gives a biased estimate of $\nabla J(\theta)$.

- Define a slack variable w, and write the TD stationary condition as

$$0 = f_1(\theta, w) = \mathbb{E}^\pi[(\phi - \alpha\phi')\phi^\top]\,w, \quad 0 = f_2(\theta, w) = \mathbb{E}^\pi[\phi \cdot \delta(\theta)] - w$$

- We can apply **TTSA**:

$$\begin{aligned} \theta_{k+1} &= \theta_k + \beta_k \{\phi_k - \alpha\phi_{k+1}\}\phi_k^\top w_k \\ w_{k+1} &= w_k + \gamma_k \{(r_k + \alpha\theta_k^\top \phi_{k+1} - \theta_k^\top \phi_k)\phi_k - w_k\}\,. \end{aligned}$$

  Again, we set $\beta_k/\gamma_k \to 0$ and w_k is 'almost' stationary w.r.t. θ_k. Furthermore, it is a **linear TTSA** as the updates are linear.

Agenda

Linear TTSA in General Form

- We analyze the **linear TTSA** scheme in general form:

$$\theta_{k+1} = \theta_k + \beta_k \{\widetilde{b}_1(X_{k+1}) - \widetilde{A}_{11}(X_{k+1})\theta_k - \widetilde{A}_{12}(X_{k+1})w_k\},$$
$$w_{k+1} = w_k + \gamma_k \{\widetilde{b}_2(X_{k+1}) - \widetilde{A}_{21}(X_{k+1})\theta_k - \widetilde{A}_{22}(X_{k+1})w_k\}.$$

where $\widetilde{b}_i(x)$, $\widetilde{A}_{ij}(x)$ are vector/matrix functions.

Our Results

- Finite-time L_2 error bounds for each of θ_k, w_k.
- Two settings for the stochastic process $(X_k)_{k\geq 0}$: when **(a)** it is a sequence of i.i.d. samples, or **(b)** it forms an ergodic Markov chain.

Linear TTSA in General Form

- (In detail) It is possible to rewrite the linear TTSA as

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

where $b_i := \lim_{k\to\infty} \mathbb{E}[\widetilde{b}_i(X_k)]$, $A_{ij} := \lim_{k\to\infty} \mathbb{E}[\widetilde{A}_{ij}(X_k)]$, and

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

- Let $\Delta := A_{11} - A_{12} A_{22}^{-1} A_{21}$, the fixed point of TTSA is:

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

- $(X_k)_{k\geq 0}$ = i.i.d. samples $\Rightarrow \mathbb{E}^{\mathcal{F}_k}[V_{k+1}], \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$,
- $(X_k)_{k\geq 0}$ = ergodic Markov chain $\Rightarrow \mathbb{E}^{\mathcal{F}_k}[V_{k+1}], \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] \neq 0$.

Prior Works

- *Almost-sure convergence, central limit theorem and alike*
  - Borkar [1997] assumes bounded iterates.
  - Mokkadem et al. [2006] consider a restricted form of nonlinear TTSA.
  - Konda and Tsitsiklis [2004] proved steady-state rates with homoscedastic (finite variance) Martingale noise:

$$\mathbb{E}[\|\theta_k - \theta^\star\|^2] = \mathcal{O}(\beta_k), \quad \mathbb{E}[\|w_k - w^\star\|^2] = \mathcal{O}(\gamma_k) \tag{1}$$

- *Finite-time Bounds*
  - Martingale noise: Dalal et al. [2018], particularly Dalal et al. [2019] obtained high probability bounds with a **projection** step, with the same steady-state rate as (1).
  - Markovian noise: Xu et al. [2019], Doan [2019] obtained L^2 bounds of $\mathbb{E}[\|\theta_k - \theta^\star\|^2] = \mathcal{O}(\gamma_k), \mathbb{E}[\|w_k - w^\star\|^2] = \mathcal{O}(\gamma_k)$ with a projection step; Gupta et al. [2019] analyzed L^2 bounds with constant step size.

- And many others...

