University at Buffalo
The State University of New York

Finite-sample Analysis of Greedy-GQ with Linear Function Approximation under Markovian Noise

Yue Wang, Shaofeng Zou

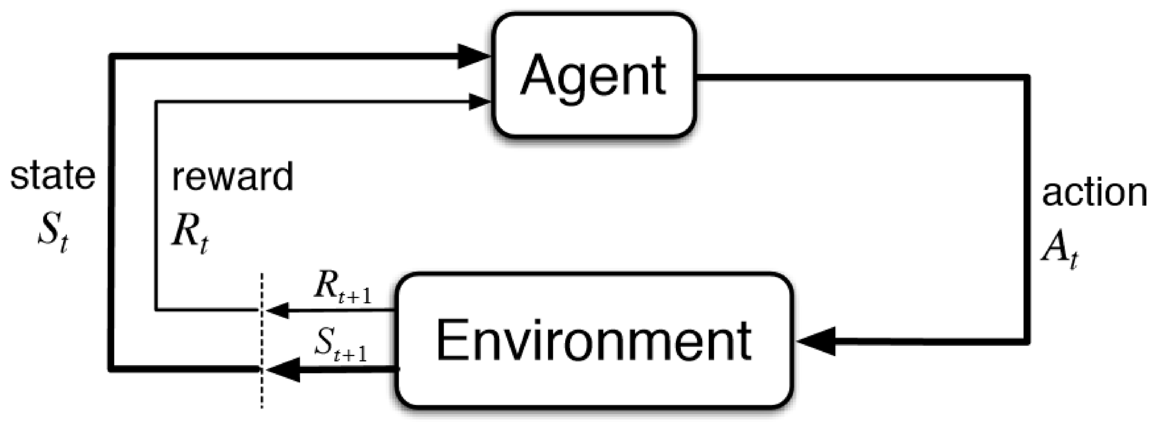
Introduction

Greedy-GQ is an off-policy two timescale algorithm for optimal control in reinforcement learning. This paper develops the first finite-sample analysis for the Greedy-GQ algorithm with linear function approximation under Markovian noise.

Keywords: Greedy-GQ, Off-policy control, non-convex optimization.

Reinforcement Learning

- An agent interacts with a stochastic environment: Markov Decision Process (MDP)
  - $\mathcal{S}$: states space
  - $\mathcal{A}$: action set
  - $\mathcal{P}$: transition kernel ($P^a_{ss'} = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$)
  - r: reward function
  - γ: discount factor
- Agent's goal: maximize cumulative discounted reward
  - Value function of a policy π: $V^\pi(s) = \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(S_t, A_t) | S_0 = s]$
  - Action-value function: $Q^\pi(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} \mathbb{P}(dx|s,a) V^\pi(x)$
  - Goal: an optimal policy that maximizes value/action value function: $Q^*(s,a) = \sup_\pi Q^\pi(s,a)$
- Linear function approximation: A set of fixed independent base functions $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$, $Q_\theta(s,a) = \phi(s,a)^\top \theta$



Greedy-GQ

Algorithm 1 Greedy-GQ [18]

Initialization:
$\theta_0, \omega_0, s_0, \phi^{(i)}$, for $i = 1, 2, ..., N$
Method:
$\pi_{\theta_0} \leftarrow \Gamma(\phi^\top \theta_0)$
for $t = 0, 1, 2, ...$ **do**
  Choose a_t according to $\pi_b(\cdot|s_t)$
  Observe s_{t+1} and r_t
  $\bar{V}_{s_{t+1}}(\theta_t) \leftarrow \sum_{a' \in \mathcal{A}} \pi_{\theta_t}(a'|s_{t+1}) \theta_t^\top \phi_{s_{t+1},a'}$
  $\delta_{t+1}(\theta_t) \leftarrow r_t + \gamma \bar{V}_{s_{t+1}}(\theta_t) - \theta_t^\top \phi_t$
  $\hat{\phi}_{t+1}(\theta_t) \leftarrow$ gradient of $\bar{V}_{s_{t+1}}(\theta_t)$
  $\theta_{t+1} \leftarrow \theta_t + \alpha_t(\delta_{t+1}(\theta_t)\phi_t - \gamma(\omega_t^\top \phi_t)\hat{\phi}_{t+1}(\theta_t))$
  $\omega_{t+1} \leftarrow \omega_t + \beta_t(\delta_{t+1}(\theta_t) - \phi_t^\top \omega_t)\phi_t$
  Policy improvement: $\pi_{\theta_{t+1}} \leftarrow \Gamma(\phi^\top \theta_{t+1})$
end for

- At time t, given s_t
- Policy: $\pi_{\theta_t} = \Gamma(\phi^\top \theta_t)$, where Γ is a policy improvement operator
- Take action a_t based on π_{θ_t}, observe s_{t+1} and r_{t+1}
- Updates:
  $\theta_{t+1} \leftarrow \theta_t + \alpha_t(\delta_{t+1}(\theta_t)\phi_t - \gamma(\omega_t^\top \phi_t)\hat{\phi}_{t+1}(\theta_t))$
  and
  $\omega_{t+1} \leftarrow \omega_t + \beta_t(\delta_{t+1}(\theta_t) - \phi_t^\top \omega_t)\phi_t$

