Trust Region Policy Optimization

Wenhao Yang

School of Mathematical Sciences
Peking University

May 4, 2017

Outline

Markov Decision Process

- Dynamic system: $(\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$
- $\rho_0 : \mathcal{S} \to \mathbb{R}$
- $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a stochastic policy
- $\eta(\pi)$ is expected discounted reward:

$$\eta(\pi) = E_{s_0, a_0, \ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t)] \tag{1}$$

where $s_0 \sim \rho_0(s_0)$, $a_t \sim \pi(a_t|s_t)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$

Actor-Critic Decomposition

- $Q_\pi(s_t, a_t) = E_{s_{t+1}, a_{t+1}, \ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$
- $V_\pi(s_t) = E_{a_t, s_{t+1} \ldots}[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})]$
- $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$
- $a_t \sim \pi(a_t|s_t)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$

Theorem

Given another policy $\tilde{\pi}$, the following equation holds:

$$\eta(\tilde{\pi}) = \eta(\pi) + E_{s_0, a_0, \ldots \sim \tilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{2}$$

- Denote $\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \ldots$ we have:

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a) \tag{3}$$

Actor-Critic Decomposition

- The computation of $\rho_{\tilde{\pi}}$ is complex, which makes the equation(2) difficult to optimize directly.
- Approximation:

$$L_{\pi}(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\pi}(s) \sum_{a} \tilde{\pi}(a|s) A_{\pi}(s,a) \tag{4}$$

- If π_{θ} is a differential funtion of θ, L_{π} matches η to first order

$$L_{\pi_{\theta_0}}(\pi_{\theta_0}) = \eta \pi_{\theta_0} \tag{5}$$

$$\nabla L_{\pi_{\theta_0}}(\pi_{\theta})_{|\theta=\theta_0} = \nabla \eta(\pi_{\theta})|_{\theta=\theta_0} \tag{6}$$

- Conservative policy iteration provide explicit lower bounds on the improvement of η.

Conservative policy iteration

- π_{old} : the current policy
- $\pi' = \arg\max_{\pi'} L_{\pi_{old}}(\pi')$
- The new policy π_{new} is defined as:

$$\pi_{new}(a|s) = (1-\alpha)\pi_{old}(a|s) + \alpha\pi'(a|s) \tag{7}$$

- Lower bound:

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{2\varepsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{8}$$

  where $\varepsilon = \max_s |E_{a\sim\pi'(a|s)}[A_\pi(s,a)]|$
- This lower bound if unwieldy and restrictive in practice.

Improvement

- The bound can be extended to general stochastic policies by replacing α with a distance measure between π and $\tilde{\pi}$ and changing the constant ε appropriately.
- Distance measure: $D_{TV}(p \parallel q) = \frac{1}{2}\sum_i |p_i - q_i|$
- $D_{TV}^{\max}(\pi, \tilde{\pi}) = \max_s D_{TV}(\pi(\cdot|s) \parallel \tilde{\pi}(\cdot|s))$

Theorem

Let $\alpha = D_{TV}^{\max}(\pi_{old}, \pi_{new})$, *then the following bound holds:*

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2 \tag{9}$$

where $\varepsilon = \max_{s,a} |A_\pi(s,a)|$

Improvement

- Note that $D_{TV}(p \parallel q)^2 \leq D_{KL}(p \parallel q)$, then the bound becomes as:

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{4\varepsilon\gamma}{(1-\gamma)^2} D_{KL}^{\max}(\pi_{new}, \pi_{old}) \quad (10)$$

- Algorithm:

Algorithm 1 Policy iteration algorithm guaranteeing non-decreasing expected return η

