CS885 Reinforcement Learning
Module 1: May 26, 2020

Trust Region and Proximal Policy Optimization

Schulman, Levine et al. (2015) Trust Region Policy Optimization
Schulman Wolski et al. (2017) Proximal Policy Optimization

Gradient policy optimization

- REINFORCE algorithm
- Advantage Actor Critic (A2C)
- Deterministic Policy Gradient (DPG)
- Trust Region Policy Optimization (TRPO)
- Proximal Policy Optimization (PPO)

Recall Policy Gradient

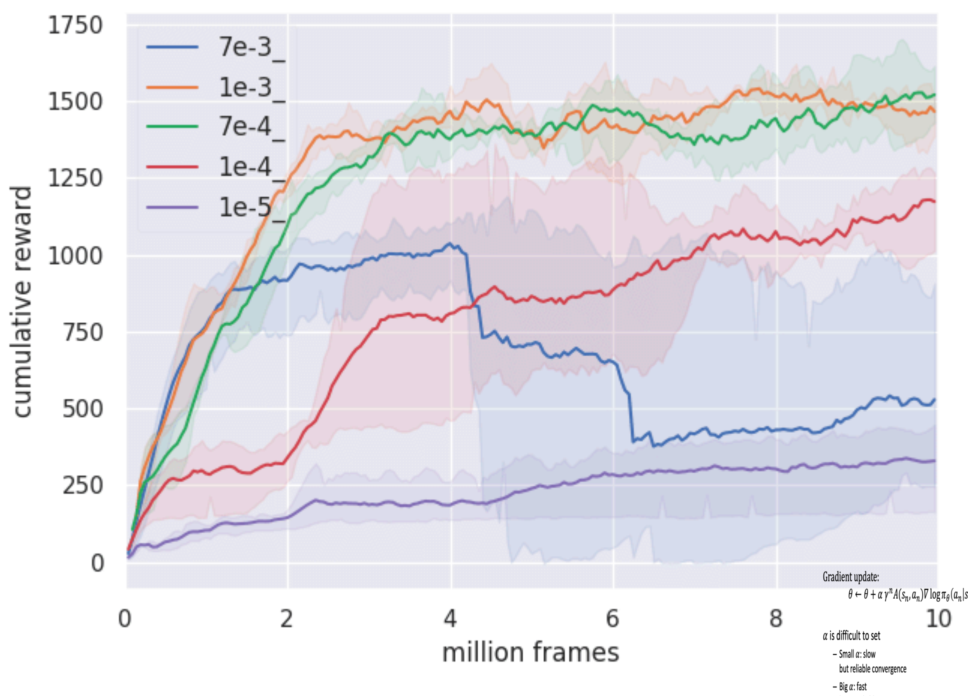
Gradient update:

$$\theta \leftarrow \theta + \alpha\, \gamma^n A(s_n, a_n) \nabla \log \pi_\theta(a_n|s_n)$$

