Finite-Time Bounds for Two-Time-Scale Stochastic Approximation with Arbitrary Norm Contractions and Markovian Noise

Siddharth Chandak

Joint work with Shaan Ul Haque (Georgia Institute of Technology) and Prof. Nicholas Bambos (Stanford University)

Outline

- Average Cost Reinforcement Learning
  - SSP Q-Learning Algorithm
- Two-time-scale Stochastic Approximation
  - Arbitrary Norm Contractions and Markov noise
- Results
  - Proof Technique

Average Cost Reinforcement Learning

Objective

- We wish to **minimize the average cost** of an MDP
- Choose actions $\{A_m\}$ such that the following cost is minimized

$$\limsup_{n\uparrow\infty} \frac{1}{n} \sum_{m=0}^{n-1} \mathbb{E}[c(S_m, A_m)]$$

  - Cost function: $c(s, a)$
  - Controlled Markov chain: $\{S_m\}$ in finite state space $\mathcal{S}$
- Interested in stationary policies

Discounted vs Average Cost

- Q-values for discounted case:

$$Q(s,a) = c(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \min_{a'} Q(s',a')$$

  - γ is the discount factor

- Q-values for average reward case:

$$Q(s,a) = c(s,a) - \rho + \sum_{s' \in \mathcal{S}} p(s'|s,a) \min_{a'} Q(s',a')$$

  - ρ is the optimal average cost

Why is average cost harder?

$$Q(s,a) = c(s,a) - \rho + \sum_{s' \in \mathcal{S}} p(s'|s,a) \min_{a'} Q(s',a')$$

- Lack of discount factor
    - Harder to obtain a contraction
- Estimating and handling the term ρ (optimal average cost)

A bit of history

- First asymptotic convergence of average cost RL studied by [Abounadi et al. (2001)[1]] for two algorithms:
  - RVI Q-Learning
  - SSP Q-Learning
- RVI Q-Learning is...
  - ...much more popular
  - ...much harder to obtain finite-time performance bounds for

[1] J. Abounadi, D. Bertsekas, and V. S. Borkar, "Learning Algorithms for Markov Decision Processes with Average Cost", (2001)

Intuition behind SSP Q-Learning

- Adaptation of the algorithm for stochastic shortest path problem
- Reference state $s_0 \in \mathcal{S}$ *(intuition: terminal state)*
- Recall Q-values for average cost MDPs:

$$Q(s,a) = c(s,a) - \rho + \sum_{s' \in \mathcal{S}} p(s'|s,a) \min_{a'} Q(s',a')$$

- Q-values for SSP:

$$Q^{SSP}(s,a) = c(s,a) - \rho + \sum_{s' \neq s_0} p(s'|s,a) \min_{a'} Q^{SSP}(s',a')$$

- **Equivalent up to additive constants**

SSP Q-Learning Algorithm

$$Q_{n+1}(s,a) = Q_n(s,a) + \alpha_n \mathbf{1}_{\{S_n=s,A_n=a\}}\Big(c(s,a)$$
$$+ \mathbf{1}_{\{S_{n+1}\neq s_0\}} \min_{a'} Q_{n+1}(S_{n+1},a) - Q_n(s,a) - \rho_n\Big)$$
$$\rho_{n+1} = \rho_n + \beta_n(\min_{a'} Q_n(s_0,a'))$$

- α_n, β_n: Stepsizes, gives rise to two-time-scale structure
  - Q-updates: faster time-scale - α_n is larger
  - Updates for average cost estimate: slower time-scale - β_n is smaller
- $\mathbf{1}_{\{S_n=s,A_n=a\}}$: Asynchronous updates
- $\mathbf{1}_{\{S_{n+1}\neq s_0\}}$: 'Terminal' state

SSP Q-Learning Algorithm

- **A two-time-scale algorithm**
- Faster time-scale can be written as fixed point iteration with contraction under max-weighted norm
    - **Arbitrary norm contractions**
- Asynchronous updates lead to **Markovian noise**
- **This Work:** *"Finite-Time Bounds for Two-Time-Scale Stochastic Approximation with Arbitrary Norm Contractions and Markovian Noise"*
    - Prior works focused on Euclidean norm

Key Result

First $O(1/n)$ mean square error bound on an algorithm for asynchronous control for average cost MDPs.

