Trust Region Policy Optimization (TRPO)

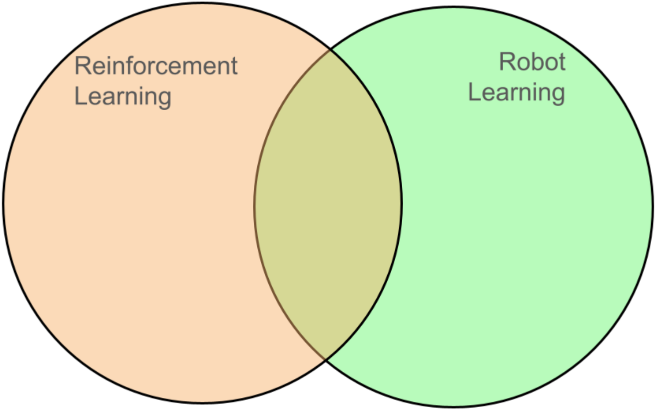
Original Paper by Schulman et al. [2017]

Matthew Vandergrift

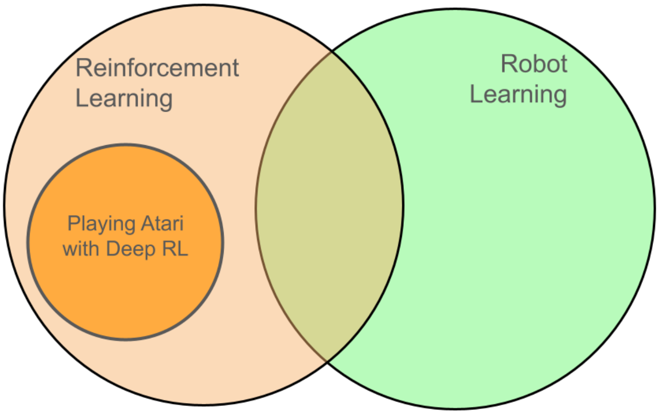
Robot Learning Seminar Presentation

March 2025

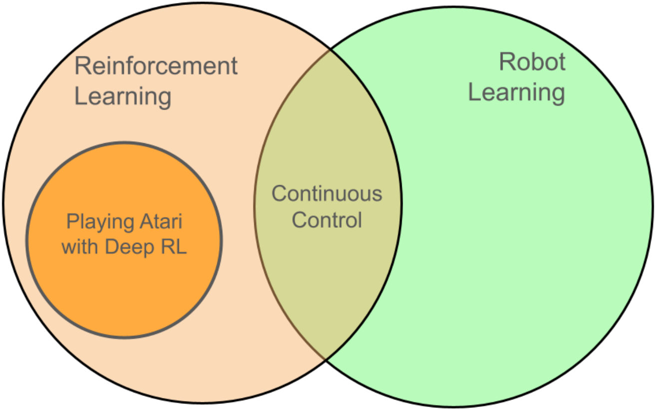
Motivation



Motivation


Reinforcement Learning
Playing Atari with Deep RL
Robot Learning

Motivation


Reinforcement
Learning
Playing Atari
with Deep RL
Continuous
Control
Robot
Learning

Existing Solutions

- Reinforce
- Basic Actor-Critic Algorithms
- Natural Policy Gradients
- Derivative Free Methods: cross-entropy method, covariance matrix adaptation.

Once again, ... RL

”RL is computational framework for learning from interaction” Sutton and Barto [2018]. The agent interacts with an environment with the goal of maximizing expected return. Let us denote expected return for a particular policy by $\eta(\pi)$.

Advantage Function

Recall the advantage function $A_\pi(s, a)$ defined as,

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$$

This function tells us how "good" taking action is compared to the average action.

Policy Improvement via Advantage

Since policies are just collections of actions, we can use advantage function to evaluate them. Let π and π' be two different policies. Equation 1 gives a way to write the performance of π' using the performance of π and the advantage function.

$$\eta(\pi') = \eta(\pi) + \sum_{s} \mu_{\pi'}(s) \sum_{a} \pi'(a|s) A_{\pi}(s, a). \tag{1}$$

Proof in appendix

RL is Solved!

At first glance we have a solution!

Algorithm The Perfect RL Algorithm

Require: π and $\eta(\pi)$

1: $\max_{\pi'} \left(\eta(\pi) + \sum_s \mu_{\pi'}(s) \sum_a \pi'(a|s) A_\pi(s,a)\right)$

This doesn't work because we have a dependence on the policy distribution which is not something we have access to when considering π'.

Dealing with $\mu_{\pi'}$

- Let's use μ_π instead of $\mu_{\pi'}$
- Define $L_\pi(\pi') = \eta(\pi) + \sum_s \mu_\pi(s) \sum_a \pi'(a|s) A_\pi(s,a)$
- Assume π is a parameterized by weights θ.
- Gives us a local first order approximation, $\nabla_\theta L_{\pi_{\theta_0}}(\theta_\theta)|_{\theta=\theta_0} = \nabla_\theta \eta_{\pi_{\theta_0}}(\theta_\theta)|_{\theta=\theta_0}$
- If we take **small** steps in θ then we can use our 'Perfect RL algorithm'.

