Trust Region Policy Optimization

Matteo De Francesco

Department of Computer Science
University of Pisa

May 26, 2022

Introduction

We introduce an iterative procedure for optimizing policies, guaranteeing monotonic improvement.
We provide the theoretical framework and by making a series of approximation we develop a practical algorithm.
Trust Region Policy Optimization (TRPO) algorithm is built upon natural policy gradient methods, showing effectiveness in improving nonlinear policies $\pi(a|s)$ such as neural networks.
TRPO comes in two different fashion way:

- *single-path* method, which can be applied in model-free setting
- *vine* method, requiring the system to be restored in particular states, typically applicable only in simulation

Model Description

We can express the expected return of another policy $\tilde{\pi}$ in terms of the expected discounted reward $\eta(\pi)$ of the policy π

$$L_{\pi}(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\pi}(s) \sum_{a} \tilde{\pi}(a|s) A_{\pi}(s, a) \tag{1}$$

where $\rho(\cdot)$ is the discounted visitation frequencies and $A(\cdot, \cdot)$ is the advantage function.
The result above (i.e. parametrised ρ w.r.t. π) ensure that we have a monotonic increase since ρ ignores changes in state visitation frequencies, i.e. initial state probabilities are fixed w.r.t. the initial policy.
In [Kakade & Langford, 2002] the authors provided a lower bound when considering the new updated policy as a mixture of the previous policy and the new one

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

$$\text{where } \epsilon = \max_{s} \left| \mathbb{E}_{a \sim \pi'(a|s)} \left[A_{\pi}(s, a) \right] \right|$$

Key Catch

One of the key aspects is the following. We extend the previous bound of having a mixture of policies driven by α to general stochastic policies, using a distance measure between the old policy and the new one.

Theorem

The following bound holds

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - CD_{KL}^{max}(\pi, \tilde{\pi}) \tag{2}$$

where $C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$

which can be proved by showing that it construct a sequence of policies $\tilde{\pi}_i$ improving at each iteration i

Theoretical Algorithm

The theoretical algorithm developed from the previous equation is an algorithm of type minorization-maximization (MM). It is only a theoretical scheme since TRPO applies some approximation to the algorithm below

Initialize π_0
for $i = 0, 1, 2, \dots$ until converge **do**
  Compute all advantage values $A_{\pi_i}(s, a)$
  Solve the constrained optimization problem
  $\pi_{i+1} = \arg\max_\pi \left[L_{\pi_i}(\pi) - CD_{KL}^{max}(\pi_i, \pi)\right]$
  where $C \frac{4\epsilon\gamma}{(1-\gamma)^2}$
  and $L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a)$
end for

TRPO

The policy update above is not suitable in practical application

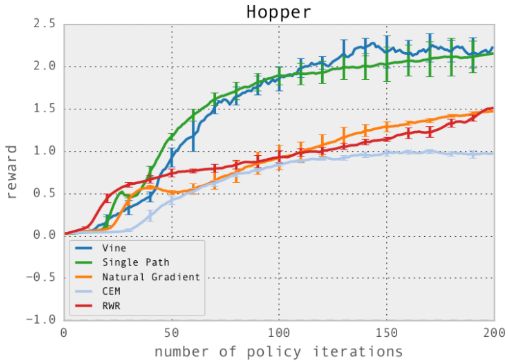
- Using the penalty coefficient C reflects on very small step sizes $\Longrightarrow$ we substitute it with a trust region constraint
- The KL divergence imposes a large number of constraints after applying the change above $\Longrightarrow$ we consider the average KL divergence

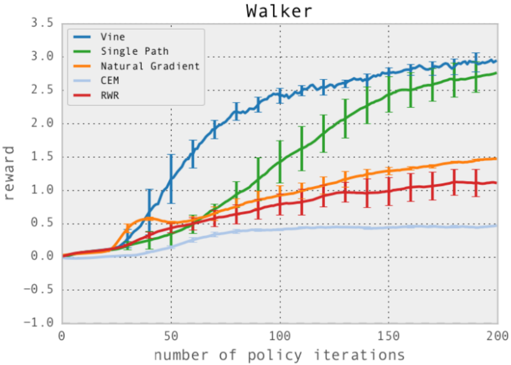
$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & L_{\theta_{old}}(\theta) \\ \text{subject to} \quad & \bar{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \leq \delta \end{aligned} \tag{3}$$

which can be solved by using Monte Carlo simulation

Results

TRPO outperforms different methods in the simulated robotic locomotion environment, in particular when considering the *Hopper* and *Walker* tasks.





Conclusion

To conclude, we can highlight some pros and cons of the described method.

Pros:

- The introduction of the trust region constraint on the KL divergence brings strong improvements in the policy update, allowing for larger step sizes

Cons:

- The *vine* method suffer from less variance but needs to evaluate multiple paths from a fixed state. Doing so is possible mostly only in simulated environments

References

- Schulman, J., Levine, S., Moritz, P., Jordan, M. & Abbeel, P. Trust Region Policy Optimization. (arXiv,2015), https://arxiv.org/abs/1502.05477
- Kakade, S. & Langford, J. Approximately Optimal Approximate Reinforcement Learning. *IN PROC. 19TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*. pp. 267-274 (2002)