Computing bounds for kernel–based policy evaluation in reinforcement learning

Raphael Fonteneau* Susan A. Murphy+
Louis Wehenkel* Damien Ernst*

*Department of EECS, University of Liège, Belgium
+Department of Statistics, University of Michigan, USA

2010

Abstract

This technical report proposes an approach for computing bounds on the finite-time return of a policy using kernel-based approximators from a sample of trajectories in a continuous state space and deterministic framework.

1 Introduction

This technical report proposes an approach for computing bounds on the finite-time return of a policy using kernel-based approximators from a sample of trajectories in a continuous state space and deterministic framework. The computation of the bounds is detailed in two different settings. The first setting (Section 3) focuses on the case of a finite action space where policies are open-loop sequences of actions. The second setting (Section 4) considers a normed continuous action space with closed-loop Lipschitz continuous policies.

2 Problem statement

We consider a deterministic discrete-time system whose dynamics over T stages is described by a time-invariant equation:

$$x_{t+1} = f(x_t, u_t) \quad t = 0, 1, \ldots, T-1, \tag{1}$$

where for all t, the state x_t is an element of the continuous normed state space $(\mathcal{X}, \|.\|_{\mathcal{X}})$ and the action u_t is an element of the finite action space $\mathcal{U}$. $T \in \mathbb{N}_0$ is referred to as the optimization horizon. The transition from t to $t+1$ is associated with an instantaneous reward

$$r_t = \rho(x_t, u_t) \in \mathbb{R} \tag{2}$$

where $\rho : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is the reward function. We assume in this technical report that the reward function is bounded by a constant $A_\rho > 0$:

Assumption 2.1

$$\exists A_\rho > 0 : \forall (x,u) \in \mathcal{X} \times \mathcal{U}, |\rho(x,u))| \leq A_\rho \ . \tag{3}$$

The system dynamics f and the reward function ρ are unknown. An arbitrary set of one-step system transitions

$$\mathcal{F} = \{(x^l, u^l, r^l, y^l)\}_{l=1}^n \tag{4}$$

is known, where each transition is such that

$$y^l = f(x^l, u^l) \tag{5}$$

and

$$r^l = \rho(x^l, u^l) \tag{6}$$

Given an initial state $x_0 \in \mathcal{X}$ and a sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, the $T-$stage return $J^{u_0,\ldots,u_{T-1}}(x_0)$ of the sequence $(u_0, \ldots, u_{T-1})$ is defined as follows.

Definition 2.2 ($T-$stage return of the sequence $(u_0, \ldots, u_{T-1})$)
$\forall x_0 \in \mathcal{X}, \forall (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T,$

$$J^{u_0,\ldots,u_{T-1}}(x_0) = \sum_{t=0}^{T-1} \rho(x_t, u_t) \ .$$

In this technical report, the goal is to compute bounds on $J^{u_0,\ldots,u_{T-1}}(x_0)$ using kernel-based approximators. We first consider a finite action space with open-loop sequences of actions in Section 3. In Section 4, we consider a continuous normed action space where the sequences of actions are chosen according to a closed-loop control policy.

3 Finite action space and open-loop control policy

In this section, we assume a finite action space $\mathcal{U}$. We consider open-loop sequences of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, u_t being the action taken at time $t \in \{0, \ldots, T-1\}$. We assume that the dynamics f and the reward function ρ are Lipschitz continuous:

Assumption 3.1 (Lipschitz continuity of f and ρ)
$\exists L_f, L_\rho \in \mathbb{R} : \forall (x, x') \in \mathcal{X}^2, \forall u \in \mathcal{U}, \forall t \in \{0, \ldots, T-1\},$

$$\|f(x,u) - f(x',u)\|_\mathcal{X} \leq L_f \|x - x'\|_\mathcal{X} \ , \tag{7}$$
$$|\rho(x,u) - \rho(x',u)| \leq L_\rho \|x - x'\|_\mathcal{X} \ , \tag{8}$$

We further assume that two constants L_f and L_ρ satisfying the above-written inequalities are known.

Under these assumptions, we want to compute for an arbitrary initial state $x_0 \in \mathcal{X}$ of the system some bounds on the $T-$stage return of any sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$.

