非均衡动态博弈：一种智能分层架构

组会报告

谭浚楷

西安交通大学电气学院

2023 年 4 月 22 日

Systems & Control Letters 125 (2019) 59–66

Contents lists available at ScienceDirect

Systems & Control Letters

journal homepage: www.elsevier.com/locate/sysconle

Non-equilibrium dynamic games and cyber–physical security: A cognitive hierarchy approach☆

Aris Kanellopoulos *, Kyriakos G. Vamvoudakis

Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

- 纳什均衡的背景是绝对理性下的理想情形
- 大多数博弈中，都会有人的存在，此时很可能无法达到纳什均衡
- 同时智能体的计算能力有限，在复杂环境下无法获取无穷智能模型

Level-k 模型介绍

Level-k 模型是什么？

Level-k 智力等级的玩家“自我假设”其他所有对手的智力等级都分布在 Level-0 $\sim$ Level-$(k-1)$，且一般服从泊松分布。[1]

实际中智力等级应该是怎样的？

实际中应该有无穷个智力等级，其中最理智的模型（即 $k \to \infty$）对应着的最优控制策略；最一般的模型（即 $k = 0$）对应着的最一般策略 ($u = 0$ 或者 $u = random$)

[1] camerer_cognitive_2004.

1 研究背景

2 研究内容

- Level-*k* 模型介绍
- 零和博弈
- 认知层分层结构
- Model-free Q-learning

3 计划进度

Preliminaries

系统：零和博弈（控制 u+ 干扰 d）

$$\dot{x} = Ax + Bu + Kd \tag{1}$$

Value function：价值函数

$$J(x(0), u, d) = \frac{1}{2}\int_0^\infty \left(x^T Qx + u^T Ru - \gamma^2 \|d\|^2 \right) dt \tag{2}$$

$$V^*(x(t)) = \min_u \max_d \int_t^\infty \frac{1}{2}\left(x^T Qx + u^T Ru - \gamma^2 \|d\|^2 \right) dt \tag{3}$$

Preliminaries

微分后得到 Hamiltatian

$$H\left(x,\frac{\partial V}{\partial x},u,d\right)\equiv x^TQx+u^TRu-\gamma^2\|d\|^2+(\frac{\partial V}{\partial x})^T(Ax+Bu+Kd)$$

极值条件

$$0=\frac{\partial H}{\partial u}\Rightarrow u=-R^{-1}B^T\frac{\partial V}{\partial x}0\quad=\frac{\partial H}{\partial d}\Rightarrow u=\frac{1}{\gamma^2}K^T\frac{\partial V}{\partial x}\qquad(4)$$

1 研究背景

2 研究内容

- Level-k 模型介绍
- 零和博弈
- 认知层分层结构
- Model-free Q-learning

3 计划进度

智能等级

Level-0：初始化选择

Level-0 可以用两种方法模拟：1、认为不存在对手；2、随机行为。本文选择第一种方法

$$V_u^0(x_0) = \min_u \int_0^\infty \left(x^{\mathrm{T}} M x + (u^0)^{\mathrm{T}} R(u^0)\right) \mathrm{d}\tau$$

$$u^0(x) = -R^{-1} B^{\mathrm{T}} \frac{\partial V_u^0(x)}{\partial x} = -R^{-1} B^{\mathrm{T}} P_u^0 x$$

$$V_d^1(x_0) = \max_d \int_0^\infty \left(x^{\mathrm{T}} M x + (u^0)^{\mathrm{T}} R u^0 - \gamma^2 \|d\|^2\right) \mathrm{d}\tau$$

$$d^1(x) = \frac{1}{\gamma^2} K^{\mathrm{T}} \frac{\partial V_d^1(x)}{\partial x} = \frac{1}{\gamma^2} K^{\mathrm{T}} P_d^1 x$$