Our Contributions

- A separation of scales in convergence rates is found in i.i.d. noise case – not found in prior works with Markovian noise.
- We close the gap in this paper (+ relax bounded iterate assumption):

L_2 error	*I.i.d. noise* [Dalal et al., 2019]	*Markovian noise* [Xu et al., 2019]	**This Work**
$\mathbb{E}[\|w_k - w^\star\|^2]$	$\mathcal{O}(\gamma_k)$	$\mathcal{O}(\gamma_k)$	$\mathcal{O}(\gamma_k)$
$\mathbb{E}[\|\theta_k - \theta^\star\|^2]$	$\mathcal{O}(\beta_k)$	$\mathcal{O}(\gamma_k)$	$\mathcal{O}(\beta_k)$

† only 'steady-state' error is shown, the exact rates will be provided later.

Highlights

- Relaxed finite-time analysis without boundedness assumption.
- Improved finite-time bounds with Markovian noise.
- Asymptotic expansion with Martingale noise.

Agenda

General Assumptions

Assumption 1

Matrices $-A_{22}$ and $-\Delta$ are *Hurwitz*.

Assumption 2, similar to [Konda and Tsitsiklis, 2004]

$(\gamma_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$ are nonincreasing positive numbers satisfying

1. There exist constants κ such that $\beta_k/\gamma_k \leq \kappa$;
2. There exist constants $\delta_1, \delta_2, \delta_3$ such that

$$\frac{\gamma_k}{\gamma_{k+1}} \leq 1 + \delta_1 \gamma_{k+1}, \quad \frac{\beta_k}{\beta_{k+1}} \leq 1 + \delta_2 \beta_{k+1}, \quad \frac{\gamma_k}{\gamma_{k+1}} \leq 1 + \delta_3 \beta_{k+1}.$$

Example

- $\beta_k = c^\beta/(k + k_0^\beta)$, $\gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3}$, **a popular choice** in lit.
- Also hold for constant, piecewise diminishing step sizes.
  (The condition will become slightly more restrictive for Markovian noise.)

Martingale Noise — Assumptions

- Let us first look at the case with Martingale noise.

Assumption 3

Noises are conditionally zero-mean, $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] = \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$.

Example

X_k are drawn i.i.d. such that $b_i = \mathbb{E}[\widetilde{b}_i(X_0)], A_{ij} = \mathbb{E}[\widetilde{A}_{ij}(X_0)]$.

Assumption 4

There exist constants m_W, m_V such that

$$\|\mathbb{E}[V_{k+1}V_{k+1}^\top]\| \leq m_V(1 + \|\mathbb{E}[\theta_k\theta_k^\top]\| + \|\mathbb{E}[w_k w_k^\top]\|),$$
$$\|\mathbb{E}[W_{k+1}W_{k+1}^\top]\| \leq m_W(1 + \|\mathbb{E}[\theta_k\theta_k^\top]\| + \|\mathbb{E}[w_k w_k^\top]\|)\,.$$

- Compared to Konda and Tsitsiklis [2004], we only need *non-homoscedastic* noise which is suitable for GTD learning.

Error Bounds, Martingale Case

Theorem

Under Assumptions 1-4, there exists $a \in (0,1)$ and for any $k \geq 0$,

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \lesssim \prod_{\ell=0}^{k-1} \Big(1 - a\beta_\ell\Big) \mathrm{V}_0 + \beta_k$$

$$\mathbb{E}[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2] \lesssim \prod_{\ell=0}^{k-1} \Big(1 - a\beta_\ell\Big) \mathrm{V}_0 + \gamma_k$$

where V_0 depends on the initialization, the inequality is up to constants not depending on k (exact expressions can be found in the paper)

- Note $w^\star(\theta_k) = A_{22}^{-1}(b_2 - A_{21}\theta_k)$ and thus **tracking error** is $\mathcal{O}(\gamma_k)$ in the steady-state; meanwhile **convergence of θ_k** is $\mathcal{O}(\beta_k)$.
- Shows a *separation of scale* similar to Dalal et al. [2019] — we analyzed the plain linear TTSA without projection.