3.1 Kernel-based policy evaluation

Given a state $x \in \mathcal{X}$, we introduce the $(T-t)-$stage return of a sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ as follows:

Definition 3.2 ($(T-t)-$stage return of a sequence of actions $(u_0, \ldots, u_{T-1})$)
Let $x \in \mathcal{X}$. For $t' \in \{T-t, \ldots, T-1\}$, we denote by $x_{t'+1}$ the state

$$x_{t'+1} = f(x_{t'}, u_{t'}) \tag{9}$$

with $x_{T-t} = x$. The $(T-t)-$stage return of the sequence $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ when starting from $x \in \mathcal{X}$ is defined as

$$J_{T-t}^{u_0,\ldots,u_{T-1}}(x) = \sum_{t'=T-t}^{T-1} \rho(x_{t'}, u_{t'}) \ . \tag{10}$$

The $T-$stage return of the sequence $(u_0, \ldots, u_{T-1})$ is thus given by

$$J^{u_0,\ldots,u_{T-1}}(x) = J_T^{u_0,\ldots,u_{T-1}}(x) \ . \tag{11}$$

We propose to approximate the sequence of mappings $\left(J_{T-t}^{u_0,\ldots,u_{T-1}}(.)\right)_{t=0}^{T-1}$ using kernels (see [1]) by a sequence $\left(\tilde{J}_{T-t}^{u_0,\ldots,u_{T-1}}(.)\right)_{t=0}^{T-1}$ computed as follows:

$$\forall x \in \mathcal{X}, \tilde{J}_0^{u_0,\ldots,u_{T-1}}(x) = J_0^{u_0,\ldots,u_{T-1}}(x) = 0 \ , \tag{12}$$

and, $\forall x \in \mathcal{X}, \ \forall t \in \{0, \ldots, T-1\}$

$$\tilde{J}_{T-t}^{u_0,\ldots,u_{T-1}}(x) = \sum_{l=1}^{n} \mathbb{I}_{\{u^l = u_t\}} k_l(x) \Big(r^l + \hat{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(y^l) \Big) \ , \tag{13}$$

with

$$k_l(x) = \frac{\Phi\left(\frac{\|x - x^l\|_{\mathcal{X}}}{b}\right)}{\sum_{i=1}^{n} \mathbb{I}_{\{u^i = u_t\}} \Phi\left(\frac{\|x - x^i\|_{\mathcal{X}}}{b}\right)} \ , \tag{14}$$

where $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a univariate non-negative “mother kernel” function, and $b > 0$ is the bandwidth parameter. We also assume that

$$\forall x > 1, \Phi(x) = 0 \ . \tag{15}$$

We suppose that the functions $\{k_l\}_{l=1}^{n}$ are Lipschitz continuous:

Assumption 3.3 (Lipschitz continuity of $\{k_l\}_{l=1}^{n}$)
$\forall l \in \{1, \ldots, n\}, \exists L_{k_l} > 0 :$

$$\forall (x', x'') \in \mathcal{X}^2, \big| k_l(x') - k_l(x'') \big| \leq L_{k_l} \left\| x' - x'' \right\|_{\mathcal{X}} \ . \tag{16}$$

Then, we define L_k such that $L_k = \max_{l \in \{1,\ldots,n\}} L_{k_l}$. The kernel-based estimator (KBE), denoted by $\mathfrak{K}^{u_0,\ldots,u_{T-1}}(x)$, is defined as follows:

Definition 3.4 (Kernel-based estimator)
$\forall x_0 \in \mathcal{X}$,

$$\mathfrak{K}^{u_0,\ldots,u_{T-1}}(x_0) = \tilde{J}_T^{u_0,\ldots,u_{T-1}}(x_0)\ . \tag{17}$$

We introduce the family of kernel operators $\left(K_{T-t}^{u_0,\ldots,u_{T-1}}\right)_{t=0}^{T-1}$ such that

Definition 3.5 (Finite action space kernel operators)
Let $g : \mathcal{X} \to \mathbb{R}$. $\forall t \in \{0,\ldots,T-1\}, \forall x \in \mathcal{X}$,

$$\left(K_{T-t}^{u_0,\ldots,u_{T-1}} \circ g\right)(x) = \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_t\}} k_l(x) \left(r^l + g(y^l)\right)\ . \tag{18}$$

One has

$$\tilde{J}_{T-t}^{u_0,\ldots,u_{T-1}}(x) \quad = \quad \left(K_{T-t}^{u_0,\ldots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}\right)(x)\ . \tag{19}$$

We also introduce the family of finite-horizon Bellman operators $\left(B_{T-t}^{u_0,\ldots,u_{T-1}}\right)_{t=0}^{T-1}$ as follows:

Definition 3.6 (Bellman operators)
Let $g : \mathcal{X} \to \mathbb{R}$. $\forall t \in \{1,\ldots,T\}, \forall x \in \mathcal{X}$,

$$\left(B_{T-t}^{u_0,\ldots,u_{T-1}} \circ g\right)(x) = \rho(x,u_t) + g(f(x,u_t))\ . \tag{20}$$

One has

$$J_{T-t}^{u_0,\ldots,u_{T-1}}(x) \quad = \quad \left(B_{T-t}^{u_0,\ldots,u_{T-1}} \circ J_{T-t-1}^{u_0,\ldots,u_{T-1}}\right)(x)\ . \tag{21}$$

We propose a first lemma that bounds the difference between the two operators $K_{T-t}^{u_0,\ldots,u_{T-1}}$ and $B_{T-t}^{u_0,\ldots,u_{T-1}}$ when applied to the approximated $(T-t-1)-$ return $\tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}$.

Lemma 3.7
$\forall t \in \{0,\ldots,T-1\}, \forall x \in \mathcal{X}$,

$$\left|\left(K_{T-t}^{u_0,\ldots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}\right)(x) - \left(B_{T-t}^{u_0,\ldots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}\right)(x)\right| \leq C_{T-t} b \tag{22}$$

with

$$C_{T-t} = L_\rho + L_k L_f A_\rho (T-t-1)\ . \tag{23}$$

Proof Let $x \in \mathcal{X}$.

- Let $t \in \{0,\ldots,T-2\}$. Since

$$\sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_t\}} k_l(x) = 1, \tag{24}$$

one can write

$$\left|\left(K_{T-t}^{u_0,\ldots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}\right)(x) - \left(B_{T-t}^{u_0,\ldots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}\right)(x)\right|$$

$$= \Bigg| \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_t\}} k_l(x) \Big[r^l - \rho(x,u_t)$$

$$+ \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(y^l) - \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(f(x,u_t)) \Big] \Bigg| \tag{25}$$

$$\leq L_\rho \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_t\}} k_l(x) \|x^l - x\|_{\mathcal{X}}$$

$$+ \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_t\}} \left| k_l(x) \left(\tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(y^l) - \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(f(x,u_t)) \right) \right| \tag{26}$$

On the one hand, since

$$\forall z > 1, \Phi(z) = 0, \tag{27}$$

one has

$$\|x^l - x\|_{\mathcal{X}} \geq b \implies k_l(x) = 0. \tag{28}$$

Thus,

$$L_\rho \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_t\}} k_l(x) \|x^l - x\|_{\mathcal{X}} \leq L_\rho b\,. \tag{29}$$

On the other hand, one has

$$\tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(y^l) - \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(f(x,u_t))$$

$$= \sum_{j=1}^{n} \mathbb{I}_{\{u^j=u_{t+1}\}} \Big[k_j(y^l) - k_j(f(x,u_t)) \Big] (r^j + \tilde{J}_{T-t-2}^{u_0,\ldots,u_{T-1}}(y^j)) \tag{30}$$

Since the reward function ρ is bounded by A_ρ, one can write

$$\left| (r^j + \tilde{J}_{T-t-2}^{u_0,\ldots,u_{T-1}}(y^j)) \right| \leq (T-t-1) A_\rho\,. \tag{31}$$

and according to the Lipschitz continuity of k_j and f, one has

$$\begin{aligned} \left| k_j(y^l) - k_j(f(x,u_t)) \right| &\leq L_{k_j} \|y^l - f(x,u_t)\|_{\mathcal{X}} && (32)\\ &\leq L_k \|y^l - f(x,u_t)\|_{\mathcal{X}} && (33)\\ &\leq L_k L_f \|x^l - x\|_{\mathcal{X}}\,. && (34) \end{aligned}$$

