Finite-time High-probability Bounds for Polyak-Ruppert Averaged Iterates of Linear Stochastic Approximation

Eric Moulines

Ecole polytechnique

June 17, 2024

Introduction

Stochastic Approximation

- Consider the problem of finding $\theta^\star \in \mathbb{R}^d$ such that

$$f(\theta^\star) = 0\,.$$

- Only "noisy" samples of $f(\theta)$ are revealed, e.g., $F(\theta; Z_n)$, such that

$$\mathbb{E}[F(\theta; Z_n)] = f(\theta) \quad \text{or, at least,} \quad \lim_{n\to+\infty} \mathbb{E}[F(\theta; Z_n)] = f(\theta).$$

- Such algorithms are called *stochastic approximation (SA)* schemes to a fixed point equation:

$$\theta_{n+1} = \theta_n + \alpha_n F(\theta_n; Z_n).$$

  Robbins and Monro [1951]

- Compare with the standard 'Euler scheme' for numerically approximating a trajectory of the o.d.e. $\dot{\theta}(t) = f(\theta(t))$

$$\theta_{t+1} = \theta_t + \alpha f(\theta_t)$$

- The simplest instance of the problem corresponds to the Linear Stochastic Approximation (LSA)

Linear Stochastic Approximation

- Given $\bar{\mathbf{A}} \in \mathbb{R}^{d \times d}$ and $\bar{\mathbf{b}} \in \mathbb{R}^d$, we aim at finding $\theta^\star \in \mathbb{R}^d$, which is a solution of
$$\bar{\mathbf{A}}\theta^\star = \bar{\mathbf{b}}\,.$$
- Our analysis is based on noisy observations $\{(\mathbf{A}(Z_n), \mathbf{b}(Z_n))\}_{n \in \mathbb{N}}$. Here $\mathbf{A} : \mathsf{Z} \to \mathbb{R}^{d \times d}$, $\mathbf{b} : \mathsf{Z} \to \mathbb{R}^d$ are measurable mappings.

LSA algorithm

For a sequence of step sizes $\{\alpha_k\}$, burn-in period $n_0 \in \mathbb{N}$, and initialization θ_0, consider the sequences of estimates $\{\theta_n\}_{n \in \mathbb{N}}, \{\bar{\theta}_{n_0,n}\}_{n \geq n_0+1}$ given by

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

Linear Stochastic Approximation

I.I.D. Noise

Sequence $\{Z_k\}_{k \in \mathbb{N}}$ is an i.i.d. sequence taking values in a state space $(\mathsf{Z}, \mathcal{Z})$ with distribution π satisfying $\mathbb{E}[\mathbf{A}(Z_1)] = \bar{\mathbf{A}}$ and $\mathbb{E}[\mathbf{b}(Z_1)] = \bar{\mathbf{b}}$;

Markovian noise

Sequence $\{Z_k\}_{k \in \mathbb{N}}$ is a Z-valued ergodic Markov chain with unique invariant distribution π, such that

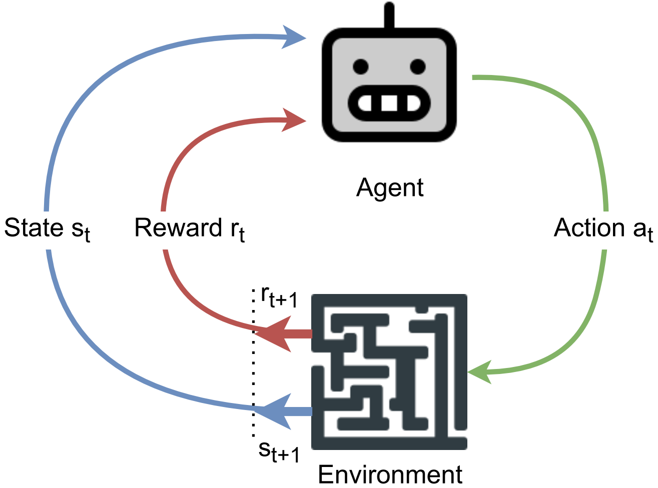
$$\lim_{n \to +\infty} \mathbb{E}[\mathbf{A}(Z_n)] = \bar{\mathbf{A}}$$

and

$$\lim_{n \to +\infty} \mathbb{E}[\mathbf{b}(Z_n)] = \bar{\mathbf{b}}$$

We write $\mathbf{A}_k$ instead of $\mathbf{A}(Z_k)$, and $\mathbf{b}_k$ instead of $\mathbf{b}(Z_k)$, respectively.

RL: general paradigm



Applications: TD learning

- Consider a problem of estimating the policy ν in a discounted MDP given by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$;
- $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces, assume that they are complete metric spaces equipped with Borel σ-algebras $\mathcal{B}(\mathcal{S})$ and $\mathcal{B}(\mathcal{A})$, respectively;
- $\gamma \in (0, 1)$ is a discount factor;
- $\mathcal{P}$ stands for the transition kernel $\mathcal{P}(\cdot|s, a)$;
- reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ - deterministic;
- policy $\nu(\cdot|s)$ - distribution over the action space $\mathcal{A}$;

Applications: TD learning

- We aim to estimate the agent's value function

$$V^{\nu}(s) = \mathbb{E}\left[\sum\nolimits_{k=0}^{\infty} \gamma^k r(s_k, a_k) | s_0 = s\right] \,,$$

where $a_k \sim \nu(\cdot|s_k)$, and $s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$;

- 1-step transition kernel:

$$\mathcal{P}_{\nu}(B|s) = \int_{\mathcal{A}} \mathcal{P}(B|s, a)\nu(\mathrm{d}a|s) \,, \quad B \in \mathcal{B}(\mathcal{S}) \,; \tag{2}$$

- *Linear functional approximation* of the true value function $V^{\nu}(s)$:

$$V^{\nu}_{\theta}(s) = \varphi^{\top}(s)\theta \,,$$

where $s \in \mathcal{S}$, $\theta \in \mathbb{R}^d$, $\varphi : \mathcal{S} \to \mathbb{R}^d$, d - feature dimension

