Hilbert Space Embeddings of POMDPs

Yu Nishiyama[1] Abdeslam Boularias[2] Arthur Gretton[2,3] Kenji Fukumizu[1]

[1]The Institute of Statistical Mathematics [2]Max Planck Institute for Intelligent Systems

[3]Gatsby Computational Neuroscience Unit, CSML, UCL

MAX-PLANCK-GESELLSCHAFT MAX-PLANCK-GESELLSCHAFT

1 Overview

A nonparametric approach for policy learning for POMDPs is proposed. The approach represents distributions over the states, observations, and actions as embeddings in feature spaces, which are reproducing kernel Hilbert spaces. Distributions over states given the observations are obtained by applying the kernel Bayes' rule to these distribution embeddings. Policies and value functions are defined on the feature space over states, which leads to a feature space expression for the Bellman equation. Value iteration may then be used to estimate the optimal value function and associated policy. Experimental results confirm that the correct policy is learned using the feature space representation.

2 Background

2.1 Partially Observable Markov Decision Process (POMDP)

A POMDP is a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, Z\rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, T is the transition function with $T(s,a,s') = \Pr(s'|s,a)$ for $s', s \in \mathcal{S}$ and $a \in \mathcal{A}$, R is the reward function with $R(s,a)$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$, $\mathcal{O}$ is the set of observations, and Z is the observation function with $Z(s,o) = \Pr(o|s)$ for $o \in \mathcal{O}$ and $s \in \mathcal{S}$.

2.2 Reinforcement Learning in POMDP Environments

Belief Update Rule:
Belief b is a distribution over $\mathcal{S}$. Given an initial belief b_0 and a history of actions and observations $h_{t+1} = \{a_0, o_1, ..., a_t, o_{t+1}\}$, belief $b_{t+1}(t \geq 0)$ is updated according to Bayes' rule

$$b_{t+1}(s_{t+1}) = \frac{Z(s_{t+1}, o_{t+1}) P(s_{t+1}|a_t; b_t)}{P(o_{t+1}|a_t; b_t)}, \tag{1}$$

where $P(s_{t+1}|a_t; b_t) = \mathbb{E}_{S_t \sim b_t}[T(S_t, a_t, s_{t+1})]$ and $P(o_{t+1}|a_t; b_t) = \mathbb{E}_{S_{t+1} \sim P(\cdot|b_t, a_t)}[Z(o_{t+1}, S_{t+1})]$. In what follows, $b_{t+1}(s_{t+1}) = b^{a,o'}$ when o' is observed after executing action a in belief b.

Goal:
A goal is to find an optimal policy $\pi^* : b \mapsto a$ that maximizes the expected sum of discounted rewards with infinite horizon $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R_t]$, where $\gamma \in (0,1)$ is a discount factor.

Bellman Optimality Equation:
The optimal policy π^* and the value function $V^*(b)$ are the fixed point of the Bellman optimality equation

$$V^*(b) = \max_{a \in \mathcal{A}} Q^*(b,a),\ \pi^*(b) = \arg\max_{a \in \mathcal{A}} Q^*(b,a),\ Q^*(b,a) = \mathbb{E}_{S \sim b}[R(S,a)] + \gamma \mathbb{E}_{O' \sim p(\cdot|b,a)}\left[V^*(b^{a,O'})\right],$$

Value Iteration Algorithm:
The optimal policy π^* and the value function $V^*(b)$ are estimated by the value iteration algorithm $V_d = HV_{d-1}(d \geq 1)$, where H is the Bellman operator and V_d is the d-step value function. H is isotonic and contractive.

Initial Value:
Initial value $V_0(b)$ is set with initial Q-value $Q_0(s,a)$ as $V_0(b) = \max_{a \in \mathcal{A}} \mathbb{E}_{S \sim b(\cdot)}[Q_0(S,a)]$. Examples are reward $Q_0(s,a) = R(s,a)$ and QMDP approximation $Q_0(s,a) = Q^{MDP}(s,a)$ [?], where $Q^{MDP}(s,a)$ is the result of MDP value iteration, approximating the POMDP by an MDP.

