Continuous-time Value Function Approximation in Reproducing Kernel Hilbert Spaces

Motoya Ohnishi, Masahiro Yukawa, Mikael Johansson, Masashi Sugiyama
Dept. Electronics and Electrical Engineering, Keio Univ., Japan
School of Electrical Engineering and Computer Science, KTH, Sweden
RIKEN Advanced Intelligence Project, Japan
Dept. Complexity Science and Engineering, Univ. Tokyo, Japan
contact: motoya.ohnishi@riken.jp

1. MOTIVATION and INTRODUCTION

• How can we efficiently tie Control Theory and ML for Physical Systems?

2. Problem Settings

Stochastic Differential Equation:
\[ dx = h(x(t), u(t)) \, dt + \eta(x(t), u(t)) \, dw \]

Value Function:
\[ V^\phi(x) := E_x \left[ \int_0^\infty e^{-\beta t} R^\phi(x(t)) \, dt \right] < \infty \]
\[ \beta \geq 0: \text{Discount Factor}, \quad R^\phi(x(t)) := R(x(t), \phi(t)) : \text{Immediate Cost} \]

Issues:
• Output of \( V^\phi \) is unobservable
• Double-sampling problems

3. Assumptions

Assumption 1:
The state space is compact with nonempty and invariant interior
\[ \mathcal{D}_x(x(t) \in \text{int}(\mathcal{X})) = 1, \forall x \in \text{int}(\mathcal{X}), \forall t \geq 0, \mathcal{X} : \text{State space} \]

Assumption 2:
The state space is compact with nonempty and invariant interior
\[ \mathcal{D}_x(x(t) \in \text{int}(\mathcal{X})) = 1, \forall x \in \text{int}(\mathcal{X}), \forall t \geq 0, \mathcal{X} : \text{State space} \]

• Continuity of sample paths
• Masking property
• All paths remain inside the state space with probability one

5. Virtues of CT formulations

• Under CT formulations, we can constrain immediate control inputs in a computationally efficient way

Example:
• Control-affine Dynamics
\[ \frac{dx}{dt} = f(x) + g(x) u \]
• Quadratic Cost
\[ R(x, u) = Q(x) + u^T M u, M > 0 \]

Lyapunov-based or Barrier-certified policy update
\[ \phi^* = \arg\min_{\phi \in \Phi} \left\{ \int_0^\infty e^{-\beta t} R^\phi(x(t)) \, dt \right\} \rightarrow S(x) \text{ defines affine constraints on the control input} \]

• A smooth control performance, an efficient policy update, no elaborative partitioning of time (Doya, 2000)
• Immune to the choice of time intervals

6. Algorithm

Algorithm: Model-based CT-VF Approximation in RKHSs with Barrier-Certified Policy Update

1. Select an RKHS \( \mathcal{H}_V \) which is supposed to contain \( V^\phi \) as one of its elements.
2. Construct another RKHS \( \mathcal{H}_R \) under one-to-one correspondence to \( \mathcal{H}_V \) through a certain bijective linear operator \( U : \mathcal{H}_V \rightarrow \mathcal{H}_R \).
3. Estimate the immediate cost function \( R^\phi \) in the RKHS \( \mathcal{H}_R \) by kernel-based supervised learning, and return its estimate \( \hat{R}^\phi \).
4. An estimate of the VF \( V^\phi \) is immediately obtained by \( U^{-1}(\hat{R}^\phi) \).

Control Theoretic Analyses:
• By constraining immediate control inputs, forward invariance of the safe set is guaranteed
• Barrier-certified policy will again be Lipschitz continuous

Hamilton-Jacobi-Bellman-Issacs Equation:
\[ \beta V^\phi(x) = \frac{1}{2} \sum_{i,j} \left[ \frac{\partial^2 V^\phi(x)}{\partial x_i \partial x_j} \right] \frac{\partial^2 H^\phi(x)}{\partial x_i \partial x_j} + \frac{\partial V^\phi(x)}{\partial x_i} \frac{\partial H^\phi(x)}{\partial x_i} + R^\phi(x), x \in \text{int}(\mathcal{X}) \]

The one-to-one mapping \( U \) can be obtained by using the HJB equation
\[ \rightarrow \text{Our model-based approach avoids the double-sampling problem!} \]

Physical System
Control Theory

Our Work

Bayesian ML

راكين

Value Function Approximation

<table>
<thead>
<tr>
<th>Stochastic</th>
<th>Deterministic</th>
<th>Static</th>
<th>Dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes ( \mathcal{H}_V )</td>
<td>No ( \mathcal{H}_R )</td>
<td>±D</td>
<td>±D</td>
</tr>
<tr>
<td>No ( \mathcal{H}_V )</td>
<td>Yes ( \mathcal{H}_R )</td>
<td>±E</td>
<td>±E</td>
</tr>
</tbody>
</table>

OUR WORK

Continuous-time Dynamics (Differential Equation)

Barrier-certified policy will again be Lipschitz continuous

Physical System
Control Theory

Bayesian ML
 Continuous-time Value Function Approximation in Reproducing Kernel Hilbert Spaces

Motoya Ohnishi, Masahiro Yukawa, Mikael Johansson, Masashi Sugiyama
Dept. Electronics and Electrical Engineering, Keio Univ., Japan
School of Electrical Engineering and Computer Science, KTH, Sweden
RIKEN Advanced Intelligence Project, Japan
Dept. Complexity Science and Engineering, Univ. Tokyo, Japan
contact: motoya.ohnishi@riken.jp

7. Theoretical Analyses

Theorem:

• Assumptions 1 and 2
• $H_V$ is an RKHS associated with the reproducing kernel $k^V(\cdot, \cdot) \in C^{2,2}(X \times X)$.
• (i) $\beta > 0$, or
• (ii) $H_V$ is a Gaussian RKHS, and there exists a point $x_{t \rightarrow \infty} \in \text{int}(X)$ which is stochastically asymptotically stable over int($X$), i.e., $P_x \left( \lim_{t \rightarrow \infty} x(t) = x_{t \rightarrow \infty} \right) = 1$ for any starting point $x \in \text{int}(X)$.

Proposition:

• Assumptions in Theorem
• Control space $U$ defines affine constraints
• $f, g, \alpha$, and the derivative of the barrier function are Lipschitz continuous over $X$.

A barrier-certified updated policy $\phi^+$ is Lipschitz continuous over $X$.

Relations to existing work

RKHS-based:
Capability of learning complicated functions, Nonparametric flexibility

Model-based:
No double-sampling problem, No sample trajectories

Continuous-time:
Immune to the choice of time intervals

Control-theoretic tools:
State constraints can be efficiently taken into account

8. CTGP and CTKF

The reproducing kernel of $H_R$ is available!

9. Policy evaluation – MountainCarContinuous

Cost: $R(x, u) + \epsilon = 1 + 0.001u^2 + \epsilon$ for $x \geq 0.45$

Barrier function: $b(x) = 0.05 + v$

(a) GPTD for $\Delta t = 20.0$
(b) GPTD for $\Delta t = 1.0$
(c) CTGP
(d) DTKF for $\Delta t = 20.0$
(e) DTKF for $\Delta t = 1.0$
(f) CTKF

10. Reinforcement learning

Inverted Pendulum

- When the time interval is too small, discrete-time methods may not work
- For both methods, we need to employ some heuristic approaches to ensure stable policy improvements
  (We observed that greedy policy-updates are unrobust for both CTGP and GPTD)

11. Future work

Employ state-of-the-art kernel methods, Employ actor-critic or other variants to ensure stable policy improvements, GPs for safety verifications