TD learning as LSA problem

- The problem of estimating $V^{\nu}(s)$ reduces to the problem of estimating $\theta \in \mathbb{R}^d$ in $V^{\nu}_{\theta}(s)$;
- Set the k-th step randomness as $Z_k = (s_k, s'_k)$;
- The corresponding LSA writes as:

$$\theta_k = \theta_{k-1} - \alpha_k(\mathbf{A}_k \theta_{k-1} - \mathbf{b}_k)\,, \tag{3}$$

where the system matrix and r.h.s. are given by

$$\mathbf{A}_k = \phi(s_k)\{\phi(s_k) - \gamma\phi(s'_k)\}^{\top},\, \mathbf{b}_k = \phi(s_k) r(s_k, a_k)\,. \tag{4}$$

- Deterministic system writes as $\bar{\mathbf{A}}\theta^{\star} = \bar{\mathbf{b}}$, where

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

Finite-time high-probability bounds for the Polyak-Ruppert averaged LSA iterates

Linear Stochastic Approximation

- Let $\{Z_k\}_{k \in \mathbb{N}}$ be an i.i.d.sequence and consider the recurrence

$$\theta_k = \theta_{k-1} - \alpha_k \{\mathbf{A}(Z_k)\theta_{k-1} - \mathbf{b}(Z_k)\} \tag{5}$$

- Set

$$\tilde{\mathbf{A}}(z) = \mathbf{A}(z) - \bar{\mathbf{A}}\,, \quad \tilde{\mathbf{b}}(z) = \mathbf{b}(z) - \bar{\mathbf{b}}\,,$$

  and introduce

$$\varepsilon(z) = \mathbf{A}(z)\theta^\star - \mathbf{b}(z)\,, \quad \Sigma_\varepsilon = \mathbb{E}[\varepsilon(Z)\varepsilon(Z)^\top]\,.$$

Assumption A1

(i) $C_A = \sup_{z \in Z} \|\mathbf{A}(z)\| \vee \sup_{z \in Z} \|\tilde{\mathbf{A}}(z)\| < \infty$ and the matrix $-\bar{\mathbf{A}}$ is Hurwitz

(ii) $\int_Z \mathbf{A}(z)\mathrm{d}\pi(z) = \bar{\mathbf{A}}$ and $\int_Z \mathbf{b}(z)\mathrm{d}\pi(z) = \bar{\mathbf{b}}$. Moreover, $\|\varepsilon\|_\infty = \sup_{z \in Z} \|\varepsilon(z)\| < +\infty$.

Why averaging: CLT view

Step size assumptions

Suppose that the sequence α_k satisfies one of the following assumptions:

(i) $\sum_{k=1}^{\infty} \alpha_k = \infty$, $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, $\log(\alpha_{k-1}/\alpha_k) = o(\alpha_k)$;

(ii) $\sum_{k=1}^{\infty} \alpha_k = \infty$, $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, $\log(\alpha_{k-1}/\alpha_k) \sim \alpha_k/\alpha_*$ for

$$\alpha_* \geq 1/(2L)\,, \text{ where } L = \min \mathrm{Re}(\lambda_i(\bar{\mathbf{A}}))\,.$$

Examples: $\alpha_k = c_0/k^\gamma$, $\gamma \in (0.5; 1)$ satisfies (i); $\alpha_k = \alpha_*/k$ satisfies (ii).

CLT

Under assumption A1 it holds that

(i) $\alpha_k^{-1/2}(\theta_k - \theta^\star) \xrightarrow{W} \mathcal{N}(0, \Sigma_1)$ if α_k satisfy (i);

(ii) $\alpha_k^{-1/2}(\theta_k - \theta^\star) \xrightarrow{W} \mathcal{N}(0, \Sigma_2)$ if α_k satisfy (ii).

Why averaging: CLT view

Covariances Σ_1 and Σ_2 are given by

$$-\Sigma_1\bar{\mathbf{A}}^\top - \bar{\mathbf{A}}\Sigma_1 = -\Sigma_\varepsilon \tag{6}$$

$$\Sigma_2(\mathrm{I} - 2\alpha_*\bar{\mathbf{A}}^\top) + (\mathrm{I} - 2\alpha_*\bar{\mathbf{A}})\Sigma_2 = -2\alpha_*\Sigma_\varepsilon\,. \tag{7}$$

Suggests that $\alpha_k = \alpha_\star/k$ is optimal. However, such a choice of step size is not implementable.

Optimal preconditioner choice

Consider now the modified LSA dynamics

$$\tilde{\theta}_k = \tilde{\theta}_{k-1} - \alpha_k \Gamma(\mathbf{A}_k\tilde{\theta}_{k-1} - \mathbf{b}_k)\,,$$

where $\alpha_k = \alpha_\star/k$ and Γ - fixed matrix. We know that

$$\alpha_k^{-1/2}(\theta_n - \theta^\star) \xrightarrow{W} \mathcal{N}(0, \Sigma_2(\Gamma))\,.$$

Can we find $\Gamma^\star$, such that for any $u \in \mathbb{R}^d$:

$$u^\top\Sigma_2(\Gamma^\star)u \le u^\top\Sigma_2(\Gamma)u$$

Optimality of Polyak-Ruppert

Optimal preconditioning

Optimal choice of $\Gamma^\star$ is given by

$$\Gamma^\star = \alpha_*^{-1} \bar{\mathbf{A}}^{-1} \,,$$

corresponding to the covariance matrix

$$\Sigma_2(\Gamma^\star) = \alpha_*^{-1} \bar{\mathbf{A}}^{-1} \Sigma_\varepsilon \bar{\mathbf{A}}^{-\top} .$$

Under A1 and (i)-th choice of step size, the Polyak-Ruppert Polyak and Juditsky [1992] averaging performs almost similarly:

$$\sqrt{n}(\bar{\theta}_n - \theta^\star) \xrightarrow{W} \mathcal{N}(0, \bar{\mathbf{A}}^{-1} \Sigma_\varepsilon \bar{\mathbf{A}}^{-\top}) \,.$$

Extensions to the Markov setting are given in Fort [2015].

