Deterministic Intra-Option Policy Gradient Theorem Technical Report

Riashat Islam
McGill University
Reasoning and Learning Lab
riashat.islam@cs.mcgill.ca

July 1, 2017

1 Deterministic Intra-Option Policy Gradient Theorem

We will consider the determinsitic version of the intra-option policy gradient theorem following work from [1]. We follow similar derivation as in [1] and [2] to derive the deterministic version of the intra-option policy gradient theorem.

Let the deterministic intra-option policy be given be $\mu_{\omega,\theta}$ such that $a = \mu_{\omega,\theta}(s)$. The cumulative reward objective function based on option $J(\mu_{\omega,\theta})$can be written as:

$$J(\mu_{\omega,\theta}) = \int_s \rho(s) V_{\Omega}^{\mu_{\omega,\theta}}(s) ds \tag{1}$$

The gradient of the expected discounted return with respect to the parameter θ of the intra-option policies is therfore:

$$\nabla_\theta J(\mu_{\omega,\theta}) = \int_s \rho(s) \nabla_\theta V_{\Omega}^{\mu_{\omega,\theta}}(s) ds \tag{2}$$

Our goal is to find the gradient of the option value function

$$\nabla_\theta Q_{\Omega}^{\mu_{\omega,\theta}}(s,\omega) \tag{3}$$

In case of deterministic intra-option policies, the following holds:

$$\nabla_\theta Q_{\Omega}^{\mu_{\omega,\theta}}(s,\omega) = \nabla_\theta Q_{U}^{\mu_{\omega,\theta}}(s,\omega,\mu_{\omega,\theta}(s)) \tag{4}$$

We can therefore derive the gradient as follows:

$$\begin{aligned}
\nabla_\theta V_\Omega^{\mu_{\omega,\theta}}(s) &= \nabla_\theta Q_\Omega^{\mu_{\omega,\theta}}(s,\omega) \\
&= \nabla_\theta [r(s,\mu_{\omega,\theta}(s)) + \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s) V^{\mu_{\omega,\theta}}(s') ds'] \\
&= \nabla_\theta r(s,\mu_{\omega,\theta}(s)) + \nabla_\theta \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s) V^{\mu_{\omega,\theta}}(s') ds' \\
&= \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a r(s,a) + \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') + \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a p(s'|s,a) V^{\mu_{\omega,\theta}}(s') ds' \\
&= \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a [r(s,a) + \int_s p(s'|s,a) V^{\mu_{\omega,\theta}}(s') ds'] + \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') ds \\
&= \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_\Omega^{\mu_{\omega,\theta}}(s,\omega) + \int_s \gamma p(s \to s', 1, \mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') ds' \\
&= \nabla_\theta \mu_{\omega,\theta}(s) Q_U^{\mu_{\omega,\theta}}(s,\omega,a) + \int_s \gamma p(s \to s', 1, \mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') ds'
\end{aligned} \tag{5}$$

By considering multiple steps ahead iterating over using the recursive relation, we can therefore write

$$\nabla_\theta V_\Omega^{\mu_{\omega,\theta}}(s) = \int_s \sum_{t=0}^{\infty} \gamma^t p(s \to s', t, \mu_{\omega,\theta}(s') \nabla_\theta \mu_{\omega,\theta}(s') \nabla_a Q_\Omega^{\mu_{\omega,\theta}}(s',\omega) ds' \tag{6}$$

Therefore, the deterministic intra-option policy gradient can be written as:

$$\begin{aligned}
\nabla_\theta J(\mu_{\omega,\theta}) &= \nabla_\theta \int_s \rho(s) V^{\mu_{\omega,\theta}}(s) ds \\
&= \int_s \int_s \sum_{t=0}^{\infty} \gamma^t \rho(s) p(s \to s', t, \mu_{\omega,\theta}) \nabla_\theta \mu_{\omega,\theta}(s') \nabla_a Q_\Omega^{\mu_{\omega,\theta}}(s',\omega) ds' ds \\
&= \int_s \rho^{\omega,\theta}(s,\omega) \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_\Omega(s,\omega) ds
\end{aligned} \tag{7}$$

Since for the deterministic intra-option policies, we will have that:

$$Q_\Omega(s,\omega) = Q_U(s,\omega,\mu_{\omega,\theta}(s)) \tag{8}$$

the final form of the deterministic intra-option policy gradient theorem can therefore be written as :

$$\nabla_\theta J(\mu_{\omega,\theta}) = \int_s \rho^{\omega,\theta}(s,\omega) \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_\Omega(s,\omega) ds \tag{9}$$

$$\nabla_\theta J(\mu_{\omega,\theta}) = \int_s \rho^{\omega,\theta}(s,\omega) \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_U(s,\omega,\mu_{\omega,\theta}(s)) ds \tag{10}$$

References

[1] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. *CoRR*, abs/1609.05140, 2016.

[2] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 387–395, 2014.