Trust Region Policy Optimization

HU Yifan

Department of Industrial and System Engineering
University of Illinois at Urbana-Champaign

IE 598, Spring 2018
Outline

- Infinite-horizon discounted MDP
- Monotonic Improvement Guarantee for General Stochastic Policies
- Approximations
- Simple path and Vine Sampling
- Algorithm and Performance
- Summary
Infinite-horizon discounted MDP

- Infinite-horizon discounted Markov Decision Process (MDP)
- \((S, A, P, r, \rho_0, \gamma)\)
- \(S\) finite set of states, \(A\) finite set of actions, \(P : S \times A \times S \to \mathbb{R}\) transition probability distribution
- \(r : S \times A \to R\) the reward function, \(\rho_0\) the distribution of the initial state \(s_0\), \(\gamma\) discounted factor.
- \(\pi\) stochastic policy : \(S \times A \to [0, 1]\), \(\eta(\pi)\) the expected discounted reward:

\[
\eta(\pi) = \mathbb{E}_{s_0, a_0, s_1, a_1 \ldots} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t) \right]
\]

- state-action value function \(Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \ldots} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t) \right]\)
- state value function \(V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, a_{t+1}, \ldots} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t) \right]\)
- advantage function \(A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)\)
Some intuition

- $\tau := (s_0, a_0, s_1, a_1 \ldots)$

- Simple calculation: $\eta(\pi_{new}) = \eta(\pi) + \mathbb{E}_{\tau \sim \pi_{new}} \left[ \sum_{t=0}^{\infty} \gamma^t A_\pi(s, a) \right]$

- Define $\tilde{A}(s) = \mathbb{E}_{a \sim \pi_{new}(\cdot|s)}[A_\pi(s, a)]$

- $\eta(\pi_{new}) = \eta(\pi) + \mathbb{E}_{\tau \sim \pi_{new}} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{A}(s_t) \right]$

- Approximation of $\eta(\pi_{new})$ as $L_\pi(\pi_{new}) := \eta(\pi) + \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t \tilde{A}(s_t) \right]$

- $\alpha$-coupled policy pair $(\pi, \pi_{new})$: if $(\pi, \pi_{new})$ defines a joint distribution $(a, a_{new})|s$, such that $\mathbb{P}(a \neq a_{new}) \leq \alpha$ for all $s$, $\pi$ and $\pi_{new}$ denote the marginal distribution of $a$ and $a_{new}$
Given that $\pi$ and $\pi_{\text{new}}$ are $\alpha$-coupled policies, for all $s$, we have:

$$|\tilde{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)| \quad (1)$$

$$|E_{s_t \sim \pi_{\text{new}}} [\tilde{A}(s_t)] - E_{s_t \sim \pi} [\tilde{A}(s_t)]| \leq 4\alpha (1 - (1 - \alpha)^t) \max_s |A_\pi(s,a)| \quad (2)$$

The proof is easy. Simply divide the LHS into two part by $P(n_t = 0)$ and $P(n_t > 0)$ where $n_t$ is the number of times that $a_i \neq a_{\text{inew}}$ for $i < t$. 
Let $\alpha = D_{TV}^{\text{max}}(\pi_{old}, \pi_{new})$, then

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{4\epsilon\gamma}{(1 - \gamma)^2} \alpha^2, \text{ where, } \epsilon = \max_{s,a} |A_\pi(s, a)| \quad (3)$$

The proof comes directly from the lemma stated above. Note that here $D_{TV}^{\text{max}}(\pi_{old}, \pi_{new}) = \max_s D_{TV}(\pi_{old}, \pi_{new})$ total variation divergence. Noticing that $D_{TV}(\pi_{old}, \pi_{new})^2 \leq D_{KL}(\pi_{old}, \pi_{new})$, we have

$$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - CD_{KL}^{\text{max}}(\pi_{old}, \pi_{new}), \text{ where, } C = \frac{4\epsilon\gamma}{(1 - \gamma)^2} \quad (4)$$
Consider parameterized policies $\pi_\xi(a|s)$, denote $\xi_n$ for $\xi_{new}$

$$\max_\xi L_{\xi_{old}}(\xi) - C D_{KL}^{\max}(\pi_{\xi_{old}}, \pi_{\xi_{new}})$$

In practice if we use the penalty coefficient $C$ recommended by the theory above, the step sizes would be very small. Thus we use a constraint on KL divergence, i.e. trust region constraint

$$\max L_{\pi_{\xi_{old}}}(\pi_\xi)$$ (5)

subject to $D_{KL}^{\max}(\pi_{\xi_{old}}, \pi_{\xi_{new}}) \leq \delta$

This new problem has a constraint that the KL divergence is bounded at every point in the state space. However, still difficult to solve large number of constraints
Thus we use a heuristic approximation which consider the average KL divergence:

$$\max_{\xi} L_{\pi_{\xi_{\text{old}}}}(\pi_{\xi})$$  \hspace{1cm} (6)

subject to \(\mathbb{E}_s D_{KL}(\pi_{\xi_{\text{old}}}, \pi_{\xi_{\text{new}}}) \leq \delta\)

it is equal to

$$\max_{\xi} \sum_s \rho_{\xi_{\text{old}}} \sum_a \pi_{\xi}(a|s) A_{\xi_{\text{old}}}(s, a)$$ \hspace{1cm} (7)

subject to \(\mathbb{E}_s D_{KL}(\pi_{\xi_{\text{old}}}, \pi_{\xi_{\text{new}}}) \leq \delta\)
Sample-Base Estimation of the Objective and constraint