Sketch of the proof

Recall

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

Highlight

- The updates are coupled together: θ_{k+1} depends on θ_k, w_k.
- Our idea: *decouple* the updates using a "Gaussian elimination" trick from Konda and Tsitsiklis [2004].

Sketch of the proof

Recall

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

Change-of-variables (by Konda and Tsitsiklis [2004]):

$$\tilde{\theta}_k := \theta_k - \theta^\star, \quad \tilde{w}_k = w_k - w^\star + C_{k-1}\tilde{\theta}_k, \quad C_k \approx A_{22}^{-1}A_{21}$$

leads to the 'decoupled' updates

$$\begin{aligned} \tilde{\theta}_{k+1} &= (\mathrm{I} - \beta_k B_{11}^k)\tilde{\theta}_k - \beta_k A_{12}\tilde{w}_k - \beta_k V_{k+1}, \quad B_{11}^k \approx \Delta, \\ \tilde{w}_{k+1} &= (\mathrm{I} - \gamma_k B_{22}^k)\tilde{w}_k - \beta_k C_k V_{k+1} - \gamma_k W_{k+1}, \quad B_{22}^k \approx A_{22} \end{aligned} \tag{2}$$

Denote

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

We bound the error terms above one by one.

Sketch of the proof

For some $a_1, a_2 > 0$, it holds

$$\mathrm{M}_{k+1}^{\tilde{w}} \lesssim \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \gamma_{k+1} + \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{M}_j^{\tilde{\theta}},$$

Highlight

- By Assumption 3

$$\mathbb{E}^{\mathcal{F}_k}[\tilde{w}_{k+1}\tilde{w}_{k+1}^\top] = (\mathrm{I} - \gamma_k B_{22}^k)\tilde{w}_k\tilde{w}_k^\top(\mathrm{I} - \gamma_k B_{22}^k)^\top$$

$$+ \mathbb{E}^{\mathcal{F}_k}[(\beta_k C_k V_{k+1} + \gamma_k W_{k+1})(\beta_k C_k V_{k+1} + \gamma_k W_{k+1})^\top]$$

- The last term can be bounded using Assumption 4, ...

Sketch of the proof

For some $a_1, a_2 > 0$, it holds

$$\mathsf{M}^{\tilde{w}}_{k+1} \lesssim \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \gamma_{k+1} + \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)\mathsf{M}^{\tilde{\theta}}_{j},$$

Similarly, for the cross-covariance:

$$\mathsf{M}^{\tilde{\theta},\tilde{w}}_{k+1} \lesssim \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \beta_{k+1} + \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)\mathsf{M}^{\tilde{\theta}}_{j},$$

Highlight

- One maybe tempted to use (Cauchy-schwarz ineq.):

$$\mathsf{M}^{\tilde{\theta},\tilde{w}}_{k+1} \leq \mathrm{C} \cdot \{\mathsf{M}^{\tilde{\theta}}_{k+1} + \mathsf{M}^{\tilde{w}}_{k+1}\}$$

to bound the cross-covariance, yet this result in a sub-optimal rate as $\mathsf{M}^{\tilde{\theta},\tilde{w}}_{k} = \mathcal{O}(\gamma_k)$.

Sketch of the proof

For some $a_1, a_2 > 0$, it holds

$$\mathsf{M}^{\tilde{w}}_{k+1} \lesssim \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \gamma_{k+1} + \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)\mathsf{M}^{\tilde{\theta}}_j,$$

Similarly, for the cross-covariance:

$$\mathsf{M}^{\tilde{\theta},\tilde{w}}_{k+1} \lesssim \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \beta_{k+1} + \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)\mathsf{M}^{\tilde{\theta}}_j,$$

$$\mathsf{M}^{\tilde{\theta}}_{k+1} \lesssim \prod_{\ell=0}^{k} \big(1 - a_2\beta_\ell\big)\mathrm{V}_0 + \beta_{k+1} + \sum_{j=0}^{k} \gamma_j\beta_j \prod_{\ell=j+1}^{k} \big(1 - a_2\beta_\ell\big)\mathsf{M}^{\tilde{\theta}}_j, \quad (3)$$

Eq. (3) is a recursive inequality. There exists a sequence $(\mathsf{U}_k)_{k\geq 0}$ satisfying $\mathsf{M}^{\tilde{\theta}}_k \leq \mathsf{U}_k$ and $\mathsf{U}_{k+1} \lesssim (1 - a_3\beta_k)\,\mathsf{U}_k + \beta_k^2$.

