Option-Critic in Reproducing Kernel Hilbert Space and Deterministic Intra-Option Policy Gradient Theorem Technical Report

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

February 24, 2017

1 Introduction

In this work, we consider deriving policy gradient theorem for options in the Reproducing Kernel Hilbert Space (RKHS). We extend work from [1] and [2] and consider modelling intra-option policies in MDPs in the vector-valued RKHS. The representation of intra-option policies in RKHS provides the ability to learn complex policies by working non-parametrically in a rich function class. Extending work from [1] and [2], we will develop gradient based intra-option policy optimisation in the RKHS by deriving the functional gradient of the return for our options.

By modelling intra-option policies in the vector valued RKHS, the policy space can be a rich functional class. The intra-option policy gradient theorem is an entire function in the RKHS which is not restricted to any a-priori chosen parameterisation of the class.

In a later section, we show derivation of the existence of the deterministic intra-option policy gradient theorem similar to [3]. We show that similar to the deterministic policy gradients [3], the intra-option deterministic gradient considers the expected gradient of the option-value function. Our hypothesis is that existence of these gradients can outperform their stochastic counterparts especially in continuous control domains such as the MuJoCo simulator.

2 Modelling Intra-Option Policies in RKHS

We will consider intra-option stochastic Gaussian policies parameterised by deterministic functions $h \in H, h : S \rightarrow A \subseteq \mathbb{R}^m$ of the form below. The function $h(.)$ is an element of an RKHS H_K of the form $h(.) = \sum_i K(s,.)\alpha_i \in H_K$

$$\pi_{\omega,h,\Sigma}(a|s) = \frac{1}{Z} e^{-\frac{1}{2}(h(s)-a)^T \Sigma^{-1}(h(s)-a)} \tag{1}$$

We denote the intra-option policy for option w parameterised by function h as $\pi_{\omega,h}$.

The linear paramterisation approach of $h(s)$ assumes that the policy π is parameterised by the parameter space θ and can depend linearly on predefined features $\phi_i(s)$ given by $h(s) = \sum_{i=1}^{d} \theta_i \phi_i(s)$. Similarly, in the non-paramteric case with a reproducing kernel K, we can represent $h(s)$ as $h(s) =< K(s), h >$ based on the reproducing property.

2.1 Intra-Option Policy Gradient Theorem in RKHS

Following work from [1], considering intra-option policies parameterised by θ, the intra-option policy gradient theorem was given by :

$$\nabla_\theta Q_\Omega(s,\omega) = \left(\sum_a \nabla_\theta \pi_{\omega,\theta}(a|s) Q_U(s,\omega,a) \right) + \left(\sum_a \pi_{\omega,\theta}(a|s) \sum_{s'} \gamma P(s'|s,a) \nabla_\theta U(\omega, s') \right) \tag{2}$$

In our work, since we consider intra-option policies parameterised by deterministic functions h, the above equation for intra-option policies in the functional space can be written as :

$$\nabla_h Q_\Omega(s,\omega) = \left(\sum_a \nabla_h \pi_{\omega,h}(a|s) Q_U(s,\omega,a) \right) + \left(\sum_a \pi_{\omega,h}(a|s) \sum_{s'} \gamma P(s'|s,a) \nabla_h U(\omega, s') \right) \tag{3}$$

and we will consider taking the functional derivative, ie, the Frechet derivative of the term $\nabla_h \pi_{\omega,h}(a|s)$ such that for the gradient of the objective functional $\nabla_h U(\pi_{\omega,h}) \in H_K$ the functional gradient update direction is given by:

$$h_{k+1} \leftarrow h_k + \alpha \nabla_h U(\pi_{\omega,h}) \tag{4}$$

Following work from [1], our Intra-Option policy gradient theorem in the RKHS is given by :

$$\sum_{s,\omega} \mu_\Omega(s,\omega|s_0,\omega_0) \sum_a \nabla_h \pi_{\omega,h} Q_U(s,\omega,a) \tag{5}$$

Note that from the above equation, we can write the following:

$$\sum_a \nabla_h \pi_{\omega,h} = \nabla_h \log \pi_{\omega,h} \tag{6}$$

Given the stochastic Gaussian policy, we now have:

$$\log \pi_{\omega,h} = -\log Z - \frac{1}{2}(h(s) - a)^T \Sigma^{-1} (h(s) - a) \tag{7}$$

The functional derivative of the $\log$ policy term can therefore be written, based on the notion of Frechet derivative as :

$$\nabla_h (\log \pi_{\omega,h}) = K(s,.) \Sigma^{-1} (a - h(s)) \in H_K \tag{8}$$

Previous work considering policy search in the RKHS space was also considered by [4]. For more details on using functional gradients, see [5].