What is a small step?

A Major Contribution of the Paper is the following bound,

Theorem

Let $D_{KL}^{max}(\pi, \pi') := max_s D_{KL}\left(\pi(*|S)||\pi'(*|s)\right)$. *We then have that,* $\eta(\pi') \geq L_\pi(\pi') - CD_{KL}^{max}(\pi, \pi')$ *where* $C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$

This means we can bound the improvement between any-two policies based on their KL divergence. This means if we optimize within a certain KL distance we can guarantee *guarantee improvement*.

A More Perfect RL Algorithm

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

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

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

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

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

end for

The constraint is not computable due to $\max_s f(s)$.

The final steps

- Computable constraint
- Cheap cost function
- Cheap constrained optimization solver

Computable Constraint

We want to make our constrained optimization solvable.

- Get rid of max KL constraint over the whole state space.
- Define 'Average' KL,
  $\bar{D}_{KL} := \mathbb{E}_{s \sim \mu_{\pi_\theta}} \left[D_{KL} \left(\pi_\theta(*|S) || \pi_{\theta_{old}}(*|s) \right) \right]$
- Estimate this Expectation using roll-outs under the policy.

This gives us,

$$\max_\theta \; L_{\theta_{old}}(\theta)$$
$$\text{subject to} \;\; \bar{D}_{KL}(\theta_{old}, \theta) \leq \delta$$

Cheap Cost Function

We want to make $L_{\theta_{old}}(\theta)$ fast to compute.

- $L_{\theta_{old}}(\theta) = \sum_s \mu_{\pi_{\theta_{old}}} \sum_a \pi_\theta(a|s) A_{\theta_{old}}(s,a)$ (Definition of L)
- $L_{\theta_{old}}(\theta) = \sum_s \mu_{\pi_{\theta_{old}}} \mathbb{E}_{a \sim \pi_{\theta_{old}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} \cdot A_{\theta_{old}}(s,a) \right]$
  (Replaced sum over actions, with expectation.)
- $L_{\theta_{old}}(\theta) = \sum_s \mu_{\pi_{\theta_{old}}} \mathbb{E}_{a \sim \pi_{\theta_{old}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} \cdot Q_{\theta_{old}}(s,a) \right]$
  $\Big($Replaced A with an estimator $\hat{A}$.$\Big)$

Cheap constrained optimization solver

$$\max_\theta L_{\theta_{old}}(\theta) \text{ subject to } \bar{D}_{KL}(\theta_{old}, \theta) \leq \delta$$

We use two steps,

- Compute Search Direction
- Line Search in found Direction

Computing Search Direction

An Introduction to
the Conjugate Gradient Method
Without the Agonizing Pain
Edition $1\frac{1}{4}$

Jonathan Richard Shewchuk

August 4, 1994

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Line Search

Trust Region Policy Optimization

Experimental Results in TRPO Paper

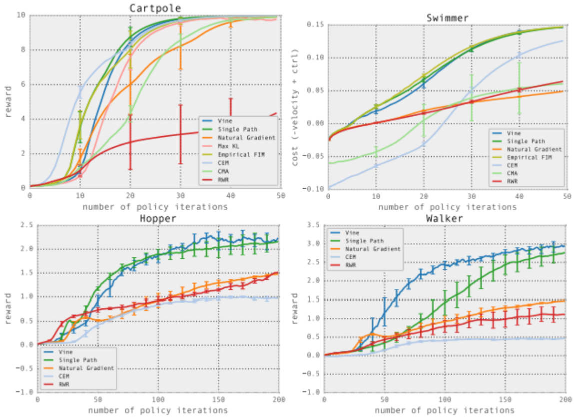


Figure 4. Learning curves for locomotion tasks, averaged across five runs of each algorithm with random initializations. Note that for the hopper and walker, a score of -1 is achievable without any forward velocity, indicating a policy that simply learned balanced standing, but not walking.

External TRPO Robotics Applications

Mahmood et al. [2018] wrote a paper bencmarking policy gradient algorithms for robotics, including TRPO.

- "TRPO achieving near-best final learning performance in all tasks."
- "Among these algorithms, the final performance of TRPO was never substantially worse compared to the best in each task."
- "TRPO's performance was the least sensitive to hyper-parameter variations with the smallest interquartile range on both tasks."

Thank you for listening

Robots following TRPO Policies

References

A. Rupam Mahmood, Dmytro Korenkevych, Gautham Vasan, William Ma, and James Bergstra. Benchmarking reinforcement learning algorithms on real-world robots, 2018. URL https://arxiv.org/abs/1809.07731.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017. URL https://arxiv.org/abs/1502.05477.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.