Markovian Noise — Assumptions

- Let $(X_k)_{k \geq 0}$ forms a Markov chain with kernel $\mathrm{P} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$.

Assumption 5

Markov kernel P is irreducible, aperiodic, with a unique invariant dist. $\mu : \mathsf{X} \to \mathbb{R}_+$. We have $b_i = \int_{\mathsf{X}} \widetilde{b}_i(x) \, \mu(\mathrm{d}x)$, $A_{ij} = \int_{\mathsf{X}} \widetilde{A}_{ij}(x) \, \mu(\mathrm{d}x)$.

Assumption 6 (Poisson equation)

For any $i, j = 1, 2$ there exist $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ which satisfy for any $x \in \mathsf{X}$.

$$\widetilde{b}_i(x) - b_i = \widehat{b}_i(x) - \mathrm{P} \, \widehat{b}_i(x), \quad \widetilde{A}_{ij}(x) - A_{ij} = \widehat{A}_{ij}(x) - \mathrm{P} \, \widehat{A}_{ij}(x). \tag{4}$$

Example

A5 **implies** A6 when $\widetilde{A}, \widetilde{b}$ are bounded functions with the solution:

$$\widehat{A}_{ij}(x) = \textstyle\sum_{k=0}^{\infty} \{\mathrm{P}^k \, \widetilde{A}_{ij}\}(x), \quad \widehat{b}_i(x) = \textstyle\sum_{k=0}^{\infty} \{\mathrm{P}^k \, \widetilde{b}_i\}(x)$$

Markovian Noise — Assumptions (cont'd)

Assumption 7

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

Example

- Previous step size $\beta_k = c^\beta/(k + k_0^\beta)$, $\gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3}$, as well as constant, piecewise diminishing step sizes, still work.
- $\beta_k = c^\beta/(k + k_0^\beta)$, $\gamma_k = c^\gamma/(k + k_0^\gamma)^\alpha$ for $\alpha < 1/2$ does not work.

  (that said, we believe this is an artifact in our proof which should be fixable.)

Assumption 8

The vector/matrix valued functions $b_i(x), A_{ij}(x)$ are uniformly bounded.

- Note we do not assume θ_k, w_k to be bounded a-priori.

Error Bounds, Markovian Case

Theorem

Under Assumptions 1-2, 5-8, there exists $a \in (0,1)$ and for any $k \geq 0$,

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \lesssim \prod_{\ell=0}^{k-1} \Big(1 - a\beta_\ell\Big)(1 + \mathrm{V}_0) + \beta_k$$

$$\mathbb{E}[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2] \lesssim \prod_{\ell=0}^{k-1} \Big(1 - a\beta_\ell\Big)(1 + \mathrm{V}_0) + \gamma_k$$

where V_0 depends on the initialization, and the inequality is up to constants not depending on k (exact expressions in the paper).

- Similar *separation of scale* to the Martingale case.
- The constants depend on mixing time of the Markov chain, upper bounds $\widetilde{A}_{ij}, \widetilde{b}_i$, etc..