α is difficult to set

- Small α: slow
  but reliable convergence
- Big α: fast
  but unreliable

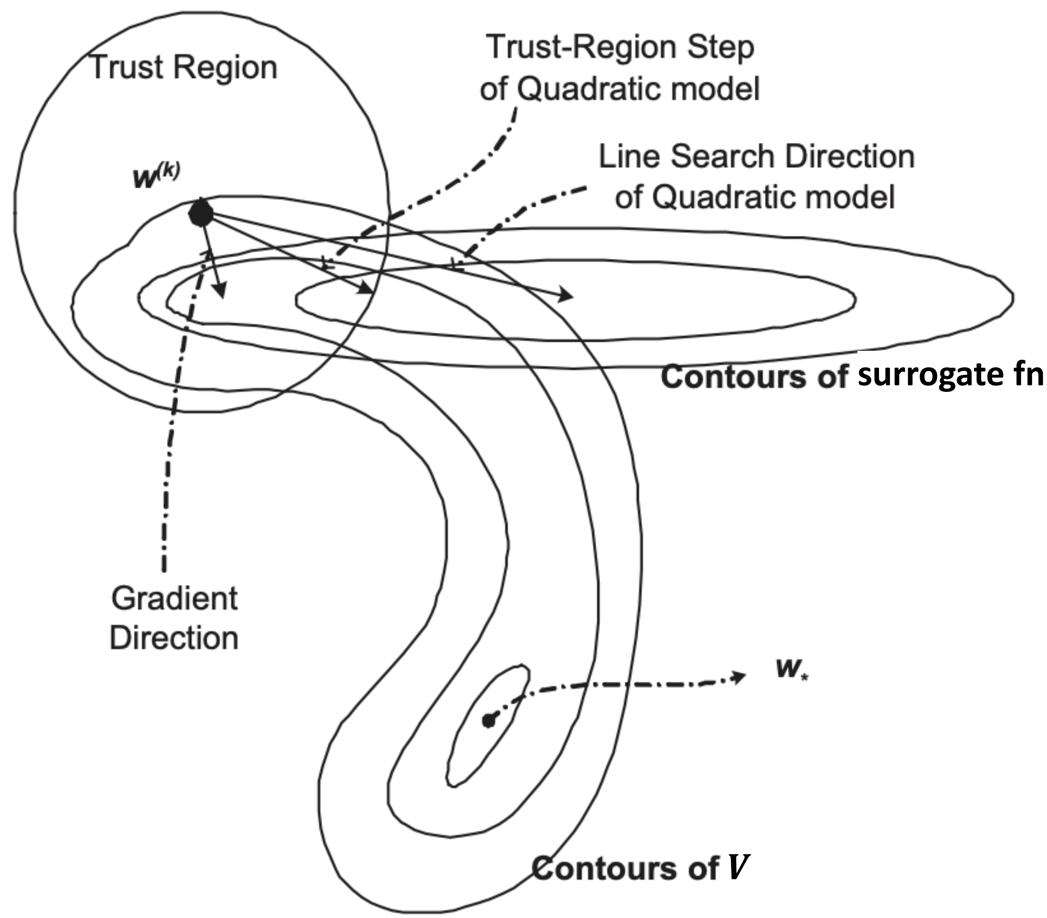
A2C on hopper-v2 with different α's
Wu, Sun et al. (2018)



Trust region method

- We often optimize a surrogate objective (approximation of V)
- Surrogate objective may be trustable (close to V) only in a small region

- Limit search to small trust region



Choi, Choi (2005)

Trust region for policies

- Let θ be the parameters for policy $\pi_\theta(s|a)$
- We can define a region

  around θ: $\{\theta' |\, D(\theta, \theta') < \delta\}$

  or around π_θ: $\{\theta' |\, D(\pi_\theta, \pi_{\theta'}) < \delta\}$

  where D is a distance measure
- V often varies more smoothly with π_θ than θ

  small change in π_θ $\xrightarrow{\text{usually}}$ small change in V

  small change in θ $\xrightarrow{\text{more often}}$ large change in V
- Hence, define policy trust regions

Kullback-Leibler Divergence

KL-Divergence is a common distance measure for distributions:

$$D_{KL}(p, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Intuition: expectation of the logarithm difference between p and q

KL-Divergence for policies at a state s:

$$D_{KL}\left(\pi_\theta(\cdot\,|s), \pi_{\tilde{\theta}}(\cdot\,|s)\right) = \sum_a \pi_\theta(a|s) \log \frac{\pi_\theta(a|s)}{\pi_{\tilde{\theta}}(a|s)}$$

Trust Region Policy Optimization

- Consider an initial state distribution $p(s_0)$

- Update step:

$$\theta \leftarrow \underset{\tilde{\theta}}{\operatorname{argmax}} E_{s_0 \sim p}[V^{\pi_{\tilde{\theta}}}(s_0) - V^{\pi_\theta}(s_0)]$$

$$\text{subject to } \max_s D_{KL}\big(\pi_\theta(\cdot\,|s), \pi_{\tilde{\theta}}(\cdot\,|s)\big) \leq \delta$$

Reformulation

- Since the objective is not directly computable, let's approximate it:

$$\underset{\tilde{\theta}}{\operatorname{argmax}}\, E_{s_0 \sim p}[V^{\pi_{\tilde{\theta}}}(s_0) - V^{\pi_\theta}(s_0)]$$

$$\approx \underset{\tilde{\theta}}{\operatorname{argmax}}\, E_{s \sim \mu_\theta,\, a \sim \pi_\theta}\left[\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)} A_\theta(s,a)\right]$$

where $\mu_\theta(s)$ is the stationary state distribution for π

- Let's also relax the bound on the max KL-divergence to a bound on the expected KL-divergence

$$\max_s D_{KL}\big(\pi_\theta(\cdot\,|s), \pi_{\tilde{\theta}}(\cdot\,|s)\big) \leq \delta$$

is relaxed to $E_{s \sim \mu_\theta}\left[D_{KL}\left(\pi_\theta(\cdot\,|s), \pi_{\tilde{\theta}}(\cdot\,|s)\right)\right] \leq \delta$

