Advanced Robotics: Homework 11

Ngo Anh Vien

July 9, 2014

1 Lecture Note

For a lecture note, I would like to refer you to a paper: *a survey paper on policy search for robotics* Deisenroth et al. [2013].

2 Finite Difference Method

Finite Difference (FD) method is a policy gradient technique. The gradient $\nabla_\theta J(\theta_k)$ is estimated by purturbing the parameter vector θ a small perturbations $\delta\theta^i$. For each $\delta\theta^i$, black-box evaluate $\delta J^i = J(\theta_k + \delta\theta^i) - J(\theta_k)$. Prove that

$$\nabla_\theta^{FD} J(\theta_k) \approx (\delta\Theta^\top \delta\Theta)^{-1} \delta\Theta^\top \delta J$$

where $\delta\Theta = \left[\delta\theta^1, \delta\theta^2, \ldots, \delta\theta^N\right]^\top$, and $\delta J = [\delta J^1, \ldots, \delta J^N]$

3 REINFORCE Algorithm

The baseline b is computed minimizing the variance of $\nabla_\theta^{RF} J(\theta)$. Derive the result of b as briefly discussed in the lecture

$$b_i = \frac{E_{p_\theta(\xi)}\left[\left(\sum_{t=0}^{T-1} \nabla_{\theta_i} \log \mu_\theta(a_t|s_t)\right)^2 R(\xi)\right]}{E_{p_\theta(\xi)}\left[\left(\sum_{t=0}^{T-1} \nabla_{\theta_i} \log \mu_\theta(a_t|s_t)\right)^2\right]}$$

where i is the ith-dim of θ.

4 G(PO)MDP Algorithm

Similarly, derive the baseline in the G(PO)MDP algorithm

$$b_{ji} = \frac{E_{p_\theta(\xi)}\left[\left(\sum_{t=0}^{j} \nabla_{\theta_i} \log \mu_\theta(a_t|s_t)\right)^2 r_j\right]}{E_{p_\theta(\xi)}\left[\left(\sum_{t=0}^{j} \nabla_{\theta_i} \log \mu_\theta(a_t|s_t)\right)^2\right]}$$

for $j = 0, 1, \ldots, T-1$, and i is the ith-dim of θ.

5 Natural Policy Gradient Algorithm

The natural policy gradient algorithm computes the steepest ascent direction at a given θ_k. Assuming that we want to maximize the expected return $J(\theta)$, and the parameter space with an A-weighted norm: $\|\theta - \theta_k\|_A = (\theta - \theta_k)^\top A(\theta - \theta_k)$. The question is: what is the steepest ascent direction:$\theta_{k+1} = \theta_k + \delta\theta^{NG}$.

The answer is: $\delta\theta^{NG}$ should be to $\arg\max_{\delta\theta} J(\theta_k + \delta\theta)$ s.t with a small purturbation from θ_k: $\|\delta\theta\| = \|\theta_k - \theta\|_A = \epsilon$. The constrained optimization problem is

$$\delta\theta^{NG} = \arg\max_{\delta\theta} J(\theta_k + \delta\theta), \text{ s.t. } \|\delta\theta\|_A = \|\theta_k - \theta\|_A = \epsilon$$

- Using Taylor expansion and Lagrange multipliers to prove the result of the steepest gradient direction
$$\delta\theta^{NG} = A^{-1}\nabla_\theta J(\theta_k)$$
- For more detail: see a note on Natural Gradient from Nathan
http://www.nathanratliff.com/pedagogy

References

Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2013.