A Law of Iterated Logarithm for Multi-Agent Reinforcement Learning

Gugan Thoppe IISc

Bhumesh Kumar WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

Highlights

- Law of iterated logarithm for distributed stochastic approx.
- Convergence rate along sample paths where algorithm converges
- Weaker assumptions on the gossip matrix and stepsizes
- A novel concentration result for a sum of martingale differences
- Applies to distributed TD(0) with linear function approximation

Background

- Reinforcement Learning
  - Train machines the same way an infant learns
  - Interact with the environment and figure out the optimal action sequence needed to complete a given task
- Stochastic Approximation (SA)
  - Theory provides toolkit for rigorously analyzing RL algorithms
  - Iterative algorithms useful to find zeroes or optimal points of functions, for which only noisy evaluations are possible
- This work: Analyze distributed SA algorithms useful in MARL

Multi-agent Reinforcement Learning

- Cooperative MARL
- Multiple agents continually interact with an environment
  - Agents picks local actions
  - Environment reacts to the joint action by transitioning to a new state and giving each agent a local reward
  - Agents gossip about local computations with each other
- Aim: Find action policies that maximize collective rewards
- Usage : Gaming, Robotics, Communications, Power Grids, Finance

Distributed Stochastic Approximation

- m agents, directed graph $\mathcal{G}$, matrix $W \equiv (W_{ij}) \in [0,1]^{m \times m}$

- $W_{ij} \in [0,1]$ denotes the strength of the edge $j \longrightarrow i$ in $\mathcal{G}$

- Update rule at agent i

$$\underbrace{x_{n+1}(i)}_{1\times d} = \sum_{j=1}^{m} W_{ij} \underbrace{x_n(j)}_{1\times d} + \alpha_n \left[\underbrace{h_i(x_n)}_{h_i:\mathbb{R}^{m\times d}\to\mathbb{R}^d} + \underbrace{M_{n+1}(i)}_{1\times d} \right],$$

where $x(i)$ denotes the i-th row of the matrix x and

$M_{n+1}(i)$ is the noise in the estimate of $h_i(x_n)$

- Joint Update Rule: $\boldsymbol{\underbrace{x_{n+1}}_{m\times d} = W x_n + \alpha_n [h(x_n) + M_{n+1}]}$

Main Result: Law of Iterated Logarithm

- Let x_* be a potential limit of the DSA algorithm

- Let $\mathcal{E}(x_*)$ be the event $\{x_n \to x_*\}$ and $t_{n+1} = \sum_{k=0}^{n} \alpha_k$

- Then, there exists some deterministic constant $C \geq 0$ such that

$$\limsup_{n\to\infty} [\alpha_n \log t_{n+1}]^{-1/2} \|x_n - x_*\| \leq C \quad \text{a.s. on } \mathcal{E}(x_*).$$

- Why Law of Iterated Logarithm (LIL)?

  Proof uses an LIL for a sum of scaled martingale differences

Assumptions on the Gossip Matrix W

$\mathcal{A}_1$. W is an irreducible aperiodic row stochastic matrix

$\exists$ a unique row vector $\pi \in \mathbb{R}^m$ such that $\pi W = \pi$

Thm. 1 in [Mathkar and Borkar, 2016]: A DSA algorithm converges to an invariant set of the m-fold product of the ODE

$$\dot{y}(t) = \underbrace{\pi}_{1\times m} \underbrace{h(\mathbf{1}^\top y(t))}_{m\times d}$$

Any such invariant set is a subset of $\mathcal{S} := \{\mathbf{1}^\top y : y \in \mathbb{R}^d\} \subset \mathbb{R}^{m\times d}$

Let $x_* = \mathbf{1}^\top y_*$, where y_* is an asymptotically stable equilibrium of the above ODE (need not be the only attractor)

Assumptions on h

$\mathcal{A}_2$. There exists a neighbourhood $\mathcal{U}$ of x_* such that, for $x \in \mathcal{U}$,

$$h(x) = -\mathbf{1}^\top \pi (x - x_*) A + \mathbf{1}^\top \pi f_1(x) + (\mathbb{I} - \mathbf{1}^\top \pi)(B + f_2(x)),$$

where

$A \in \mathbb{R}^{d \times d}$ is positive definite, i.e., $yAy^\top > 0$ for all $y \neq 0$,

$B \in \mathbb{R}^{m \times d}$ is some constant matrix,

$f_2 : \mathcal{U} \to \mathbb{R}^{m \times d}$ is some arbitrary continuous function, while

$f_1 : \mathcal{U} \to \mathbb{R}^{m \times d}$ is another continuous function such that

$$\|\mathbf{1}^\top \pi f_1(x)\| = \mathcal{O}(\|\mathbf{1}^\top \pi (x - x_*)\|^a), \qquad \text{as } x \to x_*, \tag{1}$$

under some norm $\|\cdot\|$ and for some $a > 1$

Assumptions on Stepsize

$\mathcal{A}_3$. (α_n) is either of Type 1 or Type γ.

Type 1: $\alpha(n) = \alpha_0/n$ for a suitably large α_0

Type γ : $Cn^{-\gamma}$ and $n^{-\gamma}(\log n)^{\eta}$ for $\gamma \in (0, 1)$

$\gamma > 2/b$, where b is the constant that is defined on the next slide

Assumptions on Noise

$\mathcal{A}_4$. Let $\mathcal{F}_n = \sigma(x_0, M_1, \ldots, M_n)$ and $\mathcal{E}(x_*) = \{x_n \to x_*\}$

$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = 0$ a.s.

$\exists\ C \geq 0$ s.t. $\|QM_{n+1}\| \leq C\left(1 + \|Q(x_n - x_*)\|\right)$ a.s. on $\mathcal{E}(x_*)$,

where $Q := \mathbb{I} - \mathbf{1}^\top \pi$

$\exists$ a non-random positive semi-definite matrix M such that

$$\lim_{n\to\infty} \mathbb{E}(M_{n+1}^\top \pi^\top \pi M_{n+1} \,|\, \mathcal{F}_n) = M \quad \text{a.s. on } \mathcal{E}(x_*)$$

$\exists\ b > 2$ such that $\sup_{n\geq 0} \mathbb{E}(\|\pi M_{n+1}\|^b|\mathcal{F}_n) < \infty$ a.s. on $\mathcal{E}(x_*)$.

Distributed TD(0) with Linear Function Approximation

- Useful for policy evaluation in MARL
- Our result applies since all assumptions hold in this case

Comparison to Existing Literature

- Existing results on convergence rates mainly look at expectation bounds or the CLT. However, these

  - either require the gossip matrix to be doubly stochastic

  - or require stepsizes to be square-summable

  - do not say about the decay rates along different sample paths

Future Directions

- Scaling matrix (i.e., A) in each h_i needs to be the same
- Dynamic communication protocols, i.e., W changes with time
- Two-timescale distributed SA algorithms
- Distributed Q-learning