Pure Mathematics 理论数学, 2025, 15(2), 218-226
Published Online February 2025 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2025.152062

Hans汉斯

保守策略梯度与策略改进

黄儒泽

青岛大学数学与统计学院，山东 青岛

收稿日期：2025年1月20日；录用日期：2025年2月17日；发布日期：2025年2月28日

摘　要

本文在双人非合作马尔科夫博弈模型下，引入了一种策略度量指标，将保守策略推广到了双智能体情形，给出了一种保守策略梯度和策略改进的条件。这为双人非合作博弈中寻找保守策略下的纳什均衡提供了一定基础和改进方向。

关键词

双人非合作马尔可夫博弈，保守策略，策略梯度，策略改进

Conservative Policy Gradient and Policy Improvement

Ruze Huang

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Jan. 20th, 2025; accepted: Feb. 17th, 2025; published: Feb. 28th, 2025

Abstract

In this paper, a policy metric is introduced under the two-player non-cooperative Markov game model, which generalizes the conservative policy to the two-agent case, and gives a conservative policy gradient and the conditions for policy improvement. This provides a certain foundation and improvement direction for finding Nash equilibrium under policy in two-player non-cooperative game.

Keywords

Two-Player Non-Cooperative Markov Game, Conservative Policy, Policy Gradient, Policy Improvement

文章引用: 黄儒泽. 保守策略梯度与策略改进[J]. 理论数学, 2025, 15(2): 218-226.
DOI: 10.12677/pm.2025.152062

1. 引言

强化学习(RL)已取得了有希望的实证进展，包括掌握围棋博弈[1]，扑克博弈[2]，实时策略博弈[3]，和机器人控制[4]，值得注意的是，其中许多成功都属于多智能体强化学习(MARL)领域。MARL 是关于在共享环境中交互的多个智能体，每个智能体都旨在最大化自己的长期奖励。在学习过程中，每个智能体不仅需要识别环境动态，还需要与其他智能体竞争/合作。

MARL 的一个重要子领域是非合作下的多智能体博弈。随机博弈(SG)已用于对多智能体系统进行建模。然而，大多数现有工作考虑了完全合作的情景[5]或代理之间可以进行特定通信的设置。考虑到自动驾驶汽车等系统及其与行人的交互，很明显，一些多智能体系统是部分合作的，甚至是竞争性的，在许多情况下，智能体之间无法建立通信连接。更重要的是，在这些系统中，智能体选择他们的策略时知道其他智能体也以相同的意识水平选择他们的策略。

另外，大多数工作考虑了非合作的一种特例–零和博弈[6]-[8]，然而在很多场景也存在一些非合作的情况，因此本文考虑了双人非合作马尔科夫博弈模型，在该模型下，给出了价值函数定义，策略度量指标，并给出保守策略梯度和梯度改进条件，为进一步寻找保守策略下的纳什均衡提供了一定的方向。

在双人非合作马尔科夫博弈中，两个智能体通过各自的策略在动态系统中相互作用，其目标是最大化各自的长期累积收益。该模型广泛应用于复杂系统中的决策优化问题，例如市场竞争中的资源分配、智能交通中的车辆协同控制，以及网络对抗中的安全策略制定。

2. 模型描述与预备知识

博弈论可以分为合作博弈和非合作博弈。合作博弈从决策者策略互动可能的结果出发，并规定一个合适的“解”所应满足的性质——特别是对决策者联盟(coalition)具有如“稳定”或“公平”的性质，从而预测博弈结果，决策者之间交流协商，寻求合作。非合作博弈注重从单个决策者的策略选择出发，详细规定各决策者策略互动的过程和影响，一般通过均衡概念把个体最优化决策加总，从而预测博弈结果，决策者之间互不交流单独决策，但对均衡有一致预期。

2.1. 双人非合作马尔可夫博弈

双人非合作马尔科夫博弈是一种在不完全信息和动态环境中进行决策的模型。在这一模型中，两个参与者在每个时间点根据当前状态选择策略，旨在最大化各自的收益。与传统博弈论不同，马尔科夫博弈考虑了状态的转移概率，这使得博弈的结果不仅依赖于当前的策略选择，还受到历史状态的影响。

双人非合作马尔可夫博弈(TNCMG)将单智能体马尔可夫决策过程(MDP)推广到双智能体情况。在本文中，我们考虑了一个有限视野折扣的双人非合作马尔可夫博弈，并介绍了其正式定义，如下所示。

定义 1. 一个有限视野的双人非合作马尔可夫博弈由以下元组定义：

$$\mathcal{M}=\left\{\mathcal{N},\mathcal{S},\mathcal{A},\mathcal{B},T,P,\left\{r^i\right\}_{i\in\mathcal{N}},\gamma\right\}$$

- $\mathcal{N}$：玩家集合 $\{1,2\}$。
- $\mathcal{S}$：两个玩家共享环境的状态集合。

- $\mathcal{A}$：第一个玩家的动作集合。
- $\mathcal{B}$：第二个玩家的动作集合。
- T：视野长度。
- $\{0,1,\cdots,T-1\}$：决策时刻集合，表示必须做出决策的时间点集合。当上下文明确时，我们可以将其写作$[T]$。
- $P=(P_0,P_1,\cdots,P_{T-1})$：转移模型，其中$P_t:\mathcal{S}\times\mathcal{A}\times\mathcal{B}\to\Delta(\mathcal{S}),\forall t\in[T]$表示时间步 t 的(未知的)转移概率矩阵。
- $r^i=\left(r_0^i,r_1^i,\cdots,r_{T-1}^i\right),r_t^i:\mathcal{S}\times\mathcal{A}\times\mathcal{B}\to[0,1],\forall t\in[T]$表示(未知的)确定性奖励函数，它会为第 i 个玩家返回一个标量值。
- $\gamma\in[0,1)$：表示折扣因子，它决定了未来奖励的现值。