2.3 Kernel Method for Probabilities

Mean Embedding:
The mean embedding of distribution P in $\mathcal{H}_{\mathcal{X}}$ is the mean features of P, i.e., the RKHS element $\mu_X = \mathbb{E}_{X \sim P}[k_{\mathcal{X}}(X, \cdot)]$. $\langle \mu_X, f \rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}_{X \sim P}[f(X)]$ holds for all $f \in \mathcal{H}_{\mathcal{X}}$. Empirical estimate $\hat{\mu}_X$ has a form $\hat{\mu}_X = \Upsilon \boldsymbol{\alpha}$, where $\Upsilon = (k_{\mathcal{X}}(\cdot, X_1), \ldots, k_{\mathcal{X}}(\cdot, X_n))$ is a feature matrix and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$ is a weight vector on samples $(X_1, \ldots, X_n)$. $\mathbb{E}_{X \sim P}[f(X)]$ can then be nonparametrically estimated by $\langle \hat{\mu}_X, f \rangle_{\mathcal{H}_{\mathcal{X}}} = \boldsymbol{\alpha}^\top \mathbf{f}$, where $\mathbf{f} = (f(X_1), \ldots, f(X_n))^\top$ is the sample vector of f.

Conditional Embedding Operators & Kernel Bayes' Rule (KBR):
Let $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ be RKHSs associated with $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ over $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, respectively. Let (X,Y) be a random variable taking values on $\mathcal{X} \times \mathcal{Y}$ with distribution P and the density $p(x,y)$. The conditional density functions $\{p(Y|X=x)|x \in \mathcal{X}\}$ define a family of embeddings $\{\mu_{Y|x}\}$ in $\mathcal{H}_{\mathcal{Y}}$. A mapping from $k_{\mathcal{X}}(x,\cdot) \in \mathcal{H}_{\mathcal{X}}$ to $\mu_{Y|x} \in \mathcal{H}_{\mathcal{Y}}$ for all $x \in \mathcal{X}$ can be characterized by conditional embedding operator $\mathcal{U}_{Y|X} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$,

$$\mu_{Y|x} = \mathcal{U}_{Y|X} k_{\mathcal{X}}(x,\cdot) = C_{YX} C_{XX}^{-1} k_{\mathcal{X}}(x,\cdot), \tag{2}$$

where C_{YX} and C_{XX} are uncentred covariance operators with respect to P [?].
Since a posterior distribution is a conditional distribution, the embedding of a posterior can be expressed as a conditional embedding operator [?]. Let Π be a prior distribution with density $\pi(x)$, and $(\bar{X}, \bar{Y})$ be a new random variable with distribution Q and the density $q(x,y) = p(y|x)\pi(x)$. The embedding of a posterior $q(\bar{X}|\bar{Y}=y)$ given y is expressed by a conditional embedding operator $\mathcal{U}_{\bar{X}|\bar{Y}}$

$$\mu_{\bar{X}|y} = \mathcal{U}_{\bar{X}|\bar{Y}} k_{\mathcal{Y}}(y,\cdot) = C_{\bar{X}\bar{Y}} C_{\bar{Y}\bar{Y}}^{-1} k_{\mathcal{Y}}(y,\cdot), \tag{3}$$

where $C_{\bar{X}\bar{Y}}$ and $C_{\bar{Y}\bar{Y}}$ are covariance operators with respect to Q.

3 Kernel POMDP (kPOMDP)

3.1 Kernel Bellman Equations (KBEs)

Let mean embeddings of relevant distributions $b(S)$, $P(S'|a;b)$, $P(O'|a;b)$, $b^{a,o'}(S')$ in the corresponding RKHSs $\mathcal{H}_{\mathcal{S}}$, $\mathcal{H}_{\mathcal{O}}$ be

$$\mu_S = \mathbb{E}_{S \sim b(\cdot)}[\varphi(S)], \qquad \mu_{S'|a;\mu_S} = \mathbb{E}_{S' \sim p(\cdot|a;b)}[\varphi(S')],$$
$$\mu_{O'|a;\mu_S} = \mathbb{E}_{O' \sim p(\cdot|a;b)}[\phi(O')], \qquad \mu_{S'}^{a,o'} = \mathbb{E}_{S' \sim b^{a,o'}(\cdot)}[\varphi(S')].$$