Boris Polyak (1935-2023)

Problem setting

- Goal: sharp bounds for finite-sample n and the dimension of the parameter space d;
- Constant step size α depending on the computational budget n;
- For least squares regression problems, where $\mathbf{A}(Z_n)$ is a symmetric matrix almost surely, Bach and Moulines [2013] showed that for a constant step size, the MSE of $\bar{\theta}_{n_0,n} - \theta^\star$ converges as $\mathcal{O}(1/n)$;
- General LSA: Lakshminarayanan and Szepesvari [2018] showed a rate of convergence of the MSE $\mathcal{O}(1/n)$.
- Mou et al. [2020] provided a non-asymptotic high-probability bounds for LSA-PR with independent observations. However, the proof relies on concentration bounds from Markov chain - under Log-Sobolev inequalities- $\{(\mathbf{A}(Z_n), \mathbf{b}(Z_n))\}_{n\in\mathbb{N}}$ - clear gaps in the proof.

Non-asymptotic LSA expansions

- Denote by $\Gamma_{1:n}^{(\alpha)}$ the product of random matrices

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

- The recursion $\theta_n = \theta_{n-1} - \alpha_n \{\mathbf{A}(Z_n)\theta_{n-1} - \mathbf{b}(Z_n)\}$ may be decomposed as follows

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

where $\tilde{\theta}_n^{(\mathrm{tr})}$ is the transient term and $\tilde{\theta}_n^{(\mathrm{fl})}$ is a fluctuation term

$$\tilde{\theta}_n^{(\mathrm{tr})} = \Gamma_{1:n}^{(\alpha)} \{\theta_0 - \theta^\star\} \,, \quad \tilde{\theta}_n^{(\mathrm{fl})} = -\alpha \sum_{j=1}^{n} \Gamma_{j+1:n}^{(\alpha)} \varepsilon(Z_j) \,.$$

- A cornerstone of the theoretical analysis is a tight bound for $\mathbb{E}^{1/p}[\|\Gamma_{m:n}^{(\alpha)}\|^p]$ under some assumptions on the matrix $\bar{\mathbf{A}}$.

Exponential stability of random matrix products

Key technical element:

Exponential stability of $\{\mathbf{A}(Z_i)\}_{i\in\mathbb{N}}$ (see Guo and Ljung [1995], Ljung [2002])

For $q \geq 1$, there exist $\mathrm{a}_q, \mathrm{C}_q > 0$ and $\alpha_{\infty,q} < \infty$ such that, for any step size $\alpha \leq \alpha_{\infty,q}$, $m, n \in \mathbb{N}$, $m < n$,

$$\mathbb{E}[\|\Gamma^{(\alpha)}_{m:n}\|^q] \leq \mathrm{C}_q \exp\left(-\mathrm{a}_q \alpha (n-m)\right) .$$

- Intuitively, exponential stability means that $\Gamma^{(\alpha)}_{m:n} \approx (\mathrm{I} - \alpha\bar{\mathbf{A}})^{n-m}$, for $m, n \in \mathbb{N}$, $m \leq n$;
- We handle both the setting of i.i.d.and Markov dependency in the sequence $\{Z_i\}_{i\in\mathbb{N}}$;

Lyapunov equation

Proposition

Assume that $-\bar{\mathbf{A}}$ *is Hurwitz. There exists a unique symmetric positive definite matrix* Q *satisfying the Lyapunov equation* $\bar{\mathbf{A}}^\top Q + Q\bar{\mathbf{A}} = \mathrm{I}$. *In addition, setting*

$$a = \|Q\|^{-1}/2\,, \quad \textit{and} \quad \alpha_\infty = (1/2)\|\bar{\mathbf{A}}\|_Q^{-2}\|Q\|^{-1} \wedge \|Q\|\,,$$

for any $\alpha \in [0, \alpha_\infty]$, *it holds that*

$$\|\mathrm{I} - \alpha\bar{\mathbf{A}}\|_Q^2 \leq 1 - a\alpha\,,$$

and $\alpha a \leq 1/2$.

Why Q-norm: $\mathrm{I} - \alpha\bar{\mathbf{A}}$ is a strict contraction in $\|\cdot\|_Q$, but not necessarily in $\|\cdot\|$.

Exponential stability under A1

Theorem

Assume IND and A1. For any $p, q \in \mathbb{N}$, $2 \le p \le q$, $\alpha \in (0, \alpha_{q,\infty}]$ and $n \in \mathbb{N}$, it holds

$$\mathbb{E}^{1/p}\left[\|\Gamma_{1:n}^{(\alpha)}\|^p\right] \le \sqrt{\kappa_{\mathrm{Q}}} d^{1/q}(1 - a\alpha + (q-1) b_Q^2 \alpha^2)^{n/2}\,.$$

where

$$\kappa_{\mathrm{Q}} = \lambda_{\max}(Q)/\lambda_{\min}(Q)\,, \quad b_Q = \sqrt{\kappa_{\mathrm{Q}}}\, \mathrm{C}_A\,,$$
$$\alpha_{q,\infty} = \alpha_\infty \wedge \mathrm{c}_{\mathbf{A}}\,/q\,, \quad \mathrm{c}_{\mathbf{A}} = a/\{2 b_Q^2\}\,.$$

- Note that the bound above introduces an interplay between step size α and maximal controlled moment q;
- We show that under only A1, for fixed $\alpha > 0$, $\lim_{n \to +\infty} \mathbb{E}[\|\theta_n - \theta^\star\|^p] = \infty$ for $p \ge \bar{p}(\alpha)$; cannot expect exponential tytpe HPB for $\|\theta_n - \theta^\star\|$ are not possible (see Durmus et al. [2021])

Exponential stability: sketch of the proof

- For $B \in \mathbb{R}^{d \times d}$ let $\sigma_\ell(B), \ell = 1, \ldots, d$ be its singular values;
- For $p \geq 1$, denote its Schatten p-norm

$$\|B\|_p = \{\sum_{\ell=1}^{d} \sigma_\ell^p(B)\}^{1/p}$$

- For $p, q \geq 1$ and random matrix X, we write $\|X\|_{p,q} = \{\mathbb{E}[\|X\|_p^q]\}^{1/q}$.