Two-time-scale Stochastic Approximation

Two-Time-Scale Iterations

- Coupled iterations updating on separate time-scales

$$x_{n+1} = x_n + \alpha_n(f(x_n, y_n, Z_n) - x_n + M_{n+1})$$
$$y_{n+1} = y_n + \beta_n(g(x_n, y_n, Z_n) - y_n + M'_{n+1}).$$

- Timescales dictated by the different stepsizes α_n and β_n
- Z_n is irreducible Markov chain with stationary distribution $\pi(\cdot)$ in finite state space $\mathcal{S}$
    - Define stationary averages $\overline{f}(x, y) = \sum_{s \in \mathcal{S}} \pi(s) f(x, y, s)$ and $\overline{g}(x, y)$
- Want to solve $\overline{f}(x, y) = x$ and $\overline{g}(x, y) = y$ given noisy realizations

Two-time-scale Iterations

$$x_{n+1} = x_n + \alpha_n(\overline{f}(x_n, y_n) - x_n + \omega_n + M_{n+1})$$
$$y_{n+1} = y_n + \beta_n(\overline{g}(x_n, y_n) - y_n + \omega'_n + M'_{n+1}).$$

- M_{n+1} and M'_{n+1} are martingale difference noise sequences arising from noisy observations
- $\omega_n = f(x_n, y_n, Z_n) - \overline{f}(x_n, y_n)$ and ω'_n are the Markov noise

Two-Time-Scale Iterations

$$\text{Faster: } x_{n+1} = x_n + \alpha_n(\overline{f}(x_n, y_n) - x_n + \omega_n + M_{n+1})$$
$$\text{Slower: } y_{n+1} = y_n + \beta_n(\overline{g}(x_n, y_n) - y_n + \omega_n + M'_{n+1})$$

- α_n is larger, or decays at a slower rate, e.g., $1/n^{0.6}$
- β_n is smaller, or decays at a faster rate, e.g., $1/n^{0.75}$
- Analysis
  - Faster time-scale: y_n considered quasi-static
  - Slower time-scale: x_n tracks $x^*(y_n)$, the fixed point for $\overline{f}(\cdot, y_n)$

Key Contractive Assumptions

- There exists $0 \leq \lambda < 1$ such that,

$$\|\overline{f}(x_1, y) - \overline{f}(x_2, y)\| \leq \lambda \|x_1 - x_2\|$$

  for all x_1, x_2, y

  - Unique fixed point $x^*(y)$ for each y, such that $\overline{f}(x^*(y), y) = x^*(y)$

- There exists $0 \leq \mu < 1$ such that

$$\|\overline{g}(x^*(y_1), y_1) - \overline{g}(x^*(y_2), y_2)\| \leq \mu \|y_1 - y_2\|$$

  for all y_1, y_2

  - Unique fixed point y^* such that $\overline{g}(x^*(y^*), y^*) = y^*$

- $\| \cdot \|$ **is any arbitrary norm**

Results

Mean Square Error bound

Theorem

For $\alpha_n = \Theta(1/n^{2/3})$ and $\beta_n = \Theta(1/n)$,

$$\mathbb{E}\left[\|x_n - x^*(y_n)\|^2 + \|y_n - y^*\|^2\right] = \mathcal{O}(1/n^{2/3}).$$

An Important Special Case

- SSP Q-Learning can be expressed in the following form:

$$x_{n+1} = x_n + \alpha_n(\overline{f}(x_n, y_n) - x_n + \omega_n + M_{n+1})$$
$$y_{n+1} = y_n + \beta_n(\overline{g}(x_n, y_n) - y_n).$$

- **The slower time-scale is noiseless**: no Markovian or martingale noise

Bound for special case

Theorem

For $\alpha_n = \Theta(1/n)$, $\beta_n = \Theta(1/n)$, and sufficiently small β_n/α_n,

$$\mathbb{E}\left[\|x_n - x^*(y_n)\|^2 + \|y_n - y^*\|^2\right] = \mathcal{O}(1/n),$$

when the slower time-scale is noiseless.

Tools used for Proof

- **Moreau Envelopes:** To deal with arbitrary norm contractions
  - Helps define a smooth Lyapunov function
- **Solutions of Poisson equation:** To deal with Markov noise
  - Decompose Markov noise into martingale difference sequence and an additional telescopic series

Conclusions

Conclusions

- Analyzed two-time-scale SA
- Obtained the first $O(1/n)$ bound for control of average cost MDPs
- Other applications include Q-Learning with Polyak averaging

Future Directions:

- Recent work obtained $O(1/n)$ bound for the general case (both time-scales are noisy) for the Euclidean norm [Chandak (2025)[2]]
  - Can be extended to the setting with arbitrary norm contractions

[2] Chandak, Siddharth. "$O(1/k)$ Finite-Time Bound for Non-Linear Two-Time-Scale Stochastic Approximation." *arXiv:2504.19375 (2025)*.

Thank You!