Let $\mathcal{U}_{S'|S,A}$ and $\mathcal{U}_{O|S}$ be conditional embedding operators for transition model T and observation model Z, respectively, $\mathcal{U}_{\bar{S}|\bar{O}}^{(a,\mu_S)}$ be a posterior embedding operator with a prior embedding $\mu_{S'|a;\mu_S}$, corresponding to eq.(3). These embedding operators yield relations

$$\mu_{S'|a;\mu_S} = \mathcal{U}_{S'|S,A} \mu_S \otimes k_A(a,\cdot),\ \ \mu_{O'|a;\mu_S} = \mathcal{U}_{O|S} \mu_{S'|a;\mu_S},\ \ \mu_{S'}^{a,o'} = \mathcal{U}_{\bar{S}|\bar{O}}^{(a,\mu_S)} k_O(o',\cdot). \tag{4}$$

Let $\mathcal{P}_{\mathcal{S}}$ be the set of beliefs and $\mathcal{I}_{\mathcal{S}}$ be the set of embeddings of $\mathcal{P}_{\mathcal{S}}$ in $\mathcal{H}_{\mathcal{S}}$.

Claim 1. *Let $R(\cdot,a) \in \mathcal{H}_{\mathcal{S}}$ and $V^*\left(\mu_{S'}^{a,(\cdot)}\right) \in \mathcal{H}_{\mathcal{O}}$ for all $a \in \mathcal{A}$ and $\mu_S \in \mathcal{I}_{\mathcal{S}}$. The kernel Bellman optimality equations on RKHS $\mathcal{H}_{\mathcal{S}}$ may be*

$$V^*(\mu_S) = \max_{a \in \mathcal{A}} Q^*(\mu_S, a), \quad \pi^*(\mu_S) = \arg\max_{a \in \mathcal{A}} Q^*(\mu_S, a).$$
$$Q^*(\mu_S, a) = \langle \mu_S, R(\cdot,a)\rangle_{\mathcal{H}_{\mathcal{S}}} + \gamma \left\langle \mu_{O'|a;\mu_S}, V^*\left(\mu_{S'}^{a,(\cdot)}\right)\right\rangle_{\mathcal{H}_{\mathcal{O}}},$$

4 Empirical Expression

Training samples are a set of $D_n = \{(\tilde{s}_i, \tilde{o}_i), \tilde{a}_i, \tilde{R}_i, (\tilde{s}'_i, \tilde{o}'_i)\}_{i=1}^n$ according to a POMDP. We assume that the true state samples $\{(\tilde{s}_i, \tilde{s}'_i)\}$ are available for training, but not during the test phase.

Belief Embedding Update Rule:
Empirical estimates $\hat{\mu}_S$, $\hat{\mu}_{O'|a;\hat{\mu}_S}$, $\hat{\mu}_{S'}^{a,o'}$ take respective forms $\hat{\mu}_S = \Upsilon\boldsymbol{\alpha}$, $\hat{\mu}_{O'|a;\mu_S} = \Phi \boldsymbol{\beta}'_{a;\boldsymbol{\alpha}}$, $\hat{\mu}_{S'}^{a,o'} = \Upsilon \boldsymbol{\alpha}'_{a,o'}$. Update rule $\boldsymbol{\alpha} \mapsto \boldsymbol{\beta}'_{a;\boldsymbol{\alpha}}$ is a linear transformation $\boldsymbol{\beta}'_{a;\boldsymbol{\alpha}} = L_{O|S,a}\boldsymbol{\alpha}$ for all $a \in \mathcal{A}$ by $n \times n$ matrix

$$L_{O|S,a} = (G_S + \varepsilon_S n I_n)^{-1} G_{SS'}\left(G_{(S,A)} + \varepsilon_{(S,A)} n I_n\right)^{-1} G_{(S,A)(S,a)} \tag{5}$$