Theorem (Subquadratic averages - (Huang et al., 2020))

Consider random matrices of the same sizes that satisfy $\mathbb{E}[Y|X] = 0$, $\mathbb{P}$-a.s. Then, for $2 \leq q \leq p$,

$$\|X + Y\|_{p,q}^2 \leq \|X\|_{p,q}^2 + C_p \|Y\|_{p,q}^2$$

The constant $C_p = p - 1$ is the best possible.

Proof sketch: re-write a product of matrices

$$\Gamma_{1:n}^{(\alpha)} = (\mathrm{I} - \alpha \bar{\mathbf{A}}) \Gamma_{1:n-1}^{(\alpha)} - \alpha (\mathbf{A}(Z_n) - \bar{\mathbf{A}}) \Gamma_{1:n-1}^{(\alpha)} \, ,$$

then apply the subquadratic inequality above, switch to Q-norm.

Linear Stochastic Approximation

- Recall that the error vector $\theta_n - \theta^\star$ may be decomposed as

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

- To bound $\mathbb{E}^{1/p}[\|\tilde{\theta}_n^{(\mathsf{tr})}\|^p]$, we simply apply the bound on the matrix product.
- How to proceed with $\mathbb{E}^{1/p}[\|\tilde{\theta}_n^{(\mathsf{fl})}\|^p]$?

Sketch of the proof: fluctuation term

- For any $n \in \mathbb{N}$:
$$\tilde{\theta}_n^{(\mathrm{fl})} = J_n^{(0)} + H_n^{(0)}\,, \tag{8}$$
where the latter terms are defined by the following pair of recursions
$$\begin{aligned} J_n^{(0)} &= \left(\mathrm{I} - \alpha \bar{\mathbf{A}}\right) J_{n-1}^{(0)} - \alpha \varepsilon(Z_n)\,, & J_0^{(0)} &= 0\,, \\ H_n^{(0)} &= (\mathrm{I} - \alpha \mathbf{A}(Z_n))\, H_{n-1}^{(0)} - \alpha \tilde{\mathbf{A}}(Z_n) J_{n-1}^{(0)}\,, & H_0^{(0)} &= 0\,. \end{aligned} \tag{9}$$
- Solving the recursion above,
$$J_n^{(0)} = -\alpha \sum_{j=1}^{n} \left(\mathrm{I} - \alpha \bar{\mathbf{A}}\right)^{n-j+1} \varepsilon(Z_j), \quad H_n^{(0)} = -\alpha \sum_{j=1}^{n} \Gamma_{j+1:n+1}^{(\alpha)} \tilde{\mathbf{A}}(Z_j) J_{j-1}^{(0)}\,.$$
- The term $J_n^{(0)}$ is the leading one w.r.t. α, and is a linear statistics of $\{\varepsilon(Z_j)\}_{j \geq 0}$;
- Rough bounds from (9):
$$\mathbb{E}^{1/p}[\| J_n^{(0)} \|^p] \lesssim \sqrt{\alpha}\,, \quad \mathbb{E}^{1/p}[\| H_n^{(0)} \|^p] \lesssim \sqrt{\alpha}\,.$$

Sketch of the proof: fluctuation term

The same decomposition can be applied to $H_n^{(0)}$ to obtain higher order expansions:

$$H_n^{(0)} = \sum_{\ell=1}^{L} J_n^{(\ell)} + H_n^{(L)}\,, \tag{10}$$

where for any $\ell \in \{1, \ldots, L\}$,

$$\begin{aligned} J_n^{(\ell)} &= \left(\mathrm{I} - \alpha \bar{\mathbf{A}}\right) J_{n-1}^{(\ell)} - \alpha \tilde{\mathbf{A}}(Z_n) J_{n-1}^{(\ell-1)}\,, \qquad && J_0^{(\ell)} = 0\,, \\ H_n^{(L)} &= \left(\mathrm{I} - \alpha \mathbf{A}(Z_n)\right) H_{n-1}^{(L)} - \alpha \tilde{\mathbf{A}}(Z_n) J_{n-1}^{(L)}\,, && H_0^{(L)} = 0\,. \end{aligned} \tag{11}$$

The choice of parameter L controls the desired approximation accuracy:

$$\mathbb{E}^{1/p}[\|J_n^{(\ell)}\|^p] \lesssim \alpha^{(\ell+1)/2}\,, \quad \mathbb{E}^{1/p}[\|H_n^{(L)}\|^p] \lesssim \alpha^{(L+1)/2}\,.$$

Combining (8) and (10), we obtain the decomposition which is the cornerstone of our analysis:

$$\tilde{\theta}_n^{(\mathrm{fl})} = \sum_{\ell=0}^{L} J_n^{(\ell)} + H_n^{(L)}. \tag{12}$$

p-th moment bound for the LSA error $\|\theta_n - \theta^\star\|$

Theorem

Assume IND and A1. Then, for any $p, q \in \mathbb{N}$, $2 \le p \le q$, $\alpha \in (0, \alpha_{q,\infty}]$, $n \in \mathbb{N}$, and $\theta_0 \in \mathbb{R}^d$ it holds

$$\mathbb{E}^{1/p}\left[\|\theta_n - \theta^\star\|^p\right] \le d^{1/q} \kappa_{\mathrm{Q}}^{1/2} (1 - \alpha a/4)^n \|\theta_0 - \theta^\star\| + d^{1/q} \mathrm{D}_2 \sqrt{\alpha a p} \|\varepsilon\|_\infty ,$$

where D_2 has closed-form expression.

