Finite-Sample Analysis for SARSA with Linear Function Approximation

Shaofeng Zou† Tengyu Xu* Yingbin Liang*

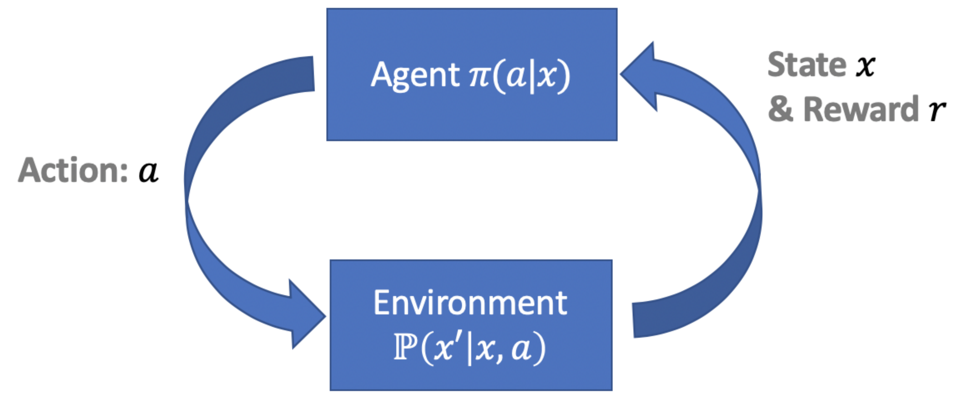
†University at Buffalo, the State University of New York *The Ohio State University

University at Buffalo, The State University of New York

The Ohio State University

Reinforcement Learning

- An agent interacts with a stochastic environment: Markov Decision Process (MDP)
- An MDP: $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$



  - $\mathcal{X}$: state space
  - $\mathcal{A}$: action space
  - P: action dependent transition kernel
    - $\mathbb{P}(X_{t+1} \in U | X_t = x, A_t = a) = \int_U \mathsf{P}(dy|x,a)$
  - $r(X_t, A_t)$: one-stage at time t
  - γ: discount factor
- A policy $\pi(a|x)$ is a conditional distribution over $\mathcal{A}$
- Agent's goal: maximize cumulative discounted reward
  - Value function for policy π: $V^\pi(x_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t)\right]$
  - Action-value function: $Q^\pi(x,a) = r(x,a) + \gamma \int_{\mathcal{X}} \mathsf{P}(dy|x,a) V^\pi(y)$
  - Goal: an optimal policy that maximizes value/action-value function
    $V^*(x) = \sup_\pi V^\pi(x), \forall x \in \mathcal{X}$
    $Q^*(x,a) = \sup_\pi Q^\pi(x,a), \forall (x,a) \in \mathcal{X} \times \mathcal{A}$
- Linear function approximation:

$$Q_\theta(x,a) = \sum_{i=1}^{N} \theta_i \phi_i(x,a) = \phi^\top(x,a)\theta$$

Related Work

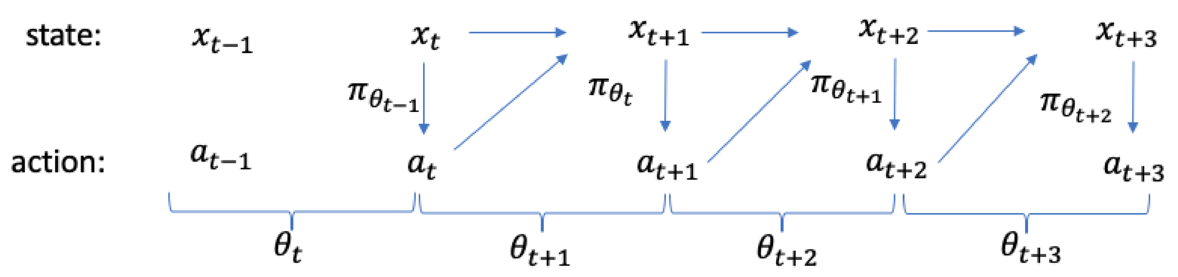
- Q-learning with a single sample path
  - Q-learning with linear function approximation [Melo et al. 2008]
  - K-nearest neighbor Q-learning: non-i.i.d. sample [Shah and Xie 2018]
  - More recent studies: [Chen et al. 2019]
- SARSA with a single sample path
  - Asymptotic analysis: SARSA with linear function approximation [Melo et al. 2008] and [Perkins & Precup 2003], which suggests convergence
- Our study: SARSA with a single sample path
  - Non-asymptotic analysis: How fast the convergence is; and how the convergence rate depends on parameters of RL algorithms and underlying MDP?