智能等级

Level-k: 与已得到的 Level-$k-1$ 智能进行博弈获得

$$V_u^k(x_0) = \min_u \int_k^{\infty} \left(x^{\mathrm{T}} M x + (u^k)^{\mathrm{T}} R(u^k) - \gamma^2 \|d^{k-1}\|^2\right) \mathrm{d}\tau$$

$$u^k(x) = -R^{-1} B^{\mathrm{T}} \frac{\partial V_u^k(x)}{\partial x} = -R^{-1} B^{\mathrm{T}} P_u^k x$$

$$V_d^{k+1}(x_0) = \max_d \int_0^{\infty} \left(x^{\mathrm{T}} M x + (u^k)^{\mathrm{T}} R u^k - \gamma^2 \|d^{k+1}\|^2\right) \mathrm{d}\tau$$

$$d^{k+1}(x) = \frac{1}{\gamma^2} K^{\mathrm{T}} \frac{\partial V_d^{k+1}(x)}{\partial x} = \frac{1}{\gamma^2} K^{\mathrm{T}} P_d^{k+1} x$$

Theorem 1

如果满足以下条件，则系统全局渐进稳定

$$
\begin{aligned}
& u^k(0)=0, d^{k+1}(0)=0, \\
& \dot{V}_u^k(\mathrm{x})=-L\left(\mathrm{x}, u^k, d^{k-1}\right)<0, \forall \mathrm{x} \neq 0, \\
& \dot{V}_d^{k+1}(\mathrm{x})=-L\left(\mathrm{x}, u^k, d^{k+1}\right)<0, \forall \mathrm{x} \neq 0, \\
& H_u^k\left(\mathrm{x}, \frac{\partial V_u^k}{\partial \mathrm{x}}, u^k, d^{k-1}\right)=0, \forall \mathrm{x}, \\
& H_d^{k+1}\left(\mathrm{x}, \frac{\partial V_d^{k+1}}{\partial \mathrm{x}} u^k, d^{k+1}\right)=0, \forall \mathrm{x}, \\
& H_u^k\left(\mathrm{x}, \frac{\partial V_u^k}{\partial \mathrm{x}}, u, d^{k-1}\right) \geqslant 0, \forall \mathrm{x}, u \\
& H_d^{k+1}\left(\mathrm{x}, \frac{\partial V_d^{k+1}}{\partial \mathrm{x}}, u^k, d\right) \leqslant 0, \forall \mathrm{x}, d
\end{aligned}
\tag{5}
$$

Q-learning

Q 函数:

$$Q_j^k\left(x, a_j^k\right) := V_j^k(x) + H_j^k\left(x, a_j^k, \nabla V_j^k\right), \forall x, a_j^k, j \in \{u, d\}$$

增广，写成矩阵形式:

$$Q_j^k\left(x, a_j^k\right) = \left(U_j^k\right)^{\mathrm{T}} \left[\begin{array}{cc} Q_{j,\mathrm{xx}}^k & Q_{j,\mathrm{xa}}^k \\ Q_{j,\mathrm{ax}}^k & Q_{j,\mathrm{aa}}^k, \end{array}\right] U_j^k := \left(U_j^k\right)^{\mathrm{T}} \tilde{Q}_j^k U_j^k$$

Q-learning

当 $j := u$ 时

即 $U_u^k = \begin{bmatrix} x^{\mathrm{T}} & \left(u^k\right)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$，此时 $\tilde{Q}$ 矩阵的元素分别为：

$$\begin{aligned}
Q_{u,\mathrm{xx}}^k =& \left(A + \frac{1}{\gamma^2} K K^{\mathrm{T}} P_d^{k-1}\right)^{\mathrm{T}} P_u^k + P_u^k \left(A + \frac{1}{\gamma^2} K K^{\mathrm{T}} P_d^{k-1}\right) \\
&+ \left(M - \frac{1}{\gamma^2} P_d^{k-1} K K^{\mathrm{T}} P_d^{k-1}\right) - P_u^k B R^{-1} B^{\mathrm{T}} P_u^k + P_u^k, && (6) \\
Q_{u,\mathrm{xa}}^k =& B^{\mathrm{T}} P_u^k, && (7) \\
Q_{u,\mathrm{ax}}^k =& B P_u^k, && (8) \\
Q_{u,\mathrm{aa}}^k =& R && (9)
\end{aligned}$$

Q-learning

当 $j := d$ 时

即 $U_d^k = \begin{bmatrix} x^{\mathrm{T}} & (d^k)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$，此时 $\tilde{Q}$ 矩阵的元素分别为：

$$\begin{aligned} Q_{d,\mathrm{xx}}^k =& \left(A - BR^{-1}B^{\mathrm{T}}P_u^k\right)^{\mathrm{T}} P_d^k + P_d^k \left(A - BR^{-1}B^{\mathrm{T}}P_u^k\right) \\ &+ \left(M + P_u^k BR^{-1}B^{\mathrm{T}}P_u^k\right) + \frac{1}{\gamma^2} P_d^k KK^{\mathrm{T}} P_d^k, \end{aligned} \tag{10}$$

$$Q_{d,\mathrm{xa}}^k = K^{\mathrm{T}} P_d^k, \tag{11}$$

$$Q_{d,\mathrm{ax}}^k = K P_d^k, \tag{12}$$

$$Q_{d,\mathrm{aa}}^k = -\gamma^2 \tag{13}$$

Q-learning

最优控制

通过庞特里亚金极小原理 $\frac{\partial Q_j^k(x,a_j^k)}{\partial a_j^k}=0$ 得每个 player 的动作：

$$a_j^k(x)=-\left(Q_{j,\mathrm{aa}}^k\right)^{-1}Q_{j,\mathrm{ax}}^k x,\quad \forall x,j\in\{u,d\}$$

Q-learning

重写 Q 函数

$$Q_j^k\left(x, a_j^k\right) = \mathrm{vech}(\tilde{Q})^{\mathrm{T}}\left(U_j^k \otimes U_j^k\right), \forall x, a_j^k, j \in \{u, d\}$$

NN 逼近

1、逼近 Q 函数：

$$\hat{Q}_j^k\left(x, a_j^k\right) = \left(\hat{W}_j^k\right)^{\mathrm{T}}\left(U_j^k \otimes U_j^k\right), \quad \forall x, a_j^k, j \in \{u, d\}$$

2、逼近控制 u：

$$\hat{a}_j^k = \hat{W}_{a,j}^k x, \quad \forall x, a_j^k, j \in \{u, d\}$$

Frame Title

积分形式 Q 函数

$$
\begin{aligned}
Q_j^k\left(x(t), a_j^k(t)\right) =& Q_j^k\left(x\left(t-T_{\mathrm{IRL}}\right), a_j^k\left(t-T_{\mathrm{IRL}}\right)\right) \\
&- \int_{t-T_{\mathrm{IRL}}}^{t}\left(x^{\mathrm{T}}\bar{M}_j^k x + \left(a_j^k\right)^{\mathrm{T}}\bar{R}_j a_j^k\right)\mathrm{d}\tau, \\
&\forall t \geq 0, j \in \{u, d\}
\end{aligned}
\tag{14, 15}
$$

其中 $T_{\mathrm{IRL}} \in \mathbb{R}^+$ 是采样时长。
对 Player-1： $\bar{M}_u^k := M - \frac{1}{\gamma^2}P_d^{k-1}KK^{\mathrm{T}}P_d^{k-1}$， $\bar{R}_u := R$；
对 Player-2： $\bar{M}_d^k := M + P_u^k BR^{-1}B^{\mathrm{T}}P_u^k$， $\bar{R}_d := -\gamma^2$

