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 $h(.) = \sum_i K(s,.)\alpha_i \in H_K$

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

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

The linear parameterisation approach of $h(s)$ assumes that the policy $\pi$ is parameterised by the parameter space $\theta$ 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 $h(s) = \langle K(s), h \rangle$ based on the reproducing property.

### 2.1 Intra-Option Policy Gradient Theorem in RKHS

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

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

(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:

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

(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:

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

(4)

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

$$
\sum_{s, \omega} \mu_\Omega(s, \omega|s_0, \omega_0) \sum_a \nabla_h \pi_\omega, hQ_U (s, \omega, a)
$$

(5)

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

$$
\sum_a \nabla_h \pi_\omega, h = \nabla_h \log \pi_\omega, h
$$

(6)

Given the stochastic Gaussian policy, we now have:

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

(7)

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

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

(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:

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

(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 gradients in the RKHS.

References