Polyak-Ruppert averaging

$$\bar{\theta}_{n_0,n} = (n-n_0)^{-1} \sum_{k=n_0}^{n-1} \theta_k \,, \quad n \geq n_0 + 1$$

Key decomposition

For any $n, n_0 \in \mathbb{N}$, $n_0 \leq n$,

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

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

Using (12), we may further decompose

$$\textstyle\sum_{t=n_0}^{n-1} e\left(\theta_t, Z_{t+1}\right) = E^{\mathrm{tr}}_{n_0,n} + E^{\mathrm{fl}}_{n_0,n} \,,$$

where we have set

$$E^{\mathrm{tr}}_{n_0,n} = \textstyle\sum_{t=n_0}^{n-1} \tilde{\mathbf{A}}(Z_{t+1}) \Gamma^{(\alpha)}_{1:t} \{\theta_0 - \theta^\star\} \,,$$

$$E^{\mathrm{fl}}_{n_0,n} = \textstyle\sum_{t=n_0}^{n-1} \varepsilon(Z_{t+1}) + \sum_{\ell=0}^{L} \sum_{t=n_0}^{n-1} \tilde{\mathbf{A}}(Z_{t+1}) J^{(\ell)}_t + \sum_{t=n_0}^{n-1} \tilde{\mathbf{A}}(Z_{t+1}) H^{(L)}_t \,.$$

Polyak-Ruppert averaging

Theorem

Assume IND and A1. Then, for any $p \geq 2$, $n \geq 2$, burn-in period $n_0 = n/2$, step size

$$\alpha(n, d, p) \asymp \frac{1}{(1 + \log d) p n^{1/2}} , \tag{13}$$

and an initial parameter $\theta_0 \in \mathbb{R}^d$, it holds that

$$\mathbb{E}^{1/p} \left[\| \bar{\mathbf{A}} \left(\bar{\theta}_{n_0,n} - \theta^\star \right) \|^p \right] \lesssim_d \frac{\{\operatorname{Tr} \Sigma_\varepsilon\}^{1/2} p^{1/2}}{n^{1/2}} + \|\varepsilon\|_\infty \left(\frac{p}{n^{3/4}} + \frac{p^2}{n} \right) + p \|\theta_0 - \theta^\star\| \exp\left\{ -\frac{(\alpha_\infty \wedge c_{\mathbf{A}})\sqrt{n}}{8p(1 + \log d)} \right\} , \tag{14}$$

where $\Sigma_\varepsilon = \int_{\mathsf{Z}} \varepsilon(z)\varepsilon(z)^\top \mathrm{d}\pi(z)$.

The leading term is the p-moment of the Gaussian appearing in the CLT !

Markovian Setting

Markovian setting

For any $\mathsf{A} \in \mathcal{Z}$, $\mathbb{P}_\xi(Z_k \in \mathsf{A} | Z_{k-1}) = \mathrm{Q}(Z_{k-1}, \mathsf{A})$, $\mathbb{P}_\xi$-a.s.

Assumption UGE

The Markov kernel Q is Uniformly Geometrically Ergodic, i.e., there exists $t_{\mathrm{mix}} \in \mathbb{N}^*$ such that for all $k \in \mathbb{N}^*$,

$$\Delta(\mathrm{Q}^k) = \sup_{z, z' \in \mathsf{Z}} (1/2) \| \mathrm{Q}^k(z, \cdot) - \mathrm{Q}^k(z', \cdot) \|_{\mathrm{TV}} \leq (1/4)^{\lfloor k / t_{\mathrm{mix}} \rfloor} . \tag{15}$$

Here, t_{mix} is the mixing time of Q.

- UGE implies that π is the unique invariant distribution of Q;
- UGE is equivalent to the uniform minorization condition, i.e., there exists a probability measure ν such that for all $z \in \mathsf{Z}$, $\mathsf{A} \in \mathcal{Z}$,

$$\mathrm{Q}^{t_{\mathrm{mix}}}(z, \mathsf{A}) \geq (1/4) \nu(\mathsf{A}) .$$

Exponential stability: Markovian case

Define the quantities

$$\alpha^{(\mathrm{M})}_{q,\infty} = \alpha^{(\mathrm{M})}_{\infty} \wedge \mathsf{c}_{\mathbf{A}} / q \,, \tag{16}$$

where $\alpha^{(\mathrm{M})}_{\infty}$ depends upon constants from A1 and κ_{Q}. Then:

Theorem

Assume UGE and A1. Then, for any $2 \leq p \leq q$, $\alpha \in (0, \alpha^{(\mathrm{M})}_{\infty} t_{\mathsf{mix}}^{-1}]$, $n \in \mathbb{N}$, and probability distribution ξ on $(\mathsf{Z}, \mathcal{Z})$, it holds

$$\mathbb{E}_{\xi}^{1/p}\left[\|\Gamma^{(\alpha)}_{1:n}\|^{p}\right] \leq \sqrt{\kappa_{\mathsf{Q}}}\mathrm{e}^{2} d^{1/q} \exp\{-n\alpha a/6 + n(q-1)\alpha^{2}\, \mathsf{C}_{\Gamma}\} \,, \tag{17}$$

where $\alpha^{(\mathrm{M})}_{\infty}$ is some constant. Moreover, for $\alpha \in (0, \alpha^{(\mathrm{M})}_{q,\infty} t_{\mathsf{mix}}^{-1}]$, it holds

$$\mathbb{E}_{\xi}^{1/p}\left[\|\Gamma^{(\alpha)}_{1:n}\|^{p}\right] \leq \sqrt{\kappa_{\mathsf{Q}}}\mathrm{e}^{2} d^{1/q} \mathrm{e}^{-a\alpha n/12} \,. \tag{18}$$

Covariance matrix

Noise covariance matrix

Under A1 and UGE, we define the matrix $\Sigma_\varepsilon^{(\mathrm{M})}$ as

$$\Sigma_\varepsilon^{(\mathrm{M})} = \mathbb{E}_\pi[\varepsilon(Z_0)\varepsilon(Z_0)^\top] + 2\sum_{\ell=0}^{\infty} \mathbb{E}_\pi[\varepsilon(Z_0)\varepsilon(Z_\ell)^\top]. \tag{19}$$

- For any initial probability measure ξ on $(\mathsf{Z}, \mathcal{Z})$, $n^{-1/2}\sum_{t=0}^{n-1}\varepsilon(Z_t)$ converges in distribution to $\mathcal{N}(0, \Sigma_\varepsilon^{(\mathrm{M})})$;
- We expect that this is also the leading term in the bound for $\mathbb{E}_\xi^{1/p}\left[\|\bar{\mathbf{A}}\left(\bar{\theta}_n - \theta^\star\right)\|^p\right]$

p-th moment bound for the LSA error $\|\theta_n - \theta^\star\|$

Theorem

Assume A1 and UGE. Let $2 \le p \le q/2$ and $\alpha^{(\mathrm{M})}_{q,\infty}$ be defined in (16). Then, for any $\alpha \in (0, \alpha^{(\mathrm{M})}_{q,\infty} t_{\mathsf{mix}}^{-1}]$, $\theta_0 \in \mathbb{R}^d$, initial probability measure ξ on $(\mathsf{Z}, \mathcal{Z})$, and $n \in \mathbb{N}$, it holds

$$\mathbb{E}_{\xi}^{1/p}\left[\|\theta_n - \theta^\star\|^p\right] \le \sqrt{\kappa_{\mathsf{Q}}}\mathrm{e}^2 d^{1/q} \mathrm{e}^{-\alpha a n/12}\|\theta_0 - \theta^\star\| + \mathsf{D}_2^{(\mathsf{M})} d^{1/q}\sqrt{\alpha a p t_{\mathsf{mix}}}\|\varepsilon\|_\infty\,,$$

where $\mathsf{D}_2^{(\mathsf{M})}$ is some constant.

Polyak-Ruppert averaging

Theorem

Assume UGE and A1. Then, for any $p \geq 2$, $n \geq 4 \vee t_{\mathrm{mix}}$, step size

$$\alpha^{(\mathrm{M})}(n, d, p, t_{\mathrm{mix}}) \asymp \frac{1}{(1+\log d) p n^{2/3} t_{\mathrm{mix}}^{1/3}}\,,$$

initial parameter $\theta_0 \in \mathbb{R}^d$, and initial probability measure ξ on $(\mathsf{Z}, \mathcal{Z})$, it holds that

$$\mathbb{E}_{\xi}^{1/p}\left[\|\bar{\mathbf{A}}\left(\bar{\theta}_n - \theta^\star\right)\|^p\right] \lesssim_{d,n} \frac{\{\operatorname{Tr} \Sigma_{\varepsilon}^{(\mathrm{M})}\}^{1/2} p^{1/2}}{n^{1/2}} + \|\varepsilon\|_\infty \left(\frac{t_{\mathrm{mix}}^{2/3} p}{n^{2/3}} + \frac{t_{\mathrm{mix}} p^2}{n}\right)$$
$$+ p n^{1/2} \|\theta_0 - \theta^\star\| \exp\left\{ -\frac{(\alpha_\infty^{(\mathrm{M})} \wedge \mathsf{c}_{\mathbf{A}}^{(\mathrm{M})}) n^{1/3}}{24 p t_{\mathrm{mix}}^{1/3} (1+\log d)} \right\}.$$

Remark: unlike the i.i.d. noise scenario,

$$\mathbb{E}_\pi[\bar{\theta}_n] \neq \theta^\star\,, \quad \textit{moreover}, \mathbb{E}_\pi[\bar{\theta}_n] = \mathcal{O}(\alpha)\,.$$

Rosenthal-type inequality for Markov chains

Key technical innovation - novel Rosenthal-type inequalities of Durmus et al. [2023].

Rosenthal type inequality

Let $\{Z_k\}_{k \geq 1}$ be a Markov chain on $(\mathsf{Z}, \mathcal{Z})$ with Markov kernel Q, satisfying UGE. Then, for any bounded $f : \mathsf{Z} \to \mathbb{R}$, and $p \geq 2$ it holds

$$\mathbb{E}_{\pi}^{1/p}[|\sum_{\ell=1}^{n}(f(Z_\ell) - \pi(f))|^p] \lesssim p^{1/2} n^{1/2} \sigma_\infty(f) + \\ n^{1/4} t_{\mathrm{mix}}^{3/4} p \log_2(2p) \|f\|_\infty + t_{\mathrm{mix}} p \log_2(2p) \|f\|_\infty \,. \quad (20)$$

Applications to TD learning

TD learning

Optimal parameter

Define $\theta^\star$ as a solution of the minimization problem

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

Error norm

Consider the following distance between the parameters:

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

Previous results: discussion

- Bhandari et al. [2018]: RSA ("robust stochastic approximation") framework following Nemirovski et al. [2009]. Here

$$\mathbb{E}^{1/2}[\|\bar{\theta}_n - \theta^\star\|^2_{\Sigma_\varphi}] = \mathcal{O}(1/\sqrt{n})\,.$$

  Advantages: step size α and bounds independent of conditioning;
- Li et al. [2023b]: Lower bounds on the MSE for policy evaluation problems and optimal MSE for the variance-reduced TD-learning algorithm (based on control variates);
- Li et al. [2023a]: HPB and sample complexity for TD(0) and off-policy counterpart (TDC). Step size α scales with the minimal eigenvalue of the feature matrix and covers i.i.d. setting only;
- Patil et al. [2023]: second moment for TD(0) and high-probability bounds for projected TD (0) iterates. HPBs require a projection on a ball - radius depends on $\|\theta^\star\|$.

Matrix stability in TD

Checking matrix stability for TD learning

Let $\{\theta_k\}_{k \in \mathbb{N}}$ be a sequence of TD(0) updates under TD1 and TD2. Then this update scheme satisfies the stability assumption A2(p) with

$$a = \frac{(1-\gamma)\lambda_{\min}}{2}\,, \quad \varkappa_p = 1\,, \quad \alpha_{p,\infty} = \frac{1-\gamma}{128p}\,. \tag{21}$$

Discussion: Previous results – Huang et al. [2021] and Durmus et al. [2021] – yield an instance-dependent stability threshold

$$\alpha_{p,\infty} = \frac{(1-\gamma)\lambda_{\min}}{c_0 p} \tag{22}$$

for some absolute constant $c_0 > 0$. The same order of magnitude of the step size is predicted in [Li et al., 2023a, Theorem 1].