Sketch of the proof

- Observe that

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

- The Poisson equations (A6) allow us to write

$$\widetilde{b}_i(X_{k+1}) - b_i = \underbrace{\widehat{b}_i(X_{k+1}) - \mathrm{P}\,\widehat{b}_i(X_k)}_{\text{a martingale}} + \underbrace{\widehat{b}_i(X_k) - \mathrm{P}\,\widehat{b}_i(X_{k+1})}_{\text{finite difference}}$$

- Split V_k, W_k to martingale $V_k^{(0)}, W_k^{(0)}$ & finite-difference $V_k^{(1)}, W_k^{(1)}$.
- We also split the error terms for TTSA:

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

Sketch of the proof (cont'd)

- Observe that

$$\tilde{\theta}_{k+1} = \tilde{\theta}^{(0)}_{k+1} + \tilde{\theta}^{(1)}_{k+1}, \quad \tilde{w}_{k+1} = \tilde{w}^{(0)}_{k+1} + \tilde{w}^{(1)}_{k+1}.$$

- The error terms can be analyzed separately. E.g., for some $a_1 > 0$:

$$\begin{aligned} \mathsf{M}^{\tilde{w}^{(0)}}_{k+1} &\leq \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \gamma_{k+1} \\ &\quad + \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)(\mathsf{M}^{\tilde{w}}_j + \mathsf{M}^{\tilde{\theta}}_j), \\ \mathsf{M}^{\tilde{w}^{(1)}}_{k+1} &\lesssim \prod_{\ell=0}^{k} \big(1 - a_1\gamma_\ell\big)\mathrm{V}_0 + \gamma_{k+1}^2(\mathsf{M}^{\tilde{\theta}}_{k+1} + \mathsf{M}^{\tilde{w}}_{k+1}) + \gamma_{k+1}^2 \\ &\quad + \gamma_{k+1} \sum_{j=0}^{k} \gamma_j^2 \prod_{\ell=j+1}^{k} \big(1 - a_1\gamma_\ell\big)(\mathsf{M}^{\tilde{\theta}}_j + \mathsf{M}^{\tilde{w}}_j), \end{aligned}$$

$\Longrightarrow$ *Martingale-driven errors* $\gg$ *finite-difference-driven errors.*

- Finally, we repeat the proof of Theorem 1 to bound $\mathsf{M}^{\tilde{\theta}^{(0)}}_k$, and subsequently it can be shown that $\mathsf{M}^{\tilde{\theta}^{(1)}}_k$ is small.

Asymptotic Expansion of Error for Slow Timescale

Theorem

Under some mild assumptions and Assumptions 1-4 for sufficiently small stepsizes and for all $k \in \mathbb{N}$ the following expansion holds

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = I_k + J_k,$$
$$I_k := \sum_{j=0}^{k} \beta_j^2 \operatorname{Tr}\left(\prod_{\ell=j+1}^{k} (\mathrm{I} - \beta_\ell \Delta) \Sigma \left\{ \prod_{\ell=j+1}^{k} (\mathrm{I} - \beta_\ell \Delta) \right\}^\top \right),$$

and Σ depends on the Martingale noise covariance, A_{ij}; importantly,

$$\beta_k \cdot C_1 \operatorname{Tr}(\Sigma) \leq I_k \leq \beta_k \cdot C_2 \operatorname{Tr}(\Sigma),$$
$$|J_k| \lesssim \prod_{\ell=0}^{k-1} (1 - a\beta_\ell) \mathrm{V}_0 + \beta_k \left(\gamma_k + \frac{\beta_k}{\gamma_k}\right).$$

- Focus on the martingale noise setting, we have that I_k dominates J_k as $k \to \infty$.
- Importantly, $I_k = \Theta(\beta_k)$ which matches the upper bound and it can be computed in **closed form**.

Agenda

1. Two-Timscales Stochastic Approximation (TTSA)
2. Linear TTSA
3. Convergence Analysis of Linear TTSA
4. Numerical Experiments and Summary

Experiments: Toy Example with Martingale Noise

Toy scheme with fixed A_{ij}, b_i and i.i.d. noise V_k, W_k. Key parameters:

1. Dimensions $d_\theta = d_\omega = 10$;
2. Step sizes $\beta_k = c^\beta/(k_0^\beta + k), \gamma_k = c^\gamma/(k_0^\beta + k)^\sigma$ with $\sigma \in \{0.5, 0.67, 0.75\}$ and $k_0^\beta = 10^4, k_0^\gamma = 10^7, c^\beta = 140, c^\gamma = 300$.

