ISPR Mid Term 4:
Trust Region Policy Optimization

Luca Moroni
Mat: 635966

ISPR

2022

Introduction

In a Reinforcement Learning Scenario the policy optimization methods are those technique used to optimize the policy, or rather the probability distribution over possible actions given a state ($\pi(a, s)$).

The work described in [1] introduce a method for policy optimization (learns a parametrized policy), defining a policy gradient method similar to the natural policy gradient one.

The **Natural policy gradient methods** want to change the parameters of the policy ($\theta_{old} \rightarrow \theta_{new}$) without changing to much the relative distributions $\pi_{\theta_{old}}$, $\pi_{\theta_{new}}$.

Model Description (I)

Let π a policy, the following identity is an approximation (simpler to compute w.r.t the exact one) of the expected discounted reward of another policy $\widetilde{\pi}$ in terms of that of π.

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

Where η is expected discounted reward, ρ_π is the discounted visitation frequencies and $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$ is the advantage function.

The following statements hold for any θ_0,

$$L_{\pi_{\theta_0}}(\pi_{\theta_0}) = \eta(\pi_{\theta_0}),$$

$$\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_0}.$$

So, a suffcient small step $(\pi_{\theta_0} \to \widetilde{\pi})$ that improve $L_{\pi_{\theta_0}}$ will improve η.

Model Description (II)

Can be defined a class of methods called minorization-maximization that iterativelly (moving from $\pi_{old} \rightarrow \pi_{new}$) trying to raise the advantage of π_{new} over π_{old}.

In [1] it is defined the TRPO method that find the new policy solving the following maximization problem,

$$\begin{gathered} maximize_{\theta}\ L_{\theta_{old}}(\theta) \\ subject\ to\ D_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \leq \delta. \end{gathered}$$

Where $D_{KL}^{\rho_{\theta_{old}}}$ is the average KL divergence.
The optimization problem is solved using conjugate gradient descent followed by a line search.

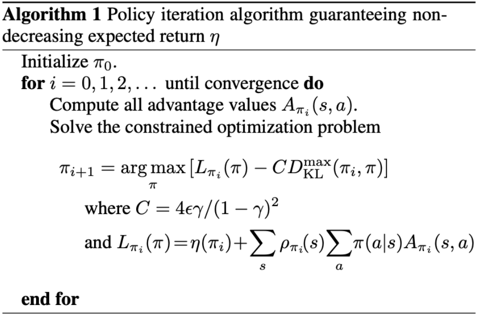
Model Description (III)

In [1] is also stated how the objective function and the constraint, of the just showed optimizations problem, can be approximated using Monte Carlo method.

- **single path**: sampling individual trajectories, model free settings.
- **vine**: construct a rollout set, performing multiple actions from each state in the rollout set, possible in simulation.

Key catch (I)

The key catch of the model described until now is the principled methodologies applied in a minorization-maximization algorithm, from which is derived the TRPO methodology, since the following algorithm doesn't permit large steps in the policy update.



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

```
Initialize $\pi_0$.
for $i = 0, 1, 2, \dots$ 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
```

Figure: Algorithm 1 from [1]

Key catch (II)

As a corollary of the Theorem1 in [1] holds that,

$$\eta(\widetilde{\pi}) \geq L_{\pi}(\widetilde{\pi}) - C\, D_{KL}^{max}(\pi, \widetilde{\pi}),$$
$$where\ C = \frac{4\epsilon\gamma}{(1-\gamma)^2},$$
$$and\ D_{KL}^{max}(\pi, \widetilde{\pi}) = \max_s D_{KL}(\pi(\cdot|s) \,\|\, \widetilde{\pi}(\cdot|s)).$$

From that equation the policies generated from the previous algorithm are monotonically improving w.r.t η.

Let $M_i(\pi) = L_{\pi_i}(\pi) - C\, D_{KL}^{max}(\pi_i, \pi)$, then
$\eta(\pi_{i+1}) - \eta(\pi_i) \geq M_i(\pi_{i+1}) - M_i(\pi_i)$, maximizing M_i at each iteration gives a non-decreasing η.
TRPO is derived from this methodoligy, taking out the $-C\, D_{KL}^{max}(\pi, \widetilde{\pi})$ factor from the objective and use an heuristic approximation of it in the constraint.

Results

The method described so far was tested over three high dimensional models of locomotion controller (hard problems), compared with other methods of gradient-free type and natural policy nature, TRPO outperforms all of them.

TRPO can do well even in the case of playing Atari games (using CNN with tens of thousand of parameters).

The results in [1] demonstrate good generalization properties, since there aren't used engineered policy, no locomotion notion inside the policy used for TRPO, and no engineered features in the case of Atari game playing as in other methods.

Comments

As demonstrated by the empirical results this method can do very well in high dimensional problems without having prior knowledges.

Moreover TRPO is a method derived by a principled approach that is rearranged, in a formal way, to deal with some unpractice numerical limitations given by the previous methods, such as small steps in previous principled techniques.

To the other hand in [1] is ignored the estimation error for the advantage function (A_π) in the inner optimization problem, so a deeper study on this fact could be done but was omitted for simplicity by the authors.

Bibliography

[1] John Schulman et al. "Trust region policy optimization". In: *International conference on machine learning*. PMLR. 2015, pp. 1889–1897.