Equations (30), (31) and (34) allow to write

$$\left| \left(\tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(y^l) - \tilde{J}_{T-t-1}^{u_0,\ldots,u_{T-1}}(f(x,u_t)) \right) \right|$$

$$\leq L_k L_f (T-t-1) A_\rho \|x^l - x\|_{\mathcal{X}}\,. \tag{35}$$

Equations (28) and (35) give

$$\left|\left(\tilde{J}^{u_0,\ldots,u_{T-1}}_{T-t-1}(y^l) - \tilde{J}^{u_0,\ldots,u_{T-1}}_{T-t-1}(f(x,u_t))\right)\right| \leq L_k L_f (T-t-1) A_\rho b \tag{36}$$

and since

$$\sum_{l=1}^{n} \mathbb{I}_{u^l=u_t} k_l(x) = 1 \ , \tag{37}$$

one has

$$\sum_{l=1}^{n} \mathbb{I}_{u^l=u_t} \left\| k_l(x)(\tilde{J}^{u_0,\ldots,u_{T-1}}_{T-t-1}(y^l) - \tilde{J}^{u_0,\ldots,u_{T-1}}_{T-t-1}(f(x,u_t)))\right\| \\ \leq L_k L_f b (T-t-1) A_\rho \tag{38}$$

Using Equations (26), (29) and (38), we can finally write
$\forall (x,t) \in \mathcal{X} \times \{0,\ldots,T-2\}$,

$$\left|K^{u_0,\ldots,u_{T-1}}_{T-t} \circ \tilde{J}^{u_0,\ldots,u_{T-1}}_{T-t-1}(x) - B^{u_0,\ldots,u_{T-1}}_{T-t} \circ \tilde{J}^{u_0,\ldots,u_{T-1}}_{T-t-1}(x)\right| \\ \leq (L_\rho + L_k L_f (T-t-1) A_\rho) b \ , \tag{39}$$

which proves the lemma for $t \in \{0,\ldots,T-2\}$.

- Let $t = T-1$. One has

$$\left|\left(K^{u_0,\ldots,u_{T-1}}_{1} \circ \tilde{J}^{u_0,\ldots,u_{T-1}}_{0}\right)(x) - \left(B^{u_0,\ldots,u_{T-1}}_{1} \circ \tilde{J}^{u_0,\ldots,u_{T-1}}_{0}\right)(x)\right|$$

$$\leq \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_{T-1}\}} k_l(x) \left| r^l - \rho(x,u_t)\right| \tag{40}$$

$$\leq \sum_{l=1}^{n} \mathbb{I}_{\{u^l=u_{T-1}\}} k_l(x) L_\rho \|x - x^l\| \leq L_\rho b \ , \tag{41}$$

since

$$\|x - x^l\| \geq b \implies k_l(x) = 0 \tag{42}$$

and

$$\sum_{l=1}^{n} \mathbb{I}_{u^l=u_t} k_l(x) = 1. \tag{43}$$

This shows that Equation (39) is also valid for $t = T-1$, and ends the proof. ■

Then, we have the following theorem.

Theorem 3.8 (Bounds on the actual return of a sequence $(u_0,\ldots,u_{T-1})$**)**
Let $x_0 \in \mathcal{X}$ *be a given initial state. Then,*

$$|\mathfrak{K}^{u_0,\ldots,u_{T-1}}(x_0) - J^{u_0,\ldots,u_{T-1}}(x_0)| \leq \beta b \ , \tag{44}$$

with

$$\beta = \sum_{t=0}^{T-1} C_{T-t} \ . \tag{45}$$

Proof We use the notation $x_{t+1} = f(x_t, u_t),\ \forall t \in \{0, \dots, T-1\}$. One has

$$J_T^{u_0,\dots,u_{T-1}}(x_0) - \tilde{J}_T^{u_0,\dots,u_{T-1}}(x_0)$$
$$= B_T^{u_0,\dots,u_{T-1}} \circ J_{T-1}^{u_0,\dots,u_{T-1}}(x_0) - K_T^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-1}^{u_0,\dots,u_{T-1}}(x_0) \tag{46}$$
$$= B_T^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-1}^{u_0,\dots,u_{T-1}}(x_0) - K_T^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-1}^{u_0,\dots,u_{T-1}}(x_0)$$
$$+ B_T^{u_0,\dots,u_{T-1}} J_{T-t-1}^{u_0,\dots,u_{T-1}}(x_0) - B_T^{u_0,\dots u_{T-1}} \tilde{J}_{T-t-1}^{u_0,\dots,u_{T-1}}(x_0) \tag{47}$$
$$= B_T^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-1}^{u_0,\dots,u_{T-1}}(x_0) - K_T^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-1}^{u_0,\dots,u_{T-1}}(x_0)$$
$$+ J_{T-1}^{u_0,\dots,u_{T-1}}(x_1) - \tilde{J}_{T-1}^{u_0,\dots,u_{T-1}}(x_1)\,. \tag{48}$$

Using the recursive form of Equation (48), one has

$$J^{u_0,\dots,u_{T-1}}(x) - \mathfrak{K}^{u_0,\dots,u_{T-1}}(x) = J_T^{u_0,\dots,u_{T-1}}(x) - \tilde{J}_T^{u_0,\dots,u_{T-1}}(x) \tag{49}$$
$$= \sum_{t=0}^{T-1} B_{T-t}^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\dots,u_{T-1}}(x_t) - K_{T-t}^{u_0,\dots,u_{T-1}} \circ \tilde{J}_{T-t-1}^{u_0,\dots,u_{T-1}}(x_t) \tag{50}$$

Equation (50) and Lemma 3.7 allow to write

$$\left| J_T^{u_0,\dots,u_{T-1}}(x_0) - \mathfrak{K}^{u_0,\dots,u_{T-1}}(x_0) \right| \le \sum_{t=0}^{T-1} C_{T-t} b\,, \tag{51}$$

which ends the proof. ∎

4 Continuous action space and closed-loop control policy

In this section, the action space $(\mathcal{U}, \|.\|_{\mathcal{U}})$ is assumed to be continuous and normed. We consider a deterministic time-varying control policy

$$h : \{0, 1, \dots, T-1\} \times X \to U \tag{52}$$

that selects at time t the action u_t based on the current time and the current state ($u_t = h(t, x_t)$). The $T-$stage return of the policy h when starting from x_0 is defined as follows.

Definition 4.1 (**$T-$stage return of the policy h**)
$\forall x_0 \in \mathcal{X}$,

$$J^h(x_0) = \sum_{t=0}^{T-1} \rho(x_t, h(t, x_t)). \tag{53}$$

where

$$x_{t+1} = f(x_t, h(t, x_t)) \quad \forall t \in \{0, \dots, T-1\}\,. \tag{54}$$

We assume that the dynamics f, the reward function ρ and the policy h are Lipschitz continuous:

Assumption 4.2 (Lipschitz continuity of f, ρ and h)
$\exists L_f, L_\rho, L_h \in \mathbb{R} : \forall (x,x') \in X^2, \forall (u,u') \in U^2, \forall t \in \{0,\ldots,T-1\}$,

$$\begin{aligned} \|f(x,u) - f(x',u')\|_{\mathcal{X}} &\leq L_f\big(\|x-x'\|_{\mathcal{X}} + \|u-u'\|_{\mathcal{U}}\big) , &(55)\\ |\rho(x,u) - \rho(x',u')| &\leq L_\rho\big(\|x-x'\|_{\mathcal{X}} + \|u-u'\|_{\mathcal{U}}\big) , &(56)\\ \|h(t,x) - h(t,x')\|_{\mathcal{U}} &\leq L_h\|x-x'\|_{\mathcal{X}} . &(57)\end{aligned}$$

The dynamics and the reward function are unknown, but we assume that three constants L_f, L_ρ, L_h satisfying the above-written inequalities are known. Under those assumptions, we want to compute bounds on the $T-$stage return of a given policy h.

4.1 Kernel-based policy evaluation

Given a state $x \in \mathcal{X}$, we also introduce the $(T-t)-$stage return of a policy h when starting from $x \in \mathcal{X}$ as follows:

Definition 4.3 ($(T-t)-$stage return of a policy h)
Let $x \in \mathcal{X}$. For $t' \in \{t,\ldots,T-1\}$, we denote by $x_{t'+1}$ the state

$$x_{t'+1} = f(x_{t'}, u_{t'}) \tag{58}$$

with

$$u_{t'} = h(t', x_{t'}) \tag{59}$$

and $x_t = x$. The $(T-t)-$stage return of the policy h when starting from x is defined as follows:

$$J^h_{T-t}(x) = \sum_{t'=t}^{T-1} \rho(x_{t'}, u_{t'}) .$$

The stage return of the policy h is thus given by

$$J^h(x_0) = J^h_T(x_0). \tag{60}$$

The sequence of functions $\big(J^h_{T-t}(.)\big)_{t=0}^{T-1}$ is approximated using kernels ([1]) by a sequence $\big(\tilde{J}^h_{T-t}(.)\big)_{t=0}^{T-1}$ computed as follows

$$\forall x \in \mathcal{X}, \tilde{J}^h_0(x) = J^h_0(x) = 0 , \tag{61}$$

and, $\forall x \in \mathcal{X}, \forall t \in \{0,\ldots,T-1\}$,

$$\tilde{J}^h_{T-t}(x) = \sum_{l=1}^{n} k_l(x, h(t,x))\Big(r^l + \tilde{J}^h_{T-t-1}(y^l)\Big) , \tag{62}$$

where $k_l : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is defined as follows:

$$k_l(x,u) = \frac{\Phi\left(\frac{\|x-x^l\|_{\mathcal{X}} + \|u-u^l\|_{\mathcal{U}}}{b}\right)}{\sum_{i=1}^{n} \Phi\left(\frac{\|x-x^i\|_{\mathcal{X}} + \|u-u^i\|_{\mathcal{U}})}{b}\right)} , \tag{63}$$

where $b > 0$ is the bandwidth parameter and $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a univariate non-negative "mother kernel" function. We also assume that

$$\forall x > 1, \Phi(x) = 0 \ , \tag{64}$$

and we suppose that each function k_l is Lipschitz continuous.

Assumption 4.4 (Lipschitz continuity of $\{k_l\}_{l=1}^n$)
$\forall l \in \{1, \ldots, n\}, \exists L_{k_l} > 0 :$

$$\begin{aligned} &\forall (x', x'', u', u'') \in \mathcal{X}^2 \times \mathcal{U}^2, \\ &|k_l(x', u') - k_l(x'', u'')| \leq L_{k_l} \left(\|x' - x''\|_{\mathcal{X}} + \|u' - u''\|_{\mathcal{U}} \right) \ . \end{aligned} \tag{65}$$

We define L_k such that

$$L_k = \max_{l \in \{1,\ldots,n\}} L_{k_l} \ . \tag{66}$$

The kernel-based estimator KBE, denoted by $\mathfrak{K}^h(x_0)$, is defined as follows:

Definition 4.5 (Kernel-based estimator)
$\forall x_0 \in \mathcal{X}$,

$$\mathfrak{K}^h(x_0) = \tilde{J}^h_T(x_0) \ . \tag{67}$$

We introduce the family of kernel operators $\left(K^h_{T-t}\right)_{t=0}^{T-1}$ such that

Definition 4.6 (Continuous action space kernel operators)
Let $g : \mathcal{X} \to \mathbb{R}$. $\forall t \in \{0, \ldots, T-1\}, \forall x \in \mathcal{X}$,

$$\left(K^h_{T-t} \circ g\right)(x) = \sum_{l=1}^{n} k_l(x, h(t, x)) \left(r^l + g(y^l)\right) \ . \tag{68}$$

One has

$$\tilde{J}^h_{T-t}(x) \quad = \quad \left(K^h_{T-t} \circ \tilde{J}^h_{T-t-1}\right)(x) \ . \tag{69}$$

We also introduce the family of finite-horizon Bellman operators $\left(B^h_{T-t}\right)_{t=0}^{T-1}$ as follows:

Definition 4.7 (Continuous Bellman operator)
Let $g : \mathcal{X} \to \mathbb{R}$. $\forall t \in \{1, \ldots, T\}, \forall x \in \mathcal{X}$,

$$\left(B^h_{T-t} \circ g\right)(x) = \rho(x, h(t, x)) + g(f(x, h(t, x))) \ . \tag{70}$$

One has

$$J^h_{T-t}(x) \quad = \quad \left(B^h_{T-t} \circ J^h_{T-t-1}\right)(x) \ . \tag{71}$$

We propose a second lemma that bounds the distance between the two operators K^h_{T-t} and B^h_{T-t} when applied to the approximated $(T-t-1)-$ return $\tilde{J}^h_{T-t-1}$.

Lemma 4.8
$\forall t \in \{1, \ldots, T-1\}, \forall x \in \mathcal{X}$,

$$\left| \left(K^h_{T-t} \circ \tilde{J}^h_{T-t-1}\right)(x) - \left(B^h_{T-t} \circ \tilde{J}^h_{T-t-1}\right)(x) \right| \leq C_{T-t} b \tag{72}$$

with

$$C_{T-t} = L_\rho + L_k L_f A_\rho (1 + L_h)(T - t - 1) \ . \tag{73}$$

Proof Let $x \in \mathcal{X}$.