Derivation

$$\underset{\tilde{\theta}}{\operatorname{argmax}} E_{s\sim\mu_\theta,\, a\sim\pi_\theta}\left[\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)} A_\theta(s,a)\right]$$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} \sum_s \mu_\theta(s) \sum_a \pi_\theta(a|s) \left[\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)} A_\theta(s,a)\right]$$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} \sum_s \mu_\theta(s) \sum_a \pi_{\tilde{\theta}}(a|s)\, A_\theta(s,a)$$

since $\mu_{\tilde{\theta}} \approx \mu_\theta$

$$\approx \underset{\tilde{\theta}}{\operatorname{argmax}} \sum_s \mu_{\tilde{\theta}}(s) \sum_a \pi_{\tilde{\theta}}(a|s) A_\theta(s,a)$$

since $\mu_{\tilde{\theta}}(s) \propto \sum_{n=0}^{\infty} \gamma^n P_{\tilde{\theta}}(s_n = s)$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} \sum_s \sum_{n=0}^{\infty} \gamma^n P_{\tilde{\theta}}(s_n = s) \sum_a \pi_{\tilde{\theta}}(a|s) A_\theta(s,a)$$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} E_{s_0,s_1,\ldots\sim P_{\tilde{\theta}},\; a_0,a_1,\ldots\sim\pi_{\tilde{\theta}}}\left[\sum_{n=0}^{\infty} \gamma^n A_\theta(s_n,a_n)\right]$$

Derivation (continued)

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} E_{s_0,s_1,\ldots\sim P_{\tilde{\theta}}\,,\,a_0,a_1,\ldots\sim\pi_{\tilde{\theta}}}\left[\sum_{n=0}^{\infty}\gamma^n A_\theta(s_n,a_n)\right]$$

since $A_\theta(s,a) = E_{s'\sim P(s'|s,a)}[r(s) + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s)]$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} E_{s_0,s_1,\ldots\sim P_{\tilde{\theta}}\,,\,a_0,a_1,\ldots\sim\pi_{\tilde{\theta}}}\left[\sum_{n=0}^{\infty}\gamma^n (r(s_n) + \gamma V^{\pi_\theta}(s_{n+1}) - V^{\pi_\theta}(s_n))\right]$$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} E_{s_0,s_1,\ldots\sim P_{\tilde{\theta}}\,,\,a_0,a_1,\ldots\sim\pi_{\tilde{\theta}}}\left[\sum_{n=0}^{\infty}\gamma^n r(s_n) - V^{\pi_\theta}(s_0)\right]$$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}} E_{s_0,s_1,\ldots\sim P_{\tilde{\theta}}\,,\,a_0,a_1,\ldots\sim\pi_{\tilde{\theta}}}\left[V^{\pi_{\tilde{\theta}}}(s_0) - V^{\pi_\theta}(s_0)\right]$$

$$= \underset{\tilde{\theta}}{\operatorname{argmax}}\ E_{s_0\sim P}\left[V^{\pi_{\tilde{\theta}}}(s_0) - V^{\pi_\theta}(s_0)\right]$$

Trust Region Policy Optimization (TRPO)

TRPO()

Initialize π_θ to anything

Loop forever (for each episode)

Sample s_0 and set $n \leftarrow 0$

Repeat N times

Sample $a_n \sim \pi_\theta(a|s_n)$

Execute a_n, observe s_{n+1}, r_n

$\delta \leftarrow r_n + \gamma \max_{a_{n+1}} Q_w(s_{n+1}, a_{n+1}) - Q_w(s_n, a_n)$

$A(s_n, a_n) \leftarrow r_n + \gamma \max_{a_{n+1}} Q_w(s_{n+1}, a_{n+1}) - \sum_a \pi_\theta(a|s_n) Q_w(s_n, a)$

Update Q: $w \leftarrow w + \alpha_w \gamma^n \delta \, \nabla_w Q_w(s_n, a_n)$

$n \leftarrow n + 1$

Update π: $\theta \leftarrow \operatorname{argmax}_{\tilde{\theta}} \frac{1}{N} \sum_{n=0}^{N-1} \frac{\pi_{\tilde{\theta}}(a_n|s_n)}{\pi_\theta(a_n|s_n)} A_\theta(s_n, a_n)$

subject to $\frac{1}{N} \sum_{n=0}^{N-1} D_{KL}\left(\pi_\theta(\cdot\,|s_n), \pi_{\tilde{\theta}}(\cdot\,|s_n)\right) \leq \delta$

linear approximation

quadratic approximation

Constrained Optimization

- TRPO is conceptually and computationally challenging in large part because of the constraint in the optimization.

