Advanced Policy-Gradient Algorithms

Adrien Bolland (adrien.bolland@uliege.be)

Outline

Advanced On-Policy Algorithms

Policy Gradient Theorem

Theorem (Policy Gradient Theorem)

For any differentiable policy π_θ, the policy gradient of $J(\pi_\theta)$ is

$$\nabla_\theta J(\pi_\theta) = \mathop{\mathbb{E}}_{\substack{s_0 \sim p_0(\cdot) \\ a_t \sim \pi_\theta(\cdot|s_t) \\ s_{t+1} \sim T(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right].$$

Theorem (Policy Gradient Theorem 2)

For any differentiable policy π_θ, the policy gradient of $J(\pi_\theta)$ is

$$\nabla_\theta J(\pi_\theta) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\gamma,\pi_\theta}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}} \left[Q^{\pi_\theta}(s, a) \nabla_\theta \log \pi_\theta(a|s) \right],$$

where d^{γ,π_θ} is the discounted state visitation probability.

Advantage Actor Critic

Actor update direction:

$$\hat{\nabla}_\theta J(\pi_\theta) = \left\langle \sum_{t=0}^{T-1} \gamma^t \left((\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} + \gamma^T V_\phi(s_T)) - V_\phi(s_t) \right) \nabla_\theta \log \pi_\theta(a_t|s_t) \right\rangle_n .$$

Critic update direction:

$$\hat{\nabla} \mathcal{L}(\phi) = \left\langle \left(\sum_{t=0}^{T-1} V_\phi(s_t) - \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} - \gamma^{T-t} V_\phi(s_T) \right) \left(\sum_{t=0}^{T-1} \nabla_\phi V_\phi(s_t) \right) \right\rangle_n .$$

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 $|d\theta| \to 0$

$$\begin{array}{llll} \max\limits_{d\theta} & J(\theta + d\theta) & & \\ \text{s.t.} & |d\theta|^2 & = & \varepsilon^2 \end{array}$$

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

$$|d\theta|^2 = d\theta^T I d\theta = d\theta^T d\theta$$

But how does a parameter change influence the distribution π_θ?

Natural Policy Gradient

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

$$
\begin{aligned}
|d\theta|_f^2 &= d\theta^T F(\theta)\, d\theta \\
F(\theta) &= \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi_\theta}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}} \left[(\nabla_\theta \log \pi_\theta(a|s))(\nabla_\theta \log \pi_\theta(a|s))^T \right]
\end{aligned}
$$

- 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

$$\begin{array}{ccc} \max_{d\theta} & J(\theta + d\theta) & \\ \text{s.t.} & |d\theta|_f^2 & = \quad \varepsilon^2 \end{array}$$

This optimization problem has a closed form for small ε:

$$d\theta = aF(\theta)^{-1}\nabla_\theta J(\theta)$$
$$a = \frac{\varepsilon}{\sqrt{(\nabla_\theta J(\theta))^T F(\theta)^{-1} \nabla_\theta J(\theta)}} \ .$$

Natural Policy Gradient

Theorem (Natural Policy Gradient)

The natural policy gradient is given by [Kakade, 2001]

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

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

- Natural policy gradient ascent is more stable.
- Nevertheless computing $F(\theta)^{-1} \nabla_\theta J(\theta)$ is expensive !

NPG in practice – Naive approach

Natural policy gradient needs to (1) estimate the (expected) Fisher information matrix and (2) solve a linear system.

- The matrix is estimated based on samples and can be singular or ill-defined...
- Compute the Moore–Penrose (pseudo) inverse with, e.g., singular value decomposition.

$$F(\theta) = U \operatorname{diag}(\sigma)\ V^T$$
$$F(\theta)^{-1} = V \operatorname{diag}(\sigma)^{-1}\ U^T$$

- We can afterwards solve the linear system by matrix multiplication.

As such the method is inefficient and prone to numerical errors.

NPG in practice – Efficient solution 1

Approximate the linear system solution with the conjugate gradient method.

Can be further accelerated in practice, see readings.

NPG in practice – Efficient solution 2

The natural policy gradient can be found by solving directly a least-squared minimization problem, typically by stochastic gradient descent.