Stability of matrix product

Theorem: Matrix stability for TD learning

Let $\{\theta_k\}_{k \in \mathbb{N}}$ be a sequence of TD(0) updates under TD1 and TD2. Then, for any $n \in \mathbb{N}$, $1 \leq j \leq n$, $p \geq 2$, step size $\alpha \in \left(0; \frac{1-\gamma}{128p}\right]$, it holds $\mathbb{P}$-a.s. that

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

TD learning: Proof of matrix stability

Result from Patil et al. [2023]

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

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

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

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

Hence, $\mathbb{E}[\mathbf{A} + \mathbf{A}^\top] \succeq 2(1-\gamma)\Sigma_\varphi$. Similarly, one can show by direct computations that

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

TD learning: Proof of matrix stability

Lemma

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

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

Lemma

For random matrix $\mathbf{A}$ defined in (4) and $\mathbf{B} = \mathbf{A} + \mathbf{A}^\top - \alpha \mathbf{A}^\top \mathbf{A}$, for $p \in \mathbb{N}$ and step size $\alpha \in (0; \frac{1-\gamma}{(1+\gamma)^2}]$ it holds that

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

TD learning: Proof of matrix stability

Lemma: key lemma for p-th moment stability

Let $\mathbf{A} = \varphi(s)\{\varphi(s) - \gamma\varphi(s')\}^\top$ be a random TD update matrix defined in (4), where $s' \sim P^\pi(\cdot|s)$, and $s \sim \mu$. Then, for any $p \in \mathbb{N}$ and step size

$$\alpha \in (0; \frac{1-\gamma}{64p}]\,,$$

it holds that

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

TD learning: 2nd moment bound

Theorem 2: second moment error for tail-averaging

Let $\{\theta_k\}_{k \in \mathbb{N}}$ be a sequence of TD(0) updates generated by (3) under TD1 and TD2. Then for any $n \geq 2$, $\alpha \in (0; (1-\gamma)/256]$, and $\theta_0 \in \mathbb{R}^d$, it holds that

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

where $f_1(\alpha, \lambda_{\min}, n)$ is a polynomial function in $1/\alpha, 1/\lambda_{\min}, n$.

TD sample complexity: 2-nd moment

Sample complexity

Under assumptions of Theorem 2, $\mathbb{E}[\|\bar{\theta}_n - \theta^\star\|^2_{\Sigma_\varphi}] \le \varepsilon^2$ requires where $\mathsf{R}_1(1/\varepsilon) = \frac{\|\theta^\star\|_{\Sigma_\varphi}+1}{\sqrt{\alpha}(1-\gamma)^{3/2}\lambda_{\min}\varepsilon}$.

- Set $\alpha \simeq 1-\gamma$, sample complexity (agrees with Patil et al. [2023]):

$$\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2\lambda_{\min}} \cdot \log\frac{\|\theta_0-\theta^\star\|}{\varepsilon} + \underbrace{\frac{1+\|\theta^\star\|^2_{\Sigma_\varphi}}{(1-\gamma)^2\lambda^2_{\min}\varepsilon^2}}_{\text{suboptimal by a factor } \lambda_{\min}^{-1}}\right).$$

- Set $\alpha \simeq (1-\gamma)\lambda_{\min}$, sample complexity (agrees with Li et al. [2023a]):

$$\widetilde{\mathcal{O}}\left(\underbrace{\frac{1}{(1-\gamma)^2\lambda^2_{\min}} \cdot \log\frac{\|\theta_0-\theta^\star\|}{\varepsilon}}_{\text{suboptimal by a factor } \lambda_{\min}^{-1}} + \frac{1+\|\theta^\star\|^2_{\Sigma_\varphi}}{(1-\gamma)^2\lambda_{\min}\varepsilon^2}\right).$$

TD learning: HPB

Theorem 3: high-probability error bounds for tail-averaging

Fix $\varepsilon > 0$, $\delta > 0$, assume TD1 and TD2. Let $\{\theta_k\}_{k \in \mathbb{N}}$ be a sequence of TD(0) updates generated by (3). Then for any $n \geq 2$, and step size

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

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

$$\widetilde{\mathcal{O}}\left(\frac{(\|\theta^\star\|^2_{\Sigma_\varphi}+1)\log\left(1/\delta\right)}{(1-\gamma)^2\lambda_{\min}\varepsilon^2}\left(1+\frac{\alpha\log\left(1/\delta\right)}{(1-\gamma)\lambda_{\min}}\right)+\mathrm{R}_2(1/\varepsilon,\delta)+\frac{1}{\alpha\lambda_{\min}(1-\gamma)}\log\frac{\|\theta_0-\theta^\star\|}{\varepsilon}\right).$$

TD(0) updates.

Optimizing the bound w.r.t. α yields the same dilemma as for the 2-nd moment.

Asymptotic covariance matrix

- Introduce the TD(0) covariance matrix

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

- Covariance $\Sigma_{\varepsilon}^{(TD)}$ aligns with the CLT for Polyak-Ruppert averaged iterates Fort [2015];
- Define the transformed covariance matrix

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

corresponding to $\Sigma_{\varphi}^{1/2}\bar{\mathbf{A}}^{-1}\varepsilon$.