$$\max_s D_{KL}\big(\pi_\theta(\cdot\,|s), \pi_{\tilde{\theta}}(\cdot\,|s)\big) \leq \delta$$

- What is the effect of the constraint?
- Recall KL-Divergence:

$$D_{KL}\big(\pi_\theta(\cdot\,|s), \pi_{\tilde{\theta}}(\cdot\,|s)\big) = \sum_a \pi_\theta(a|s) \log \frac{\pi_\theta(a|s)}{\pi_{\tilde{\theta}}(a|s)}$$

We are effectively constraining the ratio $\frac{\pi_\theta(a|s)}{\pi_{\tilde{\theta}}(a|s)}$

Simpler Objective

Let's design a simpler objective that directly constrains $\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)}$

$$\underset{\tilde{\theta}}{\operatorname{argmax}}\, E_{s\sim\mu_\theta,\, a\sim\pi_\theta} \min \left\{ \begin{array}{c} \frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)} A_\theta(s,a), \\ clip\left(\frac{\pi_{\tilde{\theta}}(a|s)}{\pi_\theta(a|s)}, 1-\epsilon, 1+\epsilon\right) A_\theta(s,a) \end{array} \right\}$$

where $clip(x, 1-\epsilon, 1+\epsilon) = \begin{cases} 1-\epsilon & if\ x < 1-\epsilon \\ x & if\ 1-\epsilon \le x \le 1+\epsilon \\ 1+\epsilon & if\ x > 1+\epsilon \end{cases}$

Proximal Policy Optimization (PPO)

PPO()
  Initialize π_θ to anything
  Loop forever (for each episode)
    Sample s_0 and set $n \leftarrow 0$
    Repeat N times
      Sample $a_n \sim \pi_\theta(a|s_n)$
      Execute a_n, observe s_{n+1}, r_n
      $\delta \leftarrow r_n + \gamma \max_{a_{n+1}} Q_w(s_{n+1}, a_{n+1}) - Q_w(s_n, a_n)$
      $A(s_n, a_n) \leftarrow r_n + \gamma \max_{a_{n+1}} Q_w(s_{n+1}, a_{n+1}) - \sum_a \pi_\theta(a|s_n) Q_w(s_n, a)$
      Update Q: $w \leftarrow w + \alpha_w \gamma^n \delta \nabla_w Q_w(s_n, a_n)$
      $n \leftarrow n + 1$
    Update π:

optimize by stochastic gradient descent

$$\theta \leftarrow \underset{\tilde{\theta}}{\operatorname{argmax}} \frac{1}{N} \sum_{n=0}^{N-1} \min \left\{ \begin{array}{c} \frac{\pi_{\tilde{\theta}}(a_n|s_n)}{\pi_\theta(a_n|s_n)} A(s_n, a_n), \\ clip\left(\frac{\pi_{\tilde{\theta}}(a_n|s_n)}{\pi_\theta(a_n|s_n)}, 1-\epsilon, 1+\epsilon\right) A(s_n, a_n) \end{array} \right\}$$