- Let $t \in \{0, \ldots, T-2\}$. Since

$$\sum_{l=1}^{n} \mathbb{I}_{\{u^l = h(t,x)\}} k_l(x) = 1, \tag{74}$$

one can write

$$\begin{aligned} &\left| \left(K^h_{T-t} \circ \tilde{J}^h_{T-t-1} \right)(x) - \left(B^h_{T-t} \circ \tilde{J}^h_{T-t-1} \right)(x) \right| \\ &= \left| \sum_{l=1}^{n} k_l(x, h(t,x)) \Big[r^l - \rho(x, h(t,x)) \right. \\ &\left. + \tilde{J}^h_{T-t-1}(y^l) - \tilde{J}^h_{T-t-1}(f(x, h(t,x))) \Big] \right| \end{aligned} \tag{75}$$

$$\begin{aligned} &\leq L_\rho \sum_{l=1}^{n} k_l(x, h(t,x)) \left(\|x^l - x\|_{\mathcal{X}} + \|u^l - h(t,x)\|_{\mathcal{U}} \right) \\ &+ \sum_{l=1}^{n} \left| k_l(x, h(t,x)) \left(\tilde{J}^h_{T-t-1}(y^l) - \tilde{J}^h_{T-t-1}(f(x, h(t,x))) \right) \right| \end{aligned} \tag{76}$$

Since

$$\forall z > 1, \Phi(z) = 0, \tag{77}$$

one has

$$\left(\|x^l - x\|_{\mathcal{X}} + \|u^l - h(t,x)\|_{\mathcal{U}} \right) \geq b \implies k_l(x, h(t,x)) = 0 \; . \tag{78}$$

This gives

$$L_\rho \sum_{l=1}^{n} k_l(x, h(t,x)) \left(\|x^l - x\|_{\mathcal{X}} + \|u^l - h(t,x)\|_{\mathcal{U}} \right) \leq L_\rho b \; . \tag{79}$$

On the other hand, one has

$$\begin{aligned} &\tilde{J}^h_{T-t-1}(y^l) - \tilde{J}^h_{T-t-1}(f(x, h(t,x))) = \sum_{j=1}^{n} \Big[k_j(y^l, h(t+1, y^l)) \\ &- k_j(f(x, h(t,x)), h(t+1, f(x, h(t,x)))) \Big] (r^j + \tilde{J}^h_{T-t-2}(y^j)) \end{aligned} \tag{80}$$

Since the reward function ρ is bounded by A_ρ, one can write

$$\left| (r^j + \tilde{J}^h_{T-t-2}(y^j)) \right| \leq (T-t-1) A_\rho \; . \tag{81}$$

and according to the Lipschitz continuity of k_j,f and h, one has

$$\left|k_j(y^l, h(t+1, y^l)) - k_j(f(x, u_t), h(t+1, f(x, h(t,x))))\right|$$
$$\leq L_{k_j}\left(\|y^l - f(x, h(t,x))\|_{\mathcal{X}} + \|h(t+1, y^l) - h(t+1, f(x, h(t,x)))\|_{\mathcal{U}}\right) \tag{82}$$
$$\leq L_k\left(\|y^l - f(x, h(t,x))\|_{\mathcal{X}} + \|h(t+1, y^l) - h(t+1, f(x, h(t,x)))\|_{\mathcal{U}}\right) \tag{83}$$
$$\leq L_k L_f (1 + L_h)\left(\|x^l - x\|_{\mathcal{X}} + \|u^l - h(t,x)\|_{\mathcal{U}}\right) \ . \tag{84}$$

Equations (80), (81) and (84) allow to write

$$\left|\left(\tilde{J}^h_{T-t-1}(y^l) - \tilde{J}^h_{T-t-1}(f(x, u_t))\right)\right|$$
$$\leq L_k L_f (1 + L_h)(T - t - 1)A_\rho\left(\|x^l - x\|_{\mathcal{X}} + \|u^l - h(t,x)\|_{\mathcal{U}}\right) \tag{85}$$

Equations (78) and (85) give

$$\left|\left(\tilde{J}^h_{T-t-1}(y^l) - \tilde{J}^h_{T-t-1}(f(x, h(t,x)))\right)\right|$$
$$\leq L_k L_f (1 + L_h)(T - t - 1)A_\rho b \tag{86}$$

and since

$$\sum_{l=1}^{n} k_l(x, h(t,x)) = 1 \ , \tag{87}$$

$$\sum_{l=1}^{n}\left|k_l(x, h(t,x))(\tilde{J}^h_{T-t-1}(y^l) - \tilde{J}^h_{T-t-1}(f(x, h(t,x))))\right|$$
$$\leq L_k L_f (1 + L_h) b (T - t - 1) A_\rho \tag{88}$$

