2012年6月15日　統計数理研究所　オープンハウス

# POMDP価値反復アルゴリズムのカーネル化

**西山　悠**　統計的機械学習研究センター　特任研究員

## 1　要旨

POMDP環境の強化学習は，真の状態が観測されない環境下で方策学習を行うものである．方策はベルマン方程式の解で計算される．通常の方策学習では，環境が持つ隠れ状態の遷移則や観測値を生成する観測関数を(パラメトリックな確率分布などで)陽に推定して行われる．近年，カーネル法を使ってサンプルからノンパラメトリックに推論，制御を行うアルゴリズム(e.g., 隠れマルコフモデル [Le Song et al., ICML2009]，完全観測強化学習 [Grunewalder et al., ICML2012])が研究されている．POMDP環境の強化学習のカーネル化を行い，真の状態が観測されない環境下で学習サンプルからノンパラメトリックに方策学習するカーネル価値反復アルゴリズムを提案する．

## 2　POMDP環境の強化学習

部分観測マルコフ決定過程(POMDP)は，組$\langle \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, Z \rangle$で指定される．$\mathcal{S}$：状態集合，$\mathcal{A}$：行動集合，$T(s,a,s')$：状態$s$で行動$a$をとるときの状態$s'$への遷移関数$\Pr(s'|s,a)$, $R(s,a)$：状態$s$で行動$a$をとるときの報酬関数，$\mathcal{O}$：観測集合，$Z(s,o)$：状態$s$で観測値$o$を観測する観測関数$\Pr(o|s)$.

**目的:** 割引期待報酬和$V^\pi(b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{S\sim b_t(\cdot)}\left[R(S,\pi(b_t))\right]\right]$を最大にする方策$\pi : b \mapsto a$を学習．($b$はビリーフ(状態推定分布)).

**ベルマン方程式:** 最適価値関数$V^*$，最適方策$\pi^*$は，ベルマン最適方程式

$$\begin{aligned} V^*(b) &= \max_{a\in\mathcal{A}} Q^*(b,a), \\ Q^*(b,a) &= \mathbb{E}_{S\sim b(\cdot)}\left[R(S,a)\right] + \gamma \mathbb{E}_{O'\sim p(\cdot|a;b)}\left[V^*(b^{a,O'})\right], \\ \pi^*(b) &= \arg\max_{a\in\mathcal{A}} Q^*(b,a) \end{aligned} \tag{1}$$

の解.

## 3　確率分布のカーネル化

**確率分布の埋め込み:** $(\mathcal{X}, \mathcal{B}_\mathcal{X})$上の確率分布$P$の再生核ヒルベルト空間$\mathcal{H}_\mathcal{X}$への埋め込みは，特徴量平均$\mu_X = \mathbb{E}_{X\sim P}\left[k_\mathcal{X}(X,\cdot)\right] \in \mathcal{H}_\mathcal{X}$[Smola et al., ALT2007].

**性質:** 任意の関数$f \in \mathcal{H}_\mathcal{X}$について，期待値$\mathbb{E}_{X\sim P}\left[f(X)\right]$は，確率分布$P$を陽に使わず，埋め込み$\mu_X$と$f$の内積$\langle \mu_X, f\rangle_{\mathcal{H}_\mathcal{X}}$で得られる.

**期待値計算:** $\mu_X$はサンプルの特徴量$\Upsilon = (k_\mathcal{X}(\cdot,X_1),\ldots,k_\mathcal{X}(\cdot,X_n))$の線形結合$\hat{\mu}_X = \Upsilon\boldsymbol{\alpha}$で推定される．期待値はサンプルの関数値を使ってノンパラメトリック推定$\mathbb{E}_{X\sim P}\left[f(X)\right] \sim \boldsymbol{\alpha}^\top \mathbf{f}$.

**カーネルベイズルール (KBR):** サンプルを使い，事前分布埋め込み$\mu_{prior}$から事後分布埋め込み$\mu_{posterior}$を推定 [Fukumizu et al., NIPS2011].

## 4　カーネルPOMDP

POMDPに従う学習サンプル$D_n = \{(\tilde{s}_i,\tilde{o}_i),\tilde{a}_i,\tilde{R}_i,(\tilde{s}'_i,\tilde{o}'_i)\}_{i=1}^n$を使って，ノンパラメトリックに方策を学習する [Nishiyama et al., IBISML2012].