Empirical Results

- Comparison on several robotics tasks

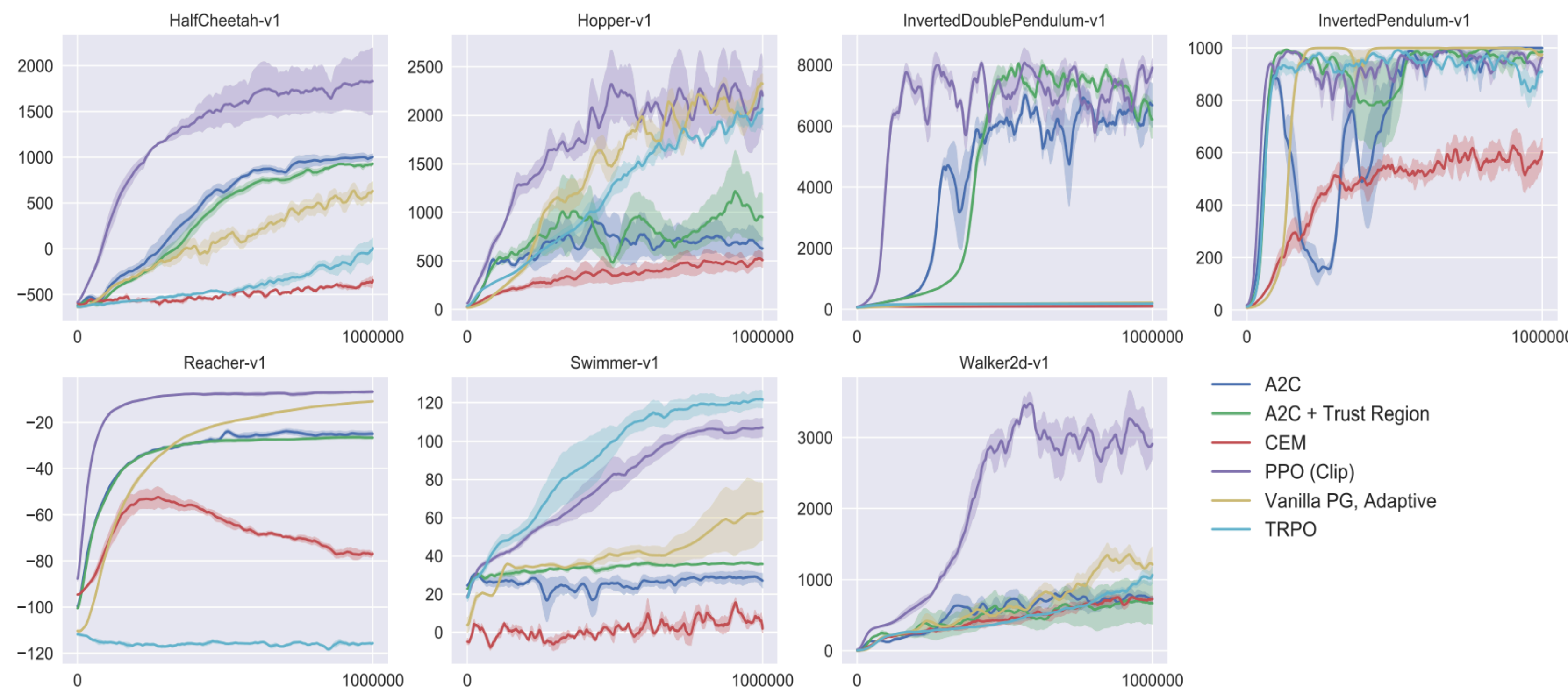
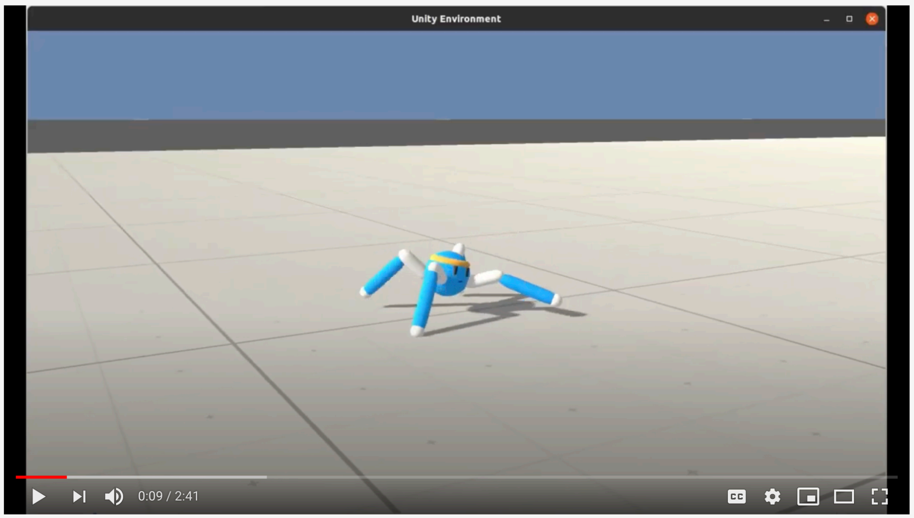


Figure 3: Comparison of several algorithms on several MuJoCo environments, training for one million timesteps.

From Schulman et al., 2017

Illustration

https://youtu.be/D6ZuxeNvkXE



Proximal Policy Optimization (PPO) trained on the Unity Crawler Environment

https://youtu.be/bqdjsmSoSgI

Proximal Policy Optimization - Robust knocked over stand up

Agent tries to reach a target, learning to walk, run, turn, recover from minor hits, and how to stand up from the ground.