Technical Assumptions

- The matrix $C = \mathbb{E}_\mu[\phi_t \phi_t^\top]$ is non-singular.
- $\|\phi_{s,a}\|_2 \leq 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
- There exists some constants $m > 0$ and $\rho \in (0,1)$ such that

  $\sup_{s \in \mathcal{S}} d_{TV}(\mathbb{P}(s_t | s_0 = s), \mu) \leq m\rho^t$,

  for any $t > 0$, where d_{TV} is the total-variation distance between the probability measures.
- The policy $\pi_\theta(a|s)$ is k_1-Lipschitz and k_2-smooth, i.e., for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

  $\|\nabla \pi_\theta(a|s)\| \leq k_1, \forall \theta$, and,

  $\|\nabla \pi_{\theta_1}(a|s) - \nabla \pi_{\theta_2}(a|s)\| \leq k_2 \|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2$

Results

Consider the following step-sizes: $\beta = \beta_t = \frac{1}{T^b}$, and $\alpha = \alpha_t = \frac{1}{T^a}$, where $\frac{1}{2} < a \leq 1$ and $0 < b \leq a$. Then we have that for $T \geq 1$,

$$\mathbb{E}[\|\nabla J(\theta_M)\|^2] = \mathcal{O}\left(\frac{1}{T^{1-a}} + \frac{\log T}{T^{\min\{b, a-b\}}}\right).$$

If we choose $a = \frac{2}{3}$ and $b = \frac{1}{3}$, then the best rate of the bound is obtained as follows:

$$\mathbb{E}[\|\nabla J(\theta_M)\|^2] = \mathcal{O}\left(\frac{\log T}{T^{\frac{1}{3}}}\right).$$

Proof Sketch

Step 1: Decompose the error recursively into two parts:

$$\mathbb{E}[\|\nabla J(\theta_M)\|^2] \leq \frac{1}{\sum_{t=0}^T \alpha_t}\Bigg(\underbrace{(J(\theta_0) - J(\theta_{T+1})) + \frac{K}{2}\sum_{t=0}^T \alpha_t^2 \mathbb{E}[\|G_{t+1}(\theta_t, \omega_t)\|^2]}_{\text{classical non-convex type analysis}} \underbrace{- \sum_{t=0}^T \frac{\alpha_t}{2} \langle \Delta_t, \nabla J(\theta_t)\rangle}_{\text{stochastic bias } (*)}\Bigg)$$

The first part is handled in many classical non-convex problems. To bound the second part stochastic bias, first bound $\|\nabla J(\theta)\|$ in stochastic bias by a constant

Step 2: Decompose stochastic bias $(*)$ into two parts: bias due to Markov noise and tracking error:

$(*) = \underbrace{\langle \nabla J(\theta_t), -2G_{t+1}(\theta_t, \omega_t) + 2G_{t+1}(\theta_t, \omega^*(\theta_t))\rangle}_{\text{tracking error}} - \underbrace{\langle \nabla J(\theta_t), \nabla J(\theta_t) + 2G_{t+1}(\theta_t, \omega^*(\theta_t))\rangle}_{\text{bias due to Markov noise } (\triangleq \zeta(\theta_t, O_t))}$ where $O_t = (S_t, A_t, R_t, S_{t+1})$

The challenge of bounding lies in that θ_t and O_t are dependent.

Step 3: Bound bias using uniform ergodicity of underlaying MDP:

Decouple the independence of θ_t and O_t by considering τ steps back: $|\zeta(\theta_t, O_t) - \zeta(\theta_{t-\tau}, O_t)| \leq \mathcal{O}\left(\sum_{k=t-\tau}^{t-1} \alpha_k\right)$

Define independent R.V. $\hat{O} = (\hat{S}, \hat{A}, \hat{R}, \hat{S}') \sim \mu \times \mathcal{P}$, then: $\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] \leq |\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] - \underbrace{\mathbb{E}[\zeta(\theta_{t-\tau}, \hat{O})]}_{=0}| \leq k_\zeta m \rho^\tau$

Combine two inequalities above to bound bias term $\zeta(\theta_t, O_t)$.

Step 4: Bound tracking error:

Rewrite tracking error recursively: $||z_{t+1}||^2 \leq ||z_t||^2 + 2\beta_t\langle z_t, f_2(\theta_t, O_t)\rangle + 2\beta_t\langle z_t, \bar{g}_2(z_t)\rangle + 2\langle z_t, \omega^*(\theta_t) - \omega^*(\theta_{t+1})\rangle + 2\beta_t\langle z_t, g_2(z_t, O_t) - \bar{g}_2(z_t)\rangle + \mathcal{O}(\beta_t^2 + \alpha_t^2)$

Bound terms above using methods similar to those in step 3

Step 5: Plug the bound of $\|\nabla J(\theta)\|$ into stochastic bias $(*)$ in step 1: $-\sum_{t=0}^T \frac{\alpha_t}{2}\langle \Delta_t, \nabla J(\theta_t)\rangle$ This can improve rate by a tighter bound of $\|\nabla J(\theta)\|$

Recursively apply step 1 to 4 until it converges