SARSA with Linear Function Approx.

Algorithm 1 SARSA

Initialization:
θ_0, x_0, R, ϕ_i, for $i = 1, 2, \ldots, N$
Method:
$\pi_{\theta_0} \leftarrow \Gamma(\phi^T \theta_0)$
Choose a_0 according to π_{θ_0}
for $t = 1, 2, \ldots$ **do**
  Observe x_t and $r(x_{t-1}, a_{t-1})$
  Choose a_t according to $\pi_{\theta_{t-1}}$
  $\theta_t \leftarrow \mathrm{proj}_{2,R}(\theta_{t-1} + \alpha_{t-1} g_{t-1}(\theta_{t-1}))$
  Policy improvement: $\pi_{\theta_t} \leftarrow \Gamma(\phi^T \theta_t)$
end for

- At time t, given (x_t, a_t, x_{t+1})
- Policy: $\pi_{\theta_t} = \Gamma(\phi^\top(x_t, a_t)\theta_t)$, where Γ is a policy improvement operator
- Take action a_{t+1} based on π_{θ_t}
- Updates: $\theta_{t+1} = \theta_t + \alpha_t \mathrm{proj}_{2,R}(g_t(\theta_t))$, where the "gradient" is given by $g_t(\theta_t) = \phi(x_t, a_t)(r(x_t, a_t) + \gamma \phi^T(x_{t+1}, a_{t+1})\theta_t - \phi^T(x_t, a_t)\theta_t)$



- As θ_t is updated, π_{θ_t} changes with time
- On-policy algorithm, changing policy
- Non-i.i.d. data
- Goal: finite-sample analysis for this algorithm

Technical Assumptions

- Lipschitz policy improvement [Perkins & Precup 2003]:

$$|\pi_{\theta_1}(a|x) - \pi_{\theta_2}(a|x)| \le C\|\theta_1 - \theta_2\|_2, \forall (x,a) \in \mathcal{X} \times \mathcal{A}$$

- Smoothness: C is small so that $A_{\theta^*} + C\lambda I$ is negative definite [Melo et al. 2008]
- Uniformly ergodic MDPs: for fixed θ, the Markov chain $\{X_t\}_{t\ge0}$ induced by π_θ and P is uniformly ergodic with invariant measure P_θ, and there are constants $m > 0$ and $\rho \in (0,1)$

$$\sup_{x\in\mathcal{X}} d_{TV}(\mathbb{P}(X_t \in \cdot | X_0 = x), \mathsf{P}_\theta) \le m\rho^t, \forall t \ge 0$$

Definitions:

- $A_\theta = \mathbb{E}_\theta[\phi(X,A)(\gamma\phi^\top(X',A') - \phi^\top(X,A)]$
- $b_\theta = \mathbb{E}_\theta[\phi(X,A)r(X,A)]$
- Limit point θ^* of SARSA satisfies [Melo et al. 2008]: $A_{\theta^*}\theta^* + b_{\theta^*} = 0$

Challenge in Technical Analysis

- Non-i.i.d. samples
  - Complicated coupling between sample path $\{X_t, A_t\}_{t\ge0}$ and $\{\theta_t\}_{t\ge0}$, which introduces bias in g_t
  - Samples are used to compute gradient g_t and θ_{t+1}
  - θ_t is further used (as in policy π_{θ_t}) to generate subsequent actions
- Convergence can be established using O.D.E approach
- For finite time bound, stochastic bias in g_t needs to be explicitly characterized
- Dynamically changing learning policy
  - Analysis in [Bhandari et al. 2018] for TD relies on the fact that the learning policy is fixed so that the Markov process reaches its stationary distribution quickly
  - Episodic SARSA in [Perkins & Precup 2003], within each episode, learning policy is fixed, and the Markov process reach its stationary distribution within each episode
  - No such nice properties for SARSA!

Convergence Results

Theorem 1 *Finite-sample bound on convergence of SARSA with **diminishing** step-size:*

$$\mathbb{E}\|\theta_T - \theta^*\|_2^2 \le c_1 \frac{\log T + 1}{T} + \frac{c_2}{T}.$$

Theorem 2 *Finite-sample bound on convergence of SARSA with **constant** step-size:*

$$\mathbb{E}\|\theta_T - \theta^*\|_2^2 \le c_3 e^{-c_4 T} + c_5 \times \textit{stepsize}.$$

- With constant step-size, SARSA converges faster to a small neighborhood of θ^*.