To solve the above optimization problem is the same to solve

\[
\max_{\xi} \mathbb{E}_{s \sim \rho_{\xi_{\text{old}}}, a \sim q} \left[ \frac{\pi_{\xi}(a|s)}{q(a|s)} Q_{\xi_{\text{old}}}(s, a) \right]
\]  

subject to \( \mathbb{E}_{s} D_{KL}(\pi_{\xi_{\text{old}}}, \pi_{\xi_{\text{new}}}) \leq \delta \)

Note that here \( q \) is the sampling distribution, from \( A \) to \( Q \) which only changes the objective by a constant. All that remains is to replace the expectations by sample averages and replace the \( Q \) value by an empirical estimate.
Single Path: collect a sequence of states by sampling $s_0 \sim \rho_0$ and then simulating the policy $\pi_{\xi_{old}}$ for $T$ timesteps to generate $s_0, a_0, s_1...a_{T_1}, s_T$. Then $q(a|s) = \pi_{\xi_{old}}(a|s)$. $Q_{\xi_{old}}(s, a)$ is computed at each state-action pair $(s_t, a_t)$.

Vine: generate the trajectory by current policy, pick a subset of $N$ states as rollout and at each state, sample $K$ action according to $q(\cdot|s_n)$. At each state we estimate $\hat{Q}_\xi(s_n, a_{n,k})$ by performing a rollout starting with state $s_n$ and action $a_{n,k}$.
Use the single path or vine procedures to collect a set of state-action pairs along with Monte Carlo estimates of their Q-values.

By averaging over samples, construct the estimated objective and constraint.

Approximately solve this constrained optimization problem to update the policy's parameter vector $\xi$. We use the conjugate gradient algorithm followed by a line search, which is altogether only slightly more expensive than computing the gradient itself.
Performance on locomotion

Cartpole

Swimmer

Hopper

Walker

number of policy iterations

number of policy iterations

reward

reward

reward

reward

Vine
Single Path
Natural Gradient
Max KL
Empirical FIM
CEM
CMA
RWR

Vine
Single Path
Natural Gradient
Empirical FIM
CEM
CMA
RWR
Performance

<table>
<thead>
<tr>
<th></th>
<th>B. Rider</th>
<th>Breakout</th>
<th>Enduro</th>
<th>Pong</th>
<th>Q*bert</th>
<th>Seaquest</th>
<th>S. Invaders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>354</td>
<td>1.2</td>
<td>0</td>
<td>-20.4</td>
<td>157</td>
<td>110</td>
<td>179</td>
</tr>
<tr>
<td>Human (Mnih et al. 2013)</td>
<td>7456</td>
<td>31.0</td>
<td>368</td>
<td>-3.0</td>
<td>18900</td>
<td>28010</td>
<td>3690</td>
</tr>
<tr>
<td>Deep Q Learning (Mnih et al. 2013)</td>
<td>4092</td>
<td>168.0</td>
<td>470</td>
<td>20.0</td>
<td>1952</td>
<td>1705</td>
<td>581</td>
</tr>
<tr>
<td>UCC-I (Guo et al. 2014)</td>
<td>5702</td>
<td>380</td>
<td>741</td>
<td>21</td>
<td>20025</td>
<td>2995</td>
<td>692</td>
</tr>
<tr>
<td>TRPO - single path</td>
<td>1425.2</td>
<td>10.8</td>
<td>534.6</td>
<td>20.9</td>
<td>1973.5</td>
<td>1908.6</td>
<td>568.4</td>
</tr>
<tr>
<td>TRPO - vine</td>
<td>859.5</td>
<td>34.2</td>
<td>430.8</td>
<td>20.9</td>
<td>7732.5</td>
<td>788.4</td>
<td>450.2</td>
</tr>
</tbody>
</table>

Performance comparison for vision-based RL algorithms on the Atari domain. Not best in all however can be very general.
We prove that minimizing a certain surrogate objective function guarantees policy improvement with non-trivial step sizes.

We make a series of approximations to the theoretically-justified algorithm, yielding a practical algorithm, which we call trust region policy optimization (TRPO)

We describe two variants of this algorithm: the single-path method, and the vine method.

In our experiments, we show that the same TRPO methods can learn complex policies for swimming, hopping, and walking, as well as playing Atari games directly from raw images.