Theorem (Natural Policy Gradient)

The natural policy gradient can be computed as

$$\tilde{\nabla}_\theta J(\theta) = \arg\min_w \mathop{\mathbb{E}}_{\substack{s\sim d^{\pi_\theta}(\cdot) \\ a\sim\pi_\theta(\cdot|s)}} \left[\left(w^T(\nabla_\theta \log \pi_\theta(a|s)) - Q^{\pi_\theta}(s,a) \right)^2 \right] .$$

Proof. We write the first-order condition of the problem.

$$\nabla_w \mathop{\mathbb{E}}_{\substack{s\sim d^{\pi_\theta}(\cdot) \\ a\sim\pi_\theta(\cdot|s)}} \left[\left(w^T(\nabla_\theta \log \pi_\theta(a|s)) - Q^{\pi_\theta}(s,a) \right)^2 \right] = 2w^T F(\theta) - 2(\nabla_\theta J(\theta))^T = 0$$

Knowing $F(\theta)$ is symmetric, the condition is satisfied for $w = F(\theta)^{-1}\nabla_\theta J(\theta)$.

Trust Region Methods

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

$$
\begin{array}{lcc}
\max\limits_{d\theta} & J(\theta + d\theta) & \\
\text{s.t.} & \mathbb{E}_{s\sim d^{\pi_\theta}(\cdot)}\left[KL\left(\pi_\theta(\cdot|s), \pi_{\theta+d\theta}(\cdot|s)\right)\right] \quad \leq & \delta
\end{array}
$$

- The problem now consists in iteratively finding $d\theta$ and updating the policy.

Let us approximate the constraint to the second order (for small $d\theta$)

$$D_{KL}(d\theta) = \mathop{\mathbb{E}}_{s \sim d^{\pi_\theta}(\cdot)} \left[KL\left(\pi_\theta(\cdot|s), \pi_{\theta+d\theta}(\cdot|s)\right)\right]$$

$$D_{KL}(d\theta) \underset{Taylor}{=} D_{KL}(d\theta = 0) + d\theta^T \nabla_{d\theta} D_{KL}(d\theta = 0) + \frac{1}{2} d\theta^T \nabla^2_{d\theta} D_{KL}(d\theta = 0)\, d\theta \;.$$

This expression simplifies as:

$$\begin{aligned} D_{KL}(d\theta) &\underset{Taylor}{=} 0 + 0 + \frac{1}{2} d\theta^T F(\theta)\, d\theta \\ &\underset{Taylor}{=} \frac{1}{2} d\theta^T F(\theta)\, d\theta \;. \end{aligned}$$

To the second order, the problem boils down to computing the natural gradient !

Trust Region Policy Optimization

TRPO [Schulman et al., 2015] follows the natural gradient with the largest step respecting the KL-constraint...

$$d\theta = \alpha^j \sqrt{\frac{2\delta}{(\nabla_\theta J(\theta))^T F(\theta)^{-1} \nabla_\theta J(\theta)}} F(\theta)^{-1} \nabla_\theta J(\theta),$$

where α is the step size, δ is an hyperparameter, and j is found by line search.

This algorithm is computationally inefficient... Why ?

Readings

Approximate trust region methods:

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

Generalized method for the critic:

Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2015). High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

Off-Policy Policy Gradient

The algorithms relying on the policy gradient theorem are on-policy... and thus sample inefficient.

Let us change the objective function and maximize

$$J_\beta(\pi_\theta) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim d^{\gamma,\beta}(\cdot)\\ a\sim\pi_\theta(\cdot|s)}} \left[Q^{\pi_\theta}(s,a)\right].$$

Maximizing $J_\beta(\pi_\theta)$ looks like a policy improvement step in policy iteration...

Theorem (Off-Policy Policy Gradient Theorem)

For any differentiable policy π_θ, the off-policy policy gradient direction is [Degris et al., 2012]

$$\nabla_\theta J_\beta(\pi_\theta) \approx \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim d^{\gamma,\beta}(\cdot) \\ a\sim\beta(\cdot|s)}} \left[\frac{\pi_\theta(\cdot|s)}{\beta(\cdot|s)} Q^{\pi_\theta}(s,a) \nabla_\theta \log \pi_\theta(a|s) \right],$$

where $d^{\gamma,\beta}$ is the discounted state visitation probability of the behaviour policy.

For a sufficiently small update step, the return of π_θ is guaranteed to improve.

Reading

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

References

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.