The option-critic in RKHS can therefore be written as:

$$\sum_{s,\omega} \mu_\Omega(s,\omega|s_0,\omega_0) K(s,.) \Sigma^{-1}(a-h(s)) Q_U(s,\omega,a) \tag{9}$$

where we can use either a linear or a non-linear function approximator for $Q_U(s,\omega,a)$.

Note that the following derivation might be more useful when considering linear functional approximators of the form $Q^w = w^T \phi(s)$, since [2] also shows, for the policy gradients in RKHS, the existence of the compatible function approximator. However, in case of non-linear function approximators such as DQNs [6], it might not be easily extensible to consider intra-option policy grdaients in the RKHS.

3 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 [3] 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:

$$J(\mu_{\omega,\theta}) = \int_s \rho(s) V_\Omega^{\mu_{\omega,\theta}}(s) ds \tag{10}$$

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

$$\nabla_\theta J(\mu_{\omega,\theta}) = \int_s \rho(s) \nabla_\theta V_\Omega^{\mu_{\omega,\theta}}(s) ds \tag{11}$$

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

$$\nabla_\theta Q_\Omega^{\mu_{\omega,\theta}}(s,\omega) \tag{12}$$

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

$$\nabla_\theta Q_\Omega^{\mu_{\omega,\theta}}(s,\omega) = \nabla_\theta Q_U^{\mu_{\omega,\theta}}(s,\omega,\mu_{\omega,\theta}(s)) \tag{13}$$

We can therefore derive the gradient as follows:

$$
\begin{aligned}
\nabla_\theta V_\Omega^{\mu_{\omega,\theta}}(s) &= \nabla_\theta Q_\Omega^{\mu_{\omega,\theta}}(s,\omega) \\
&= \nabla_\theta [r(s,\mu_{\omega,\theta}(s)) + \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s) V^{\mu_{\omega,\theta}}(s') ds'] \\
&= \nabla_\theta r(s,\mu_{\omega,\theta}(s)) + \nabla_\theta \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s) V^{\mu_{\omega,\theta}}(s') ds' \\
&= \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a r(s,a) + \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') + \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a p(s'|s,a) V^{\mu_{\omega,\theta}}(s') ds' \\
&= \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a [r(s,a) + \int_s p(s'|s,a) V^{\mu_{\omega,\theta}}(s') ds'] + \int_s \gamma p(s'|s,\mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') ds \\
&= \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_\Omega^{\mu_{\omega,\theta}}(s,\omega) + \int_s \gamma p(s \to s', 1, \mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') ds' \\
&= \nabla_\theta \mu_{\omega,\theta}(s) Q_U^{\mu_{\omega,\theta}}(s,\omega,a) + \int_s \gamma p(s \to s', 1, \mu_{\omega,\theta}(s)) \nabla_\theta V^{\mu_{\omega,\theta}}(s') ds'
\end{aligned} \tag{14}
$$

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

$$
\nabla_\theta V_\Omega^{\mu_{\omega,\theta}}(s) = \int_s \sum_{t=0}^{\infty} \gamma^t p(s \to s', t, \mu_{\omega,\theta}(s') \nabla_\theta \mu_{\omega,\theta}(s') \nabla_a Q_\Omega^{\mu_{\omega,\theta}}(s',\omega) ds' \tag{15}
$$

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

$$
\begin{aligned}
\nabla_\theta J(\mu_{\omega,\theta}) &= \nabla_\theta \int_s \rho(s) V^{\mu_{\omega,\theta}}(s) ds \\
&= \int_s \int_s \sum_{t=0}^{\infty} \gamma^t \rho(s) p(s \to s', t, \mu_{\omega,\theta}) \nabla_\theta \mu_{\omega,\theta}(s') \nabla_a Q_\Omega^{\mu_{\omega,\theta}}(s',\omega) ds' ds \\
&= \int_s \rho^{\omega,\theta}(s,\omega) \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_\Omega(s,\omega) ds
\end{aligned} \tag{16}
$$

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

$$
Q_\Omega(s,\omega) = Q_U(s,\omega,\mu_{\omega,\theta}(s)) \tag{17}
$$

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

$$
\nabla_\theta J(\mu_{\omega,\theta}) = \int_s \rho^{\omega,\theta}(s,\omega) \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_\Omega(s,\omega) ds \tag{18}
$$

$$
\nabla_\theta J(\mu_{\omega,\theta}) = \int_s \rho^{\omega,\theta}(s,\omega) \nabla_\theta \mu_{\omega,\theta}(s) \nabla_a Q_U(s,\omega,\mu_{\omega,\theta}(s)) ds \tag{19}
$$

References

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

[2] Guy Lever and Ronnie Stafford. Modelling policies in mdps in reproducing kernel hilbert space. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015.

[3] 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.

[4] Ngo Anh Vien, Peter Englert, and Marc Toussaint. Policy search in reproducing kernel hilbert space. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2089–2096, 2016.

[5] Drew Bagnell. Functional gradient descent lecture notes. pages 2089–2096, 2016.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.