Proof Sketch

Key idea: design auxiliary uniformly ergodic Markov chain to approximate original Markov chain induced by SARSA

- Step 1. Error decomposition
- Step 2. Gradient descent type analysis
- Step 3. Stochastic bias analysis
- Step 4. Putting the first three steps together and recursively apply step 1 completes the proof

Notations:

- Noiseless gradient at θ: $\bar{g}(\theta) = \mathbb{E}_\theta[g_t(\theta)]$
- Bias by using non-i.i.d. samples to estimate the gradient:

$$\Lambda_t(\theta) = \langle \theta - \theta^*, g_t(\theta) - \bar{g}(\theta) \rangle$$

Proof Sketch

- **Step 1.** Error decomposition

$$\mathbb{E}[\|\theta_{t+1} - \theta^*\|_2^2] \le \underbrace{\mathbb{E}[\|\theta_t - \theta^*\|_2^2] + 2\alpha_t\mathbb{E}[\langle\theta_t - \theta^*, \bar{g}(\theta_t) - \bar{g}(\theta^*)\rangle] + \alpha_t^2\mathbb{E}[\|g_t(\theta_t)\|_2^2]}_{\text{Gradient descent type analysis}} + \underbrace{2\alpha_t\mathbb{E}[\Lambda_t(\theta_t)]}_{\text{Stochastic bias}}$$

- **Step 2.** Gradient descent type analysis because the accurate gradient $\bar{g}_t$ at θ_t is used
  - 2.1 $\|g_t(\theta_t)\|_2$ is upper bounded by G.
  - 2.2 $\mathbb{E}[\langle\theta_t - \theta^*, \bar{g}(\theta_t) - \bar{g}(\theta^*)\rangle] \le -w_s\mathbb{E}[\|\theta_t - \theta^*\|_2^2]$
- **Step 3.** Stochastic bias analysis. $\mathbb{E}[\mathbf{\Lambda}_t(\theta_t)]$ is bias caused by using a single sample path with non-i.i.d. data and time-varying behavior policy

  Rewrite $\mathbf{\Lambda}_t(\theta_t)$ as $\mathbf{\Lambda}_t(\theta_t, O_t)$, where $O_t = (X_t, A_t, X_{t+1}, A_{t+1})$

  Challenge: complicated dependency between θ_t and O_t

  - 3.1 Pre-decoupling dependency between θ_t and O_t by looking τ steps back

$$\mathbf{\Lambda}_t(\theta_t, O_t) \le \mathbf{\Lambda}_t(\theta_{t-\tau}, O_t) + (6 + \lambda C)G^2 \sum_{i=t-\tau}^{t-1} \alpha_i$$

    - If Markov chain induced by SARSA is uniformly ergodic, then given any $\theta_{t-\tau}$, O_t would reach its stationary distribution quickly for large τ
    - This argument is ***not*** necessarily true since policy π_{θ_t} changes with time.
  - 3.2 Decoupling by Auxiliary Markov Chain
    - Key idea: design an auxiliary Markov chain to assist proof
    - Auxiliary Markov chain design:
      (i) Before time $t - \tau + 1$, everything is the same as SARSA
      (ii) After time $t - \tau + 1$, fix behavior policy as $\pi_{\theta_{t-\tau}}$ to generate all subsequent actions
      Denote new observations as $\tilde{O}_t = (\tilde{X}_t, \tilde{A}_t, \tilde{X}_{t+1}, \tilde{A}_{t+1})$
      Since $\pi_{\theta_{t-\tau}}$ is kept fixed, for large τ, $\tilde{O}_t$ reaches stationary distribution induced by policy $\pi_{\theta_{t-\tau}}$ and P
    - $\mathbb{E}[\mathbf{\Lambda}_t(\theta_{t-\tau}, \tilde{O}_t)] \le 4G^2 m\rho^{\tau-1}$
  - 3.3 Stochastic Bias Analysis
    - Bound difference between SARSA Markov chain and auxiliary Markov chain
    - θ_t changes slowly
    - Due to Lipschitz property of $\pi_\theta(a|x)$, the two Markov chain should not deviate from each other too much
    - $\mathbb{E}[\mathbf{\Lambda}_t(\theta_{t-\tau}, O_t)] - \mathbb{E}[\mathbf{\Lambda}_t(\theta_{t-\tau}, \tilde{O}_t)] \le \frac{C|\mathcal{A}|G^3\tau}{w} \log \frac{t}{t-\tau}$

Step 4. Putting the first three steps together and recursively applying Step 1 complete the proof.