Noname manuscript No.
(will be inserted by the editor)

Natural Actor Critic

Do you have a subtitle? If so, write it here

First Author · Second Author

Received: date / Accepted: date

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

1 Paper

Natural Actor Critic:

- Main Version [6].
- 2nd Version: Natural Actor-Critic in Neurocomputing [4].
- 3rd version: RL of motor skills with policy gradients in NN [5].

Recommended by Jan:

- Policy Evaluation with TD [3].
- Incremental NAC algorithms [1].
- Jan said that a paper form C. Dann is very important. Did he mean Policy Evaluation with TD by Dann or did he mean a second paper?

Research:

- Comparison of four natural gradient algorithms (co-author Sutton) [2].

2 Meetings & Notes

Meetings:

- 12.12.18: Notes from Jan can be found in ".\Notes Jan 12.12.18"

3 Introduction

- Steepest ascent direction of performance object with respect to any metric $M(\theta)$: $M(\theta)^{-1}\nabla_\theta J(\mu_\theta)$
- The natural gradient ist the steepest ascent direction with respect to the Fisher information metric $M_\pi(\theta) = E_{s\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta log\pi_\theta(a|s)^T \nabla_\theta log\pi_\theta(a|s)]$
- For deterministic policies: $M_\mu(\theta) = E_{s\sim\rho^\mu}[\nabla_\theta\mu_\theta(s)\nabla_\theta\mu_\theta(s)^T w]$
  - Limiting case of the Fisher information metric: policy variance reduced to zero
- Combining DPG theorem with compatible function approximation gives $\nabla_\theta J(\mu_\theta) = E_{s\sim\rho^\mu}[\nabla_\theta\mu_\theta(s)\nabla_\theta\mu_\theta(s)^T w]$ so steepest ascent direction reduces to $M_\mu(\theta)^{-1}\nabla_\theta J_\beta(\mu_\theta) = w$

4 Episodic NAC

Important to understand beforehand: In episodic NAC, our system of equations has one equation per trajectory and not one equation per action as in the normal NAC algorithm.

First we start by adding together the advantage function across an episode e where we made N steps.

$$A(s,a) = r(s,a) + \gamma V(s') - V(s) \tag{1}$$
$$\gamma A(s',a') = \gamma r(s',a') + \gamma^2 V(s'') - \gamma V(s') \tag{2}$$
$$A(s,a) + \gamma A(s',a') = r(s,a) + \gamma r(s',a') + \gamma^2 V(s'') - V(s) \tag{3}$$
$$\sum_{i=0}^{N} \gamma^i A(s_i,a_i) = \sum_{i=0}^{N} \gamma^i r(s_i,a_i) + \gamma^N V(S_{N+1}) - V(S_0) \tag{4}$$

If we assume $\gamma \neq 1$, we can remove the term $\gamma^N V(S_{N+1})$, because in the limit the term becomes zero ($\gamma^N \to 0$). Additionally, if we assume that we always start in the same start S_0, we can write $V(S_0)$ as our cost function J because it will exactly sum up the expected Reward/cost of our problem.

$$\Rightarrow \sum_{i=0}^{N} \gamma^i A(s_i,a_i) = \sum_{i=0}^{N} \gamma^i r(s_i,a_i) - J \tag{5}$$

Now we can plug in the parametrisized gradient descent for the advantage function. That this works and is indeed the same has been proven by reference. Additionally we bring the cost J to the other side of the equation.

$$\Rightarrow \sum_{i=0}^{N} \gamma^i \nabla_{\Theta} \left[\log \pi(a_i|s_i)^T\right] \cdot w + 1 \cdot J = \sum_{i=0}^{N} \gamma^i r(s_i,a_i) \tag{6}$$

Let's do some rewriting. We define the following two terms:

$$\Phi_e = \left[\sum_{i=0}^{N} \gamma^i \nabla_{\Theta} \left[\log \pi(a_i|s_i)^T\right], 1\right] \tag{7}$$
$$R_e = \sum_{i=0}^{N} \gamma^i r(s_i,a_i) \tag{8}$$

This let's us rewrite equation 6 as:

$$\Phi_e \cdot \begin{bmatrix} w \\ J \end{bmatrix} = R_e \tag{9}$$

An easy way to solve this system of equations is by taking the pseudo inverse of Φ_e.

$$\begin{bmatrix} w \\ J \end{bmatrix} = (\Phi_e^T \Phi_e)^{-1} \Phi_e^T R_e \tag{10}$$

Algorithm 1 Episodic Natural Actor-Critic Algorithm (eNAC)

```
Require: $n \geq 0 \vee x \neq 0$
Ensure: $y = x^n$
  for $u = 1, 2, 3, \ldots$ do
    for $e = 1, 2, 3, \ldots$ do
      Execute Rollout: Draw initial state $s_0 \sim p(s_0)$
      for $t = 1, 2, 3, \ldots, N$ do
        Draw action $u_t \sim \pi(a_t|s_t)$, observe next state $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$, and reward $r_t = r(s_t, a_t)$.
      end for
    end for
  end for
```

References

1. Bhatnagar S, Ghavamzadeh M, Lee M, Sutton RS (2008) Incremental natural actor-critic algorithms. In: Advances in neural information processing systems, pp 105–112
2. Bhatnagar S, Sutton RS, Ghavamzadeh M, Lee M (2009) Natural actor–critic algorithms. Automatica 45(11):2471–2482
3. Dann C, Neumann G, Peters J (2014) Policy evaluation with temporal differences: A survey and comparison. The Journal of Machine Learning Research 15(1):809–883
4. Peters J, Schaal S (2008) Natural actor-critic. Neurocomputing 71(7-9):1180–1190
5. Peters J, Schaal S (2008) Reinforcement learning of motor skills with policy gradients. Neural networks 21(4):682–697
6. Peters J, Vijayakumar S, Schaal S (2005) Natural actor-critic. In: European Conference on Machine Learning, Springer, pp 280–291