Initialize π_0.
for $i = 0, 1, 2, \ldots$ until convergence **do**
Compute all advantage values $A_{\pi_i}(s, a)$.
Solve the constrained optimization problem

$$\pi_{i+1} = \arg\max_{\pi} \left[L_{\pi_i}(\pi) - C D_{\mathrm{KL}}^{\max}(\pi_i, \pi) \right]$$

where $C = 4\epsilon\gamma/(1-\gamma)^2$

and $L_{\pi_i}(\pi) = \eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s, a)$

end for

Improvement

- This algorithm is guaranteed to generate a monotonically improving sequence of policies $\eta(\pi_0) \leq \eta(\pi_1) \leq \eta(\pi_2) \leq ...$
- In fact, let $M_i(\pi) = L_{\pi_i}(\pi) - CD_{KL}^{\max}(\pi_i, \pi)$, then:

$$\eta(\pi_{i+1}) \geq M_i(\pi_{i+1}) \tag{11}$$
$$\eta(\pi_i) = M_i(\pi_i) \tag{12}$$
$$\eta(\pi_{i+1}) - \eta(\pi_i) \geq M_i(\pi_{i+1}) - M_i(\pi_i) \tag{13}$$

- Trust Region Policy Optimization is an approximation to this Algorithm.

Trust Region Policy Optimization

- Consider $\pi_\theta(a|s)$ as a function of parameter θ, the algorithm is aim to optimize:

$$\max_\theta L_{\theta_{old}}(\theta) - CD_{KL}^{\max}(\theta_{old}, \theta) \tag{14}$$

- Problem: Step size would be very small.
- To take larger steps in a robust way:

$$\max_\theta L_{\theta_{old}}(\theta) \tag{15}$$

$$s.t.\ D_{KL}^{\max}(\theta_{old}, \theta) \leq \delta \tag{16}$$

- Problem: Too many constrains.
- Do a heuristic approximation which considers the average KL divergence:

$$\bar{D}_{KL}^{\rho_{old}}(\theta_{old}, \theta) = E_{s\sim\rho_0}[D_{KL}(\pi_{\theta_{old}}(\cdot|s) \parallel \pi_\theta(\cdot|s))] \leq \delta \tag{17}$$

Approximation by MC

- Optimization problem:

$$\max_{\theta} \sum_{s} \rho_{\theta_{old}}(s) \sum_{a} \pi_{\theta}(a|s) A_{\theta_{old}}(s,a) \tag{18}$$

$$s.t.\ \bar{D}_{KL}^{\rho_{old}}(\theta_{old}, \theta) \leq \delta \tag{19}$$

- Sample formation:

$$\max_{\theta} E_{s\sim\rho_{old}, a\sim q}[\frac{\pi_{\theta}(a|s)}{q(a|s)} Q_{\theta_{old}}(s,a)] \tag{20}$$

$$s.t.\ \bar{D}_{KL}^{\rho_{old}}(\theta_{old}, \theta) \leq \delta \tag{21}$$

- How to sample: 1. Single Path; 2. Vine
- Single Path is just sampling $s_0 \sim \rho_0$ and simulate a sequence by $\pi_{\theta_{old}}$. Hence, $q = \pi_{\theta_{old}}$, $Q_{\theta_{old}}(s,a)$ is computed at each step.

Vine

- Like A3C. Explaination on black-board
- In small, finite action spaces, every possible action can be generated from a given state:

$$L_n(\theta) = \sum_{k=1}^{K} \pi_\theta(a_k|s_n)\hat{Q}(s_n, a_k) \tag{22}$$

where action space is $\mathcal{A} = \{a_1, a_2, ..., a_K\}$

- In large or continuous action spaces:

$$L_n(\theta) = \frac{\sum_{k=1}^{K} \frac{\pi_\theta(a_{n,k}|s_n)}{\pi_{\theta_{old}}(a_{n,k}|s_n)}\hat{Q}(s_n, a_{n,k})}{\sum_{k=1}^{K} \frac{\pi_\theta(a_{n,k}|s_n)}{\pi_{\theta_{old}}(a_{n,k}|s_n)}} \tag{23}$$

Practical Algorithm

- 1. Use Single Path or Vine to collect a set of state-action pairs along with MC estimates of their Q-Values
- 2. By averaging over samples, construct object function and constraint region
- 3. Solve the optimization problem to update the policy's parameter.

Reference

Reference:
[1] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, Pieter Abbeel; Trust Region Policy Optimization