Upper bounding the optimal covariance matrix

Under our assumptions,

$$\operatorname{Tr} \Sigma_{\varepsilon}^{(opt)} \leq \frac{\|\theta^\star\|_{\Sigma_\varphi}^2 + 1}{(1-\gamma)^2 \lambda_{\min}}\,.$$

Tighter 2-nd moment bound

Refined Theorem 2

Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (3) under TD1 and TD2. Then for any $n \geq 2$, $\alpha \in (0; (1-\gamma)/256]$, and $\theta_0 \in \mathbb{R}^d$, it holds that

$$\mathbb{E}^{1/2}[\|\bar{\theta}_n - \theta^\star\|^2_{\Sigma_\varphi}] \lesssim \frac{\sqrt{\operatorname{Tr}\Sigma_\varepsilon^{(opt)}}}{n^{1/2}} + \frac{1 + \|\theta^\star\|_{\Sigma_\varphi}}{(1-\gamma)^{3/2}\lambda_{\min} n^{1/2}}\left(\frac{1}{\sqrt{\alpha n}} + \sqrt{\alpha}\right) + f_2(\alpha, \lambda_{\min}, n)\big(1 - \alpha(1-\gamma)\lambda_{\min}\big)^{n/2}\|\theta_0 - \theta^\star\|\,, \tag{23}$$

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

Markovian sampling: assumptions

Trajectory-wise evaluation (instead of TD1):

Assumption TD3

Agent's learning is based on tuples (s_k, a_k, s_{k+1}) which are generated sequentially following the generative model $a_k \sim \pi(\cdot|s_k)$, $s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$.

The assumption TD3 yields that the sequence $\{s_k\}_{k\in\mathbb{N}}$ is a Markov chain with the Markov kernel $\mathcal{P}_\pi(\cdot|s)$.

Assumption TD4

The Markov kernel $\mathcal{P}_\pi$ admits a unique invariant distribution μ and is uniformly geometrically ergodic, that is, there exist $t_{\mathsf{mix}} \in \mathbb{N}$, such that for any $s \in S$ and $k \in \mathbb{N}$ it holds that

$$\left\|\mathcal{P}_\pi^k(\cdot|s) - \mu\right\|_{\mathrm{TV}} \leq (1/4)^{\lceil k/t_{\mathsf{mix}} \rceil} \,. \tag{24}$$

One can consider the generalisations of TD4 coming at a price of more technical work.

TD with Markovian sampling

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

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

for $k = 0, \ldots, n$: **do**

  Receive tuple (s_k, a_k, s'_k) following TD4

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

    Compute update

$$\tilde{\theta}_j = \tilde{\theta}_{j-1} - \alpha(\mathbf{A}_k \tilde{\theta}_{j-1} - \mathbf{b}_k)$$

    based on $\mathbf{A}_k$, $\mathbf{b}_k$ from (4)

  else

    skip current learning tuple

  end

end

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

value function estimate $V^{\pi}_{\bar{\theta}_n}(s) = \varphi^\top(s)\, \bar{\theta}_n$ Idea goes back to Nagaraj et al. [2020], Patil et al. [2023].

Markovian sampling schemes

Refined Theorem 2

Let $\{\theta_k\}_{k\in\mathbb{N}}$ be a sequence of TD(0) updates generated by (3) under TD2, TD3, and TD4, and $\bar{\theta}_n$ be a tail-averaged estimate generated by Algorithm **??** with $q = t_{\mathrm{mix}}$. Then, for the step size and sample size satisfy

$$\alpha = \frac{1-\gamma}{128\log(n/\delta)}, \quad n \geq \frac{\log(1/\delta)}{(1-\gamma)^2} \vee \frac{2t_{\mathrm{mix}}\log(4/\delta)}{\log 4}$$

in order to achieve $\|\bar{\theta}_n - \theta^\star\|_{\Sigma_\varphi} \leq \varepsilon$ with probability at least $1-3\delta$, it requires

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

observations.

Markovian sampling schemes

- Proof is based on Berbee's coupling lemma Berbee [1979];
- Bounds scale by a factor t_{mix} compared to the i.i.d. setting;
- Extra $\sqrt{\log 1/\delta}$ factor in the leading term as an artefact of applying Berbee's construction;
- Using Berbee's construction potentially can be avoided, but requires to adjust the step size $\alpha \approx t_{\mathsf{mix}}^{-1}$. Hence, the knowledge of t_{mix} is still required.

Conclusion and open questions

Adaptive version

Is it possible to come up with a version of Algorithm 1, which does not require to know t_{mix} in advance?

Optimal bounds for instance-independent step size

Is it possible to remove the extra $\lambda_{\mathsf{min}}^{-1}$ in the analysis of Theorem 2 for the step size α independent of λ_{min}? Or construct a lower bound showing that this suboptimality is not an artefact of the proof.

References:

- Durmus, A., Moulines, E., Naumov, A., Samsonov, S., Wai, H. T. (2021, July). On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In Conference on Learning Theory (pp. 1711-1752). PMLR.

- Durmus, A., Moulines, E., Naumov, A., Samsonov, S., Wai, H. T. (2021, July). On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In Conference on Learning Theory (pp. 1711-1752). PMLR.

- Durmus, A., Moulines, E., Naumov, A., Samsonov, S. (2024). Finite-Time High-Probability Bounds for Polyak–Ruppert Averaged Iterates of Linear Stochastic Approximation. Mathematics of Operations Research. https://arxiv.org/abs/2207.04475
- Samsonov, S., Tiapkin, D., Naumov, A., Moulines, E. (2024), Improved high-probability bounds for temporal difference learning algorithms wia exponential stability, https://arxiv.org/abs/2310.14286 - To appear in COLT-2024.

Thank you!

References I

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *NeurIPS*, volume 26, 2013.

H.C.P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*. Mathematical Centre tracts. Centrum Voor Wiskunde en Informatica, 1979. ISBN 9789061961826. URL https://books.google.ru/books?id=ivUAtQEACAAJ.

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

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

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize. In *NeurIPS*, 2021.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Marina Sheshukova. Rosenthal-type inequalities for linear statistics of markov chains. *arXiv preprint arXiv:2303.05838*, 2023.

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

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

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

C.r Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *AISTATS*, volume 84, pages 1347–1355, 2018.

References II

Gen Li, Weichen Wu, Yuejie Chi, Cong Ma, Alessandro Rinaldo, and Yuting Wei. Sharp high-probability sample complexities for policy evaluation with linear function approximation. *arXiv preprint arXiv:2305.19001*, 2023a.

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

Lennart Ljung. Recursive identification algorithms. *Circuits, Systems and Signal Processing*, 21(1): 57–68, 2002.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on learning theory*, volume 125, pages 2947–2997. PMLR, 2020.

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

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

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

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

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