with $G_{SS'} := \Upsilon^\top \Upsilon'$, $G_{(S,A)(S,a)} := D(\mathbf{k}_A(a)) G_S$, and $\mathbf{k}_A(a) = \Psi^\top \psi(a)$. Update rule $\boldsymbol{\beta}'_{a;\boldsymbol{\alpha}} \mapsto \boldsymbol{\alpha}'_{a,o'}$ is a transformation $\boldsymbol{\alpha}'_{a,o'} = R_{S|O}(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}})\mathbf{k}_O(o')$ with a non-negative vector $\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}}$ and $n \times n$ matrix

$$R_{S|O}(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}}) = \left(D(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}}) G_O + \epsilon n I_n\right)^{-1} D(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}}). \tag{6}$$

Claim 2. *Given samples D_n, the empirical expression of the kernel Bellman optimality equation (Claim 1) is*

$$\hat{V}^*(\boldsymbol{\alpha}) = \max_{a \in \mathcal{A}} \hat{Q}^*(\boldsymbol{\alpha}, a),\ \hat{\pi}^*(\boldsymbol{\alpha}) = \arg\max_{a \in \mathcal{A}} \hat{Q}^*(\boldsymbol{\alpha}, a),\ \hat{Q}^*(\boldsymbol{\alpha}, a) = \boldsymbol{\alpha}^\top \boldsymbol{R}_a + \gamma \boldsymbol{\beta}'^{\top}_{a;\boldsymbol{\alpha}} \mathbf{V}^*\left(\boldsymbol{\alpha}'_{a,\mathcal{O}_0}\right),$$

where $\mathbf{R}_a = (R(\tilde{s}_1, a), \ldots, R(\tilde{s}_n, a))^\top \in \mathbb{R}^n$ is the reward vector on samples $\mathcal{S}_0$ for action a and $\hat{\mathbf{V}}^*\left(\boldsymbol{\alpha}'_{a,\mathcal{O}_0}\right) = \left(\hat{V}^*\left(\boldsymbol{\alpha}'_{a,\tilde{o}_1}\right), \ldots, \hat{V}^*\left(\boldsymbol{\alpha}'_{a,\tilde{o}_n}\right)\right)^\top \in \mathbb{R}^n$ is the posterior belief value vector on samples $\mathcal{O}_0$ given action a.

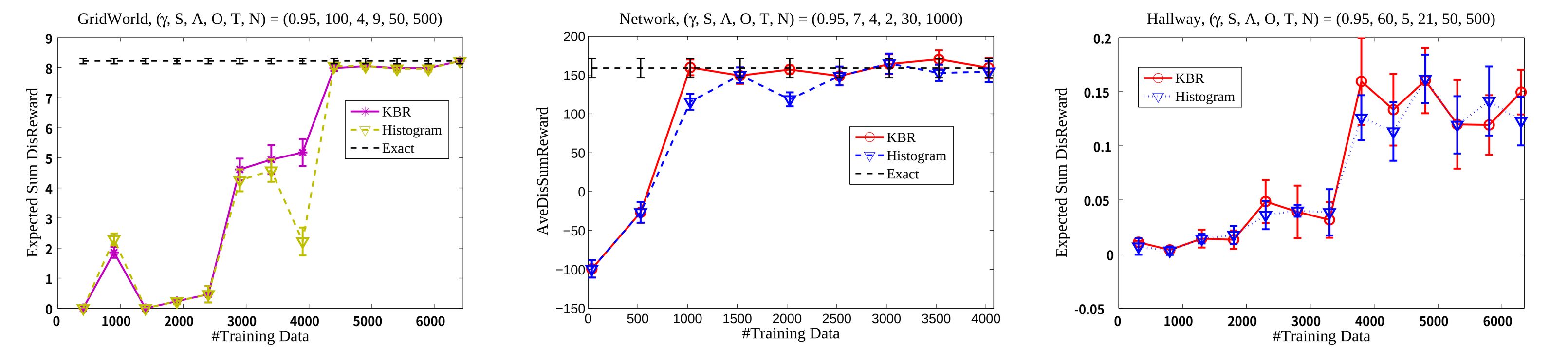
Kernel Value Iteration Algorithm:
$\hat{V}_d = \hat{H}_n \hat{V}_{d-1}(d \geq 1)$, where $\hat{H}_n$ is the kernel Bellman operator

$$(\hat{H}_n V)(\boldsymbol{\alpha}) = \max_{a \in \mathcal{A}}\left[\boldsymbol{\alpha}^\top \boldsymbol{R}_a + \gamma \boldsymbol{\beta}'^{\top}_{a;\boldsymbol{\alpha}} \mathbf{V}\left(\boldsymbol{\alpha}'_{a,\mathcal{O}_0}\right)\right]. \tag{7}$$