Frame Title

定义："当前 Q 函数误差项"

$$\begin{aligned}
e_j^k =& \hat{Q}_j^k\left(x(t), a_j^k(t)\right) - \hat{Q}_j^k\left(x\left(t - T_{\mathrm{IRL}}\right), a_j^k\left(t - T_{\mathrm{IRL}}\right)\right) \\
&+ \int_{t-T_{\mathrm{ILL}}}^{t} \left(x^{\mathrm{T}} \bar{M}_j^k x + \left(a_j^k\right)^{\mathrm{T}} \bar{R}_j a_j^k\right) \mathrm{d}\tau \\
=& \left(\hat{W}_j^k\right)^{\mathrm{T}} \left(U_j^k(t) \otimes U_j^k(t)\right) - \left(\hat{W}_k^k\right)^{\mathrm{T}} \left(U_j^k\left(t - T_{\mathrm{IRL}}\right) \otimes U_j^k\left(t - T_{\mathrm{ILL}}\right)\right) \\
&+ \int_{t-T_{\mathrm{IRL}}}^{t} \left(x^{\mathrm{T}} \bar{M}_j^k x + \left(a_j^k\right)^{\mathrm{T}} \bar{R}_j a_j^k\right) \mathrm{d}\tau,
\end{aligned}$$

定义："控制策略 u 的误差项"

$$e_{j,a}^k = \left(W_{a,j}^k\right)^{\mathrm{T}} x + \left(\hat{Q}_{j,\mathrm{aa}}^k\right)^{-1} \hat{Q}_{j,\mathrm{ax}} x, \forall x, j \in \{u, d\} \tag{16}$$

Q-learning

更新律

定义 Q 函数的误差：$K_1 = \frac{1}{2}\left\|e_j^k\right\|^2$，和控制策略 u 的误差：$K_2 = \frac{1}{2}\left\|e_{j,a}^k\right\|^2$，通过梯度下降得到更新律：

$$\dot{\hat{W}}_j^k = -\alpha \frac{\sigma_j^k}{\left(1+\left(\sigma_j^k\right)^{\mathrm{T}}\sigma_j^k\right)^2}\left(e_j^k\right)^{\mathrm{T}}, \forall t \geq 0, j \in \{u, d\} \qquad (17)$$

$$\dot{\hat{W}}_{j,a}^k = -\alpha_a x\left(e_{j,a}^k\right)^{\mathrm{T}}, \forall t \geq 0, j \in \{u, d\} \qquad (18)$$

其中 $\sigma_j^k = \left(U_j^k(t) \otimes U_j^k(t)\right) - \left(U_j^k\left(t - T_{\mathrm{IRL}}\right) \otimes U_j^k\left(t - T_{\mathrm{IRL}}\right)\right)$

Lemma 1

误差全局渐进稳定

当增益 α 远远大于增益 α_a 时，即满足：

$$1 < \alpha_a < \frac{1}{\delta\bar{\lambda}\left(\bar{R}^{-1}\right)}\left(2\underline{\lambda}\left(\bar{M}_j^k + Q_{j,xa}^k \bar{R}^{-1}\left(Q_{j,xa}^k\right)^T\right) - \bar{\lambda}\left(Q_{j,xa}^k\left(Q_{j,xa}^k\right)^T\right)\right)$$

其中 $\Delta = \frac{\sigma_j^k}{\left(1+\left(\sigma_j^k\right)^T\sigma_j^k\right)^2}$ 在时间 $[t, t+T_{\exp}]$ 内满足激励条件：

$\int_t^{t+T_{\exp}} \Delta\Delta^T d\tau \geq \beta I$

那么 $\psi = \left[x^T\left(\hat{W}_j^k - W_j^k\right)^T\left(\hat{W}_{j,a}^k - W_{j,a}^k\right)^T\right]^T$ 全局渐进稳定

具体证明来自[2]

[2] vamvoudakis_q-learning_2017.

- 研究对象："Human-Machine"非零和协作博弈
- 研究目标：safe+stablisation

Contribution

1、Barrier Transformation：建立障碍转换系统，首先保障安全
2、Bounded Level-k Rationality：先学习机器和人的 k 个智能等级的行为，以及最优控制行为（$k=\infty$）
3、Human Impact Modeling：将人（player-1）的行为利用概率模型进行建模，得到综合行为（模拟人动作）
4、Transfer Learning：基于机器的先验知识，与模拟人的综合行为交互，迭代机器的控制策略（Online Learning），实现 stablisation

Thanks!