Natural Actor Critic Components and Extensions

Maximilian Gehrke
Tabea Wilke
Yannik Frisch

Group 19 Oleg Arenz

TECHNISCHE UNIVERSITÄT DARMSTADT

Natural Gradient

TECHNISCHE UNIVERSITÄT DARMSTADT

- Optimization problem:

$$\max_{\delta\theta} J(\theta + \delta\theta) \approx J(\theta) + \delta\theta^T \nabla_\theta J(\theta)$$
$$\text{s.t. } \epsilon = D_{KL}(\pi_\theta || \pi_{\theta+\delta\theta}) \approx \tfrac{1}{2}\delta\theta^T F_\theta \delta\theta$$

- Solution:

$$\widetilde{\nabla}_\theta J(\theta) = F_\theta^{-1} \nabla_\theta J(\theta)$$

- Fisher Information Matrix:

$$F_\theta = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^T\right]$$

$\Rightarrow$ *Parametrization invariant, data efficient & fast convergence*

The Natural Actor Critic algorithm

TECHNISCHE UNIVERSITÄT DARMSTADT

Algorithm 1 Episodic Natural Actor Critic (eNAC)

Require: Parameterized policy $\pi_\theta(a|s)$ and it's derivative $\nabla_\theta \log \pi_\theta(a|s)$ with initial parameters $\theta = \theta_0$.

1: **for** $u = 1, 2, 3, \ldots$ **do**
2: **for** $e = 1, 2, 3, \ldots$ **do**
3: **Execute roll-out:** Draw initial state $s_0 \sim p(s_0)$
4: **for** $t = 1, 2, 3, \ldots, N$ **do**
5: Draw action $a_t \sim \pi_{\theta_t}(a_t|s_t)$, observe next state $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$ and reward $r_{t+1} = r(s_t, a_t)$.
6: **end for**
7: **end for**
8: **Critic Evaluation (repeat for each sampled trajectory):** Determine compatible function approximation of advantage function $A(s, a) \approx A_{w_t}(s, a)$.
9: Determine basis functions: $\Phi_e = \left[\sum_{t=0}^{T} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t)^T, 1\right]^T$,
reward statistics: $R_e = \sum_{t=0}^{T} \gamma^t r_t$ and solve $\begin{bmatrix} w_e \\ J \end{bmatrix} = (\Phi_e^T \Phi_e)^{-1} \Phi_e^T R_e$.
Update critic parameters: $w_{t+1} = w_t + \beta w_e$.
10: **Actor Update:** When the natural gradient is converged, $\measuredangle(w_{t+1}, w_t) \leq \epsilon$, update the policy parameters: $\theta_{t+1} = \theta_t + \alpha w_{t+1}$.
11: **end for**

Extensions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Recursive Least Squares
- Fitted NAC + Importance Sampling (FNAC)
- Incremental NAC (INAC)
- Implicit Incremental NAC (I2NAC)
- Regularization

Conclusion & Discussion

TECHNISCHE UNIVERSITÄT DARMSTADT

- NAC has several advantages over vanilla PGM
- Different approaches for:
  - Updating critic (& adjusting learning rate)
  - Fisher inverse
  - Actor update frequency
- Open Questions:
  - Inverse of Fisher Information Matrix (expensive!)
  - NPG estimation might be biased (Thomas, 2014)
  - Application to POMDP's (Jurčíček u. a., 2011)

Sources

TECHNISCHE UNIVERSITÄT DARMSTADT

For publication references please see our paper *"Natural Actor Critic: Components and Extensions"*.

[Jurčíček u. a. 2011] JURČÍČEK, Filip ; THOMSON, Blaise ; YOUNG, Steve: Natural actor and belief critic: Reinforcement algorithm for learning parameters of dialogue systems modelled as POMDPs. In: *ACM Transactions on Speech and Language Processing (TSLP)* 7 (2011), Nr. 3, S. 6

[Thomas 2014] THOMAS, Philip: Bias in natural actor-critic algorithms. In: *International Conference on Machine Learning*, 2014, S. 441–448