Trust-Region constrained Policy Optimization

Presented by :
Atmani Hanan

University mohammed vi polytechnic

July 23

Presentation plan

Introduction

- Let's add to the CPI method, which uses a small step size to ensure incremental updates to policies, another popular approach for incremental policy updates that involves explicitly enforcing a small change in the policy distribution via a trust region constraint.
- At iteration t with the current policy π_{θ_t} , we are interested in the following local trust-region constrained optimization:

$$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)} \log \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)$$
$$\text{s.t., } D_{KL}\left(\Pr_{\mu}^{\pi^{\theta_t}} \| \Pr_{\mu}^{\pi_\theta}\right) \leq \delta,$$

- where recall $\Pr_{\mu}^{\pi}(\tau)$ is the trajectory distribution induced by π starting at $s_0 \sim \mu$, and $D_{KL}(P_1 \| P_2)$ are KL-divergence between two distribution P_1 and P_2

TRPO

- As we are interested in small local update in parameters, we can perform sequential quadratic programming here, i.e., we can further linearize the objective function at θ_t and quadratize the KL constraint at θ_t to form a local quadratic programming:

$$\max_{\theta} \left\langle \mathbb{E}_{s\sim d_{\mu}^{\pi_t}} \mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \nabla_{\theta} \log \pi_{\theta_t}(a \mid s) A^{\pi^{\theta_t}}(s,a), \theta - \theta_t \right\rangle$$

$$\text{s.t.,} \left\langle \nabla_{\theta} D_{KL}\left(\Pr_{\mu}^{\pi^{\theta_t}} \| \Pr_{\mu}^{\pi_{\theta}}\right)\Big|_{\theta=\theta_t}, \theta - \theta_t \right\rangle +$$

$$\frac{1}{2}(\theta-\theta_t)^{\top}\left(\nabla_{\theta}^2 D_{KL}\left(\Pr_{\mu}^{\pi^{\theta_t}} \| \Pr_{\mu}^{\pi_{\theta}}\right)\Big|_{\theta=\theta_t}\right)(\theta-\theta_t) \leq \delta, \tag{1}$$

TRPO

Proposition

Consider a finite horizon MDP with horizon H. Consider any fixed θ_t. We have:

$$\nabla_\theta D_{KL}\left(\Pr_\mu^{\pi_{\theta_t}} \| \Pr_\mu^{\pi_\theta}\right)\Big|_{\theta=\theta_t} = 0,$$

$$\nabla_\theta^2 D_{KL}\left(\Pr_\mu^{\pi_{\theta_t}} \| \Pr_\mu^{\pi_\theta}\right)\Big|_{\theta=\theta_t} = H\mathbb{E}_{s,a\sim d^{\pi_{\theta_t}}} \nabla \log \pi_{\theta_t}(a \mid s) (\nabla \log \pi_{\theta_t}(a \mid s)$$

TRPO

- The above Proposition shows that a second order taylor expansion of the KL constraint over trajectory distribution gives a local distance metric at θ_t :

$$\frac{1}{2}(\theta - \theta_t) F_{\theta_t} (\theta - \theta_t)$$

Where

$$F_{\theta_t} = H\mathbb{E}_{s,a\sim d^{\pi_{\theta_t}}} \nabla \log \pi_{\theta_t}(a \mid s) \left(\nabla \log \pi_{\theta_t}(a \mid s)\right)^\top$$

is proportional to the fisher information matrix

TRPO

Now using the results from proposition, we can verify that the local policy optimization procedure in Eq. (1) exactly recovers the NPG update, where the step size is based on the trust region parameter δ. Denote $\Delta = (\theta - \theta_t)$, we have

$$\max_{\theta} \langle \Delta, \nabla_\theta V^{\pi_{\theta_t}} \rangle ,$$
$$\text{s.t., } \Delta^\top F_{\theta_t} \Delta^\top \leq 2\delta,$$

which gives the following update procedure:

$$\theta_{t+1} = \theta_t + \Delta = \theta_t + \sqrt{\frac{2\delta}{(\nabla V^{\pi\theta_t})^\top F_{\theta_t}^{-1} \nabla V^{\pi\theta_t}}} \cdot F_{\theta_t}^{-1} \nabla V^{\pi_{\theta_t}}$$

where note that we use the self-normalized learning rate computed using the trust region parameter δ