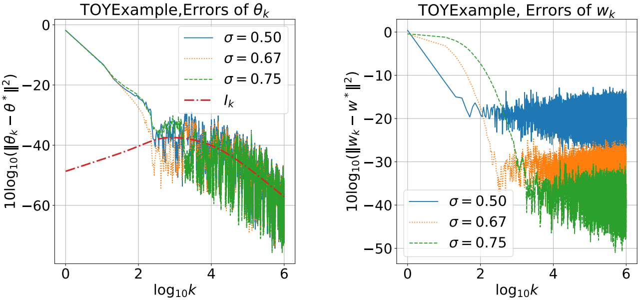


Figure: Deviations from stationary point $(\theta^\star, \omega^\star)$ *normalized* by step sizes β_k, γ_k. I_k is computed using the exact formula in the Theorem.

Experiments: Garnet Problem with Markovian Samples

Key parameters:

1. Garnet problem with $n_S = 50, n_A = 10, b = 2$;
2. Stepsizes $\beta_k = c^\beta/(k + k_0^\beta),\ \gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3}$

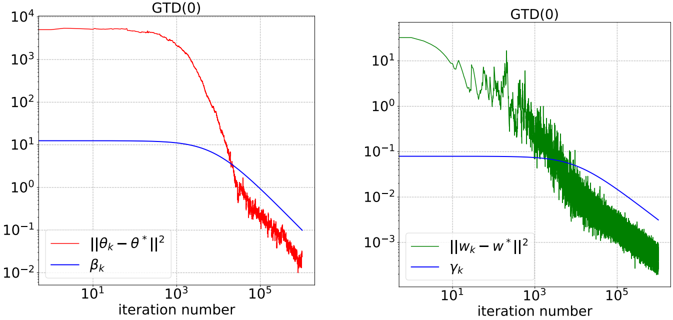


Figure: Deviations from stationary point $(\theta^\star, \omega^\star)$

Summary

- We closed a gap in the finite error bounds of TTSA – demonstrating the **separation of scales** in convergence rates with Markov noise

$$\mathbb{E}[\|\theta_k - \theta^\star\|^2] = \mathcal{O}(\beta_k), \quad \mathbb{E}[\|w_k - w^\star\|^2] = \mathcal{O}(\gamma_k)$$

- Relaxed some 'artificial' constructions made in prior works, e.g., (sparse) projection TTSA was assumed to ensure boundedness [Dalal et al., 2019].
- The martingale bound is shown to be optimal using an asymptotic expansion argument.

Future Works

- Getting rid of the Poisson equation allows us to perform a *fine-grained expansion* of linear SA similar to Aguech et al. [2000].

  — For any $p \geq 1$, we showed in the 1-TS case that

  $$\left(\mathbb{E}[\|\theta_k - \theta^\star\|^p] \right)^{\frac{1}{p}} = J_k^{(0)} + J_k^{(1)} + \cdots + J_k^{(L)} + H_k^{(L)}$$

  with a provable *separation of scale* like $J_k^{(0)} = \mathcal{O}(\sqrt{\beta})$, $J_k^{(1)} = \mathcal{O}(\beta)$, ..., $J_k^{(L)} = \mathcal{O}(\beta^{\frac{L+1}{2}})$, $H_k^{(L)} = \mathcal{O}(\beta^{\frac{L+2}{2}})$.

- A nonlinear version of TTSA allows us to tackle (possibly non-convex) *bi-level optimization problems*, see Hong et al. [2020].

  — For i.i.d. updates, we showed that a two timescale natural actor-critic algorithm converges at $\mathcal{O}(K^{-1/4})$ to optimal policy.

- $\geq$ 3-Timescale SA? ...

Thank you!

References I

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

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

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.

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

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.

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.

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

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Harsh Gupta, R Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *NeurIPS*, pages 4706–4715, 2019.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

References II

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.

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.

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

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.

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

Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *NeurIPS*, pages 1609–1616, 2009a.

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 *ICML*, pages 993–1000, 2009b.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *NeurIPS*, pages 10633–10643, 2019.