Introduction to Gradient-based Direct Policy Search

Infinite-Time Control

Adrien Bolland & Samy Aittahar

17/03/2021

Likelihood ratio policy gradient - Reminder

- Policy gradient for finite-time MDP
- Can be approximate by Monte-Carlo sampling
- On-policy or off-policy
- Can be subject to huge variance

Infinite-time control

- Let $(\mathcal{S}, \mathcal{A}, P_0, p, \rho)$ be an infinite-time MDP
- Similar to the finite time case with $T \to \infty$
- We look for the optimal stationary policy $\pi^* \in \Pi$ such that:

$$\pi^* \in \underset{\pi \in \Pi}{\text{argmax}} \underset{\substack{s_0 \sim P_0(\cdot) \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim p(\cdot|s_t, a_t) \\ r_t \sim \rho(\cdot|s_t, a_t)}}{\mathbb{E}} \{\sum_{t=0}^{\infty} \gamma^t r_t\} \tag{1}$$

Policy gradient theorem

Theorem 1 *(Policy gradient theorem)*
Let $\pi_\theta \in \Pi$ be a parametrized policy and let $J(\theta)$ be its expected cumulative reward, we then have:

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\substack{s_t \sim d^{\pi_\theta}(\cdot) \\ a_t \sim \pi_\theta(\cdot|s_t)}} \{\nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t, a_t)\} , \tag{2}$$

where $d^{\pi_\theta}(s) = \lim_{t\to\infty} \sum_{t=0}^{T-1} \gamma^t P(s|\pi_\theta, t)$ is a discounted weighting of states encountered.

Policy gradient theorem - Practice issues

- We want to perform Monte-Carlo sampling of the expectation
- But how to take into account the infinite horizon?
- And how to compute the Q-value?

Policy gradient theorem - In practice

- In order to deal with the infinite horizon we will fix a maximal length T to the trajectories
- The Q-value will be approximated with an eligible trace
- The most simple eligible trace is the cumulative reward collected

$$\hat{\nabla}_\theta J(\theta) = \left\langle \sum_{t=0}^{T-1} \gamma^t \, \nabla_\theta \log \pi_\theta(a_t|s_t) \, \hat{Q}_t^{\pi_\theta} \right\rangle_n \tag{3}$$

$$\hat{Q}_t^{\pi_\theta} = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} \; . \tag{4}$$

Policy gradient theorem - In practice

- Depending on the problem, it may be necessary to use a large T
- Even if we add a baseline, there may still be a huge variance
- The eligible trace does not really approximate the Q-function if we use scaling... more the advantage
- Sample inefficient
- The eligible trace for $0 << t \underset{\sim}{<} T$ is not very accurate... There are plenty of other eligible traces

Natural policy gradient

- The gradient $\nabla_\theta J(\theta)$ gives the direction of greater increase of the function J for a small vectorial variation $d\theta$
- What does small mean... for a norm $|d\theta| \to 0$

$$\begin{array}{lll} \max\limits_{d\theta} & J(\theta + d\theta) & \\ \text{s.t.} & |d\theta|^2 & = \quad \epsilon^2 \end{array} \tag{5}$$

- How do we compute the norm of a vector in a Euclidean space (with the usual scalar product) in an orthonormal basis?

$$|d\theta|^2 = \sum_{i,j} d\theta_i I_{i,j} d\theta_j = \sum_i d\theta_i d\theta_i \tag{6}$$

- *But how does a parameter change influence the distribution π_θ?*

Natural policy gradient

- Natural gradient are gradients accounting for small variation of the distribution
- Let us change the norm of $d\theta$ such that it accounts for changes in the underlying distribution

$$|d\theta|^2 = \sum_{i,j} d\theta_i F(\theta)_{i,j} d\theta_j \tag{7}$$

$$F(\theta) = \mathop{\mathbb{E}}_{\substack{s_t \sim d^{\pi_\theta}(\cdot) \\ a_t \sim \pi_\theta(\cdot|s)}} \{(\nabla_\theta \log \pi_\theta(a|s))(\nabla_\theta \log \pi_\theta(a|s))^T\} \tag{8}$$

- In fact, we work in a Riemannian space where the manifold is the set of distributions...

Natural policy gradient

We get the natural policy gradient by finding the direction of greater increase of J with the new norm

$$\begin{array}{lll} \max\limits_{d\theta} & J(\theta + d\theta) & \\ \text{s.t.} & |d\theta|^2 & = \quad \epsilon^2 \end{array} \tag{9}$$

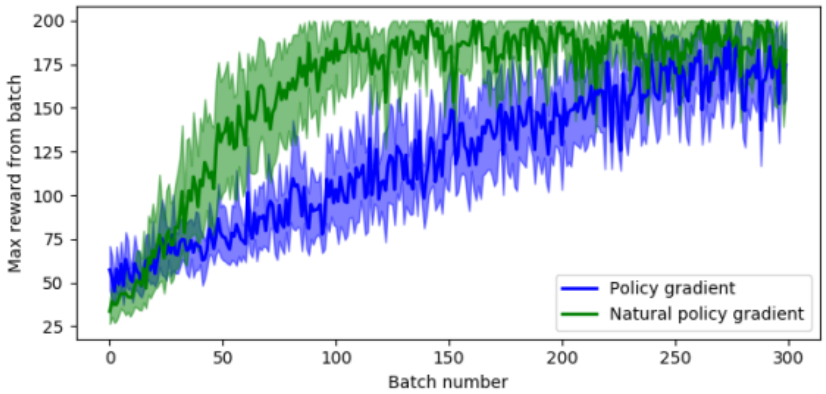
Theorem 2 *(Natural policy gradient)*
The direction of greater increase of J, called the natural gradient $\tilde{\nabla}_\theta J$, is given by:

$$\tilde{\nabla}_\theta J(\theta) = F(\theta)^{-1} \nabla_\theta J(\theta) \,, \tag{10}$$

where $F(\theta)$ is the expectation of the Fisher information matrix of the distribution π_θ.

Natural gradient versus vanilla gradient

(a) Cartpole environment (see detailed description here) (⚠not the original dynamics).



(b) Results over time (expected return, shaded area is 95% confidence interval). See implementation details.

Natural policy gradient - Discussion

- More stable algorithm with less variance
- Nevertheless computing $F(\theta)^{-1}\nabla_\theta J(\theta)$ is expensive !

Trust regions

- Trust region optimization implements a very similar idea to natural policy gradient
- We add an explicit penalization on distance between the new distribution and the previous one
- Typically, penalizing the KL-divergence

$$\begin{array}{lll} \max\limits_{d\theta} & J(\theta + d\theta) & \\ \text{s.t.} & KL(\pi_\theta, \pi_{\theta+d\theta}) & \leq \quad \epsilon \end{array} \tag{11}$$

- The problem now consists in iteratively finding $d\theta$ and updating the policy
- Natural gradients are the closed form of an approximation of this problem...

Reading

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Actor critics

- All methods we have seen relies on the computation of an approximation of the Q-value
- In order to reduce the variance, we have seen that it is common not to exactly use the Q-value as we add a baseline...
- We have focused on using the (scaled) cumulative reward
- A critic is an additional function approximator which we use to estimate the (scaled) Q-function
- What function to learn: Q-function, value function, advantage... and how to train it ?

TD actor critic

- With TD-learning, we can learn online an estimate of the value function V_ϕ
- We then iteratively update the critic and the actor

1. Sample trajectories
2. Update the critic parameters performing one gradient descent step on $\mathcal{L}(\phi)$

$$\mathcal{L}(\phi) = \left\langle \frac{1}{T} \sum_{t=0}^{T-1} (y_t - V_\phi(s_t)) \right\rangle_n \tag{12}$$

$$y_t = r_t + \gamma V_\phi(s_{t+1}) \tag{13}$$

3. Update the actor with using the advantage function as eligible trace

$$A^{\pi_\theta}(s_t, a_t) \approx r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{14}$$

Conclusion

Thank you for your attention !