**ビリーフ，予測分布，事後ビリーフの埋め込み:**

$$\hat{\mu}_S = \Upsilon\boldsymbol{\alpha},\ \hat{\mu}_{O'|a;\mu_S} = \Phi\boldsymbol{\beta}'_{a;\boldsymbol{\alpha}},\ \hat{\mu}^{a,o'}_{S'} = \Upsilon\boldsymbol{\alpha}'_{a,o'}$$

**重み推定更新アルゴリズム:**

- 予測分布重み$\boldsymbol{\beta}'_{a;\boldsymbol{\alpha}} = L_{O|S,a}\boldsymbol{\alpha}$

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

- 事後ビリーフ重み$\boldsymbol{\alpha}'_{a,o'} = R_{S|O}(\boldsymbol{\beta}'_{a;\boldsymbol{\alpha}})\mathbf{k}_O(o')$

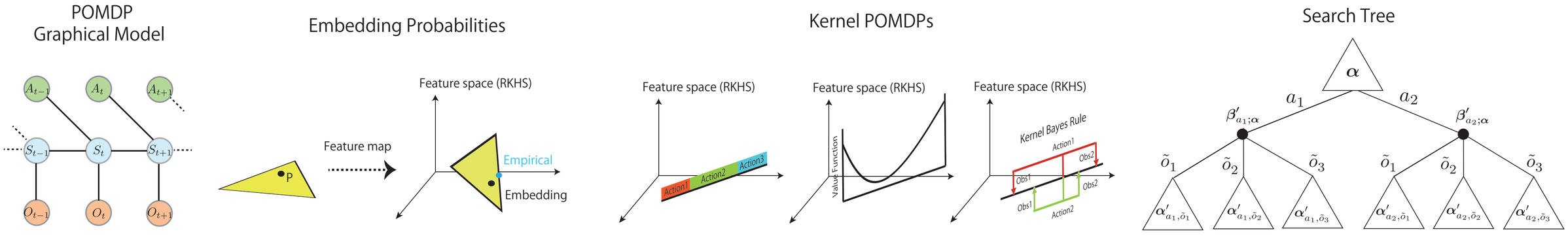
$$R_{S|O}(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}}) = \left(D(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}})G_O + \epsilon n I_n\right)^{-1} D(\hat{\boldsymbol{\beta}}'_{a;\boldsymbol{\alpha}})$$

**行動制御アルゴリズム:**

ベルマン最適方程式(1)のカーネル表現

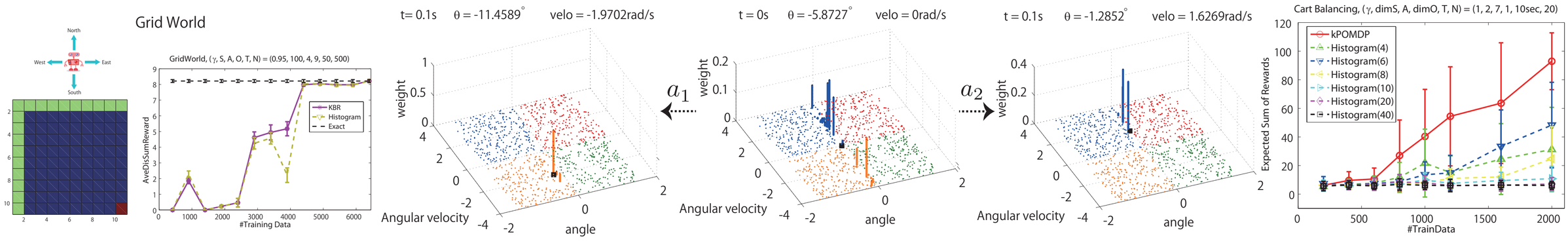
$$\begin{aligned} \hat{V}^*(\boldsymbol{\alpha}) &= \max_{a\in\mathcal{A}} \hat{Q}^*(\boldsymbol{\alpha},a), \\ \hat{Q}^*(\boldsymbol{\alpha},a) &= \boldsymbol{\alpha}^\top \boldsymbol{R}_a + \gamma \boldsymbol{\beta}'^\top_{a;\boldsymbol{\alpha}} \mathbf{V}^*\left(\boldsymbol{\alpha}'_{a,\mathcal{O}_0}\right),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

報酬サンプルベクトル$\boldsymbol{R}_a = (R(\tilde{s}_1,a),\ldots,R(\tilde{s}_n,a))^\top$. 事後ビリーフ価値サンプルベクトル$\hat{\mathbf{V}}^*\left(\boldsymbol{\alpha}'_{a,\mathcal{O}_0}\right) = \left(\hat{V}^*\left(\boldsymbol{\alpha}'_{a,\tilde{o}_1}\right),\ldots,\hat{V}^*\left(\boldsymbol{\alpha}'_{a,\tilde{o}_n}\right)\right)^\top$




## 5　数値実験

$10\times10$ Grid World (左図): 状態数100，観測数9(最近接の壁情報)，行動数5(停止と4方向への移動)．報酬は，ゴール状態での任意の行動で1.他は0.学習サンプルは，環境内での無作為行動．図は横軸：学習サンプル数，縦軸：テストで獲得した平均割引報酬和．ヒストグラム推定(Histogram)と比較．倒立振子 (右図): 状態$s = (\theta,\dot{\theta})$，観測値$o = \theta$．$3D$図はビリーフ埋め込みを表した様子(状態のサンプルとその上の重み(正規化後))．図中□は現在の真の状態．$3D$図(中央)は，観測値$o$が観測されたときのビリーフ埋め込みの初期推定．行動$a$とその後の観測値$o'$に依存して事後ビリーフ埋め込みが更新される．最右図は横軸：学習サンプル数，縦軸：振り子が獲得した累積報酬和(高さ$=\cos\theta$)．ヒストグラム推定と比較.




大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics