Our study: SARSA with a single sample path

- Linear function approximation:
  - Reinforcement Learning
  - Decision Process (MDP)
  - Value function for policy
  - One-stage at time

More recent studies: [Chen et al. 2019]

K-nearest neighbor Q-learning: non-i.i.d. sample [Shah 2008, SAE]

Goal: an optimal policy that maximizes value/action-function

\[ \pi \in \hat{\Gamma} (\theta_0^\gamma) \]

Algorithm 1 SARSA

Initialization:
- \( \pi_0 \in \hat{\Gamma} (\theta_0^\gamma) \)

Method:
- \( \pi_0 \leftarrow \Gamma (\theta_0^\gamma) \)
- Choose \( a_0 \) according to \( \pi_0 \)
- For \( t = 1, 2, \ldots \)
  - Observe \( s_t \) and \( r(s_t, a_t, s_{t+1}) \)
  - Choose \( a_t \) according to \( \pi_{s_t} \)
  - \( \theta_t \leftarrow \theta_{t-1} + \alpha (s_{t-1}, a_{t-1}) \)

Policy improvement: \( \pi_0 \leftarrow \Gamma (\theta_0^\gamma) \)

- At time \( t \), given \( (s_t, a_t, s_{t+1}) \):
- Policy: \( \pi_t = \Gamma (\theta_t, s_t, a_t) \)

- Update: \( \theta_{t+1} = \theta_t + \alpha \frac{1}{N} \sum \phi (x, a) \phi (x, \theta) \)

- As \( \theta_t \) is updated, \( \pi_t \) changes with time

- On-policy algorithm, changing policy
- Non-i.i.d. data
- Goal: finite-sample analysis for this algorithm

Convergence Results

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

\[
E(\theta_t - \theta^*)^2 \leq c_2 \frac{T}{T + 1} + c_3 \frac{2}{T}
\]

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

\[
E(\theta_t - \theta^*)^2 \leq c_2 \frac{t}{T + 1} + c_3 \times \text{stepsize}
\]

Proof Sketch

Step 1. Error decomposition

- \( E[\theta_t - \theta^*]^2 \leq 2 \frac{D}{2} + 2 \frac{|\theta(\theta, \theta^*)|}{2} \)

- \( \frac{D}{2} \) is bias caused by \( \pi_\theta \)

Step 2. Gradient descent type analysis because the accurate gradient \( g_t \) is used

- \( 2 \frac{|\theta(\theta, \theta^*)|}{2} \leq -w_t E(\theta_t - \theta^*)^2 \)

Step 3. Stochastic bias analysis: \( E(\theta_t) \) is biased by using a single sample path with non-i.i.d. data and time-varying behavior policy

Rewrite \( \theta_t = \theta_t(\pi_t, \theta_t) \), where \( \theta_t = (X_t, A_t, X_{t+1}, A_{t+1}) \)

Challenge: complicated dependency between \( \theta_t \) and \( \pi_t \)

- 3.1 Pre-decoupling dependency between \( \theta_t \) and \( \pi_t \) by looking \( T \) steps back

- If Markov chain induced by SARSA is uniformly ergodic, then given any \( \theta_t, \pi_t \), \( \theta_t \) would reach its stationary distribution quickly for large \( \tau \)

- This argument is not necessarily true since policy \( \pi_t \) 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 \( \pi_{\theta_t} \) to generate all subsequent actions
  Denote new observations as \( \theta_t = (X_t, A_t, X_{t+1}, A_{t+1}) \)
  Since \( \pi_{\theta_t} \) is kept fixed, for large \( \tau \), \( \theta_t \) reaches stationary distribution induced by policy \( \pi_{\theta_t} \), and P

- \( E(\theta_t - \theta^*) \leq 4G^2 T \frac{1}{2} \)

- 3.3 Stochastic Bias Analysis

- Bound difference between SARSA Markov chain and auxiliary Markov chain

- \( \theta_t \) changes slowly

- Due to Lipschitz property of \( \pi_\theta \), the two Markov chains should not deviate from each other too much

- \( E(\theta_t - \theta^*) \leq \frac{4G^2}{2} \frac{1}{T \log} \)

Proof Sketch

Step 4. Putting the first three steps together and recursively apply step 1 completes the proof