$\hat{H}_n$ can be enforced to be isotonic and contractive by replacing above weight vectors with probability vectors $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\alpha}}'_{a,o'}$ as $\hat{w}_i = \frac{\max\{w_i, 0\}}{\sum_{i=1}^n \max\{w_i, 0\}}$ for weights $\boldsymbol{w}$, proposed in [?].

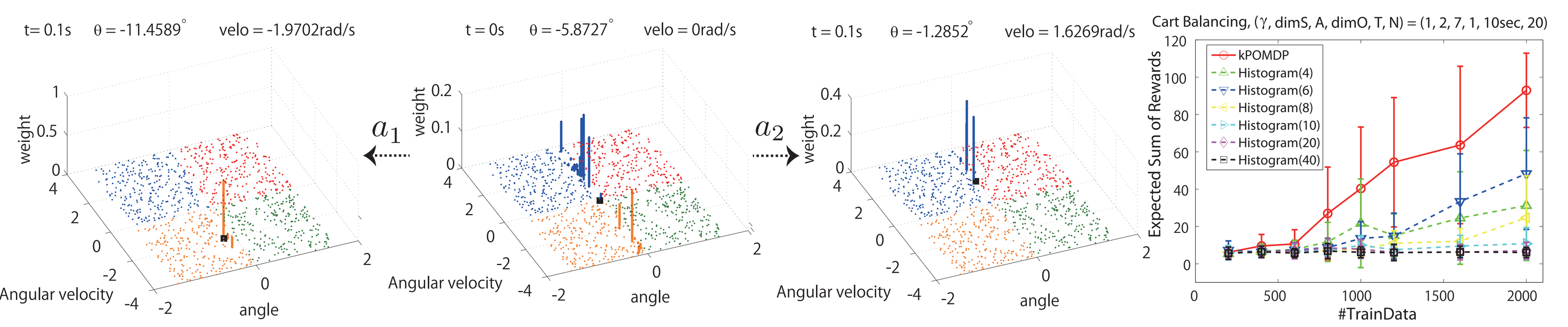
5 Experiments

5.1 Sets $\mathcal{S}$, $\mathcal{O}$, $\mathcal{A}$ are Finite



10×10 Gridworld (naive extension of 4×4 Gridworld), Network, and Hallway1 benchmark problems from left to right. Plots are averaged discounted sum of rewards earned in test experiments (vertical) against the number of training samples n (horizontal). Training samples are collected by uniform random actions. Parameters are within the title (γ, S, A, O, T, N) = (γ, $|\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{O}|$, T, N), where T and N indicates that one episode consists of T steps and results are averaged over N trials. kPOMDP are compared with hitogram methods, in which transition and observation matrices are naively estimated by histograms.

5.2 Sets $\mathcal{S}$, $\mathcal{O}$ are Euclidian spaces $\mathbb{R}^d$



An inverted pendulum problem, where hidden state s is angular and angular velocity $(\theta, \dot{\theta})$, and angular θ is only observed. All the points (colored by RGBY) in the three 3D plots indicate training samples on hidden states $\mathcal{S}$, and z axis indicates belief embedding weights $\hat{\boldsymbol{\alpha}}$ on their samples (after normalization). The true hidden state is marked by the black point in each 3D plot. Since $\dot{\theta}$ is uncertain at the initial point, positive weights spread in the direction of $\dot{\theta}$ axis in the middle 3D figure. The left and right 3D figures show that $\dot{\theta}$ is well estimated by the kPOMDPs dynamics for both of executed actions a_1 and a_2.

UAI 2012, Catalina Island, August 2012