Using Equations (76), (79) and (88), we can finally write
$\forall (x,t) \in \mathcal{X} \times \{0, \ldots, T-2\}$,

$$\left|\left(K^h_{T-t} \circ \tilde{J}^h_{T-t-1}\right)(x) - \left(B^h_{T-t} \circ \tilde{J}^h_{T-t-1}\right)(x)\right|$$
$$\leq (L_\rho + L_k L_f (1 + L_h)(T - t - 1)A_\rho) b \tag{89}$$

This proves the lemma for $t \in \{0, \ldots, T-2\}$.

- Let $t = T - 1$. One has

$$\left|\left(K^h_1 \circ \tilde{J}^h_0\right)(x) - \left(B^h_1 \circ \tilde{J}^h_0\right)(x)\right|$$
$$\leq \sum_{l=1}^{n} k_l(x, h(T-1, x))\left|r^l - \rho(x, h(T-1, x))\right| \tag{90}$$
$$\leq \sum_{l=1}^{n} k_l(x, h(T-1, x)) L_\rho\left(\|x - x^l\| + \|h(T-1, x) - u^l\|\right) \tag{91}$$
$$\leq L_\rho b \ , \tag{92}$$

since

$$\left(\|x - x^l\| + \|h(T-1,x) - u^l\|_{\mathcal{U}}\right) \geq b \implies k_l(x, h(T-1,x)) = 0 \quad (93)$$

and

$$\sum_{l=1}^{n} k_l(x, h(T-1,x)) = 1. \quad (94)$$

This shows that Equation (89) is also valid for $t = T-1$, and ends the proof.

■

According to the previous lemma, we have the following theorem.

Theorem 4.9 (Bounds on the actual return of h)
Let $x_0 \in \mathcal{X}$ be a given initial state. Then,

$$\left|\mathfrak{K}^h(x_0) - J^h(x_0)\right| \leq \beta b \,, \quad (95)$$

with

$$\beta = \sum_{t=1}^{T} C_{T-t} \,. \quad (96)$$

Proof We use the notation $x_{t+1} = f(x_t, u_t)$ with $u_t = h(t, x_t)$. One has

$$\begin{aligned} J^h_T(x_0) - \tilde{J}^h_T(x_0) &= B^h_{T-1} \circ J^h_{T-1}(x_0) - K^h_{T-1} \circ \tilde{J}^h_{T-1}(x_0) && (97) \\ &= B^h_{T-1} \circ \tilde{J}^h_{T-1}(x_0) - K^h_{T-1} \circ \tilde{J}^h_{T-1}(x_0) && (98) \\ &+ B^h_{T-1} \circ J^h_{T-1}(x_0) - B^h_{T-1} \circ \tilde{J}^h_{T-1}(x_0) \\ &= B^h_{T-1} \circ \tilde{J}^h_{T-1}(x_0) - K^h_{T-1} \circ \tilde{J}^h_{T-1}(x_0) \\ &+ J^h_{T-1}(x_1) - \tilde{J}^h_{T-1}(x_1) && (99) \end{aligned}$$

Using the recursive form of Equation (99), one has

$$\begin{aligned} J^h(x_0) - \mathfrak{K}^h(x_0) &= J^h_T(x_0) - \tilde{J}^h_T(x_0) && (100) \\ &= \sum_{t=0}^{T-1} B^h_{T-t} \circ \tilde{J}^h_{T-t-1}(x_t) - K^h_{T-t} \circ \tilde{J}^h_{T-t-1}(x_t) && (101) \end{aligned}$$

Then, according to Lemma 1, we can write

$$\left|J^h_T(x_0) - \mathfrak{K}^h(x_0)\right| \leq \sum_{t=0}^{T-1} C_{T-t} b \,, \quad (102)$$

which ends the proof. ■

Acknowledgements

Raphael Fonteneau acknowledges the financial support of the FRIA (Fund for Research in Industry and Agriculture). Damien Ernst is a research associate of the FRS-FNRS. This report presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.We also acknowledge financial support from NIH grants P50 DA10075 and R01 MH080015. The scientific responsibility rests with its authors.

References

[1] D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002.