本文关注的是状态空间$|\mathcal{S}|$，动作空间$|\mathcal{A}|,|\mathcal{B}|$和时间长度 T 都是有限的情况。在这个背景下，这个博弈可以按顺序描述如下：在每个时间步 t，环境处于状态s_t，两个玩家同时执行他们的动作，这些动作随后都被双方知晓。如果第一个玩家在时间步 t 选择动作a_t，第二个玩家选择b_t，那么两个玩家的联合动作会使环境转移到下一个状态$s_{t+1}\sim P_t(\cdot\mid s_t,a_t,b_t)$，同时环境会为每个玩家确定一个即时奖励$r_t^i(s_t,a_t,b_t)$。

定义 2. 在 TNCMG 中，一个策略由$\pi=(\mu,\nu)$定义，其中

- $\mu:=(\mu_0,\mu_1,\cdots,\mu_{T-1}),\mu_t:\mathcal{S}\to\Delta(\mathcal{A}),\forall t\in[T]$是第一个玩家的策略，其中$\Delta(\mathcal{A})$是动作集$\mathcal{A}$上的概率单纯形。
- $\nu:=(\nu_0,\nu_1,\cdots,\nu_{T-1}),\nu_t:\mathcal{S}\to\Delta(\mathcal{B}),\forall t\in[T]$是第二个玩家的策略，其中$\Delta(\mathcal{B})$是动作集$\mathcal{B}$上的概率单纯形。

其中$\mu_t(\cdot)$和$\nu_t(\cdot)$通常被称为决策规则(decision rules)。

定义 3 [9]. 策略π在马尔可夫过程$\mathcal{M}$及初始状态为s_0的情况下，会生成一个关于序列

$$\tau=\left(s_0,a_0,b_0,r_0^1,r_0^2,\cdots,s_{T-1},a_{T-1},b_{T-1},r_{T-1}^1,r_{T-1}^2,s_T\right)$$

的概率分布：

$$\Pr(\tau\mid\pi,s_0,\mathcal{M}):=\pi_0(a_0,b_0\mid s_0)\prod_{t=0}^{T-2}P_t(s_{t+1}\mid s_t,a_t,b_t)\pi_{t+1}(a_{t+1},b_{t+1}\mid s_{t+1})\times P_{T-1}(s_T\mid s_{T-1},a_{T-1},b_{T-1})$$

其中$\pi_t(a_t,b_t\mid s_t)=\mu_t(a_t\mid s_t)\nu_t(b_t\mid s_t)$。

类似地，对于序列$\tau_t=(s_t,a_t,b_t,\cdots,s_{T-1},a_{T-1},b_{T-1},s_T)$的概率分布可定义为：

$$\Pr(\tau_t\mid\pi,s_t,\mathcal{M}):=\pi_t(a_t,b_t\mid s_t)\prod_{t=t}^{T-2}P_t(s_{t+1}\mid s_t,a_t,b_t)\pi_{t+1}(a_{t+1},b_{t+1}\mid s_{t+1})\times P_{T-1}(s_T\mid s_{T-1},a_{T-1},b_{T-1})$$

定义4. 设$\mathcal{M}$是一个视野长度为T的TNCMG，t是$\mathcal{M}$中的时间步，对于与$\mathcal{M}$相关的策略$\pi=(\mu,\nu)$，第 i 个玩家($i\in\mathcal{N}$)的 t 时刻的状态价值函数$V_{\pi,t,\mathcal{M}}^i(s)$和状态－动作价值函数$Q_{\pi,t,\mathcal{M}}^i(s,a,b)$定义分别如下：

$$V_{\pi,t,\mathcal{M}}^i(s):=(1-\gamma)\mathbb{E}_{(s_t,a_t,b_t,\cdots,s_T)\sim\Pr(\cdot\mid\pi,s_t=s,\mathcal{M})}\left[\sum_{t=t}^{T-1}\gamma^{t-t}r_t^i(s_t,a_t,b_t)\right]\tag{2.1}$$

$$Q_{\pi,t,\mathcal{M}}^i(s,a,b):=(1-\gamma)r_t^i(s,a,b)+\gamma\mathbb{E}_{s'\sim P_t(\cdot\mid s,a,b)}\left[V_{\pi,t+1,\mathcal{M}}^i(s')\right]\tag{2.2}$$

容易验证$V_{\pi,t,\mathcal{M}}^i(s)=\mathbb{E}_{(a,b)\sim\pi_t(\cdot\mid s)}\left[Q_{\pi,t,\mathcal{M}}^i(s,a,b)\right]$。当上下文明确$\mathcal{M}$时，我们可以省略。有关价值函数和根据上述概率分布的事件的更多详细信息，请参阅文献[10]和[11]。

定义 5. 设$\mathcal{M},t,\pi$的定义与上述相同。对于第 i 个玩家，策略π关于状态－动作(s,a,b)在时间步长 t

的优势函数 $A_{\pi,t}^{i}(s,a,b)$ 定义如下：

$$A_{\pi,t}^{i}(s,a,b) \coloneqq Q_{\pi,t}^{i}(s,a,b) - V_{\pi,t}^{i}(s), \quad i \in \mathcal{N}. \tag{2.3}$$

注：基于上述价值函数的定义可以知道，优势函数 $A_{\pi,t}^{i}(s,a,b) \in [0,1]$。此外，它反映了一个动作对相对于其他动作对的相对优势。

2.2. 策略度量指标

对于任意状态分布 $\rho = (\rho_0, \rho_1, \cdots, \rho_{T-1})$，其中 $\rho_t \in \Delta(\mathcal{S})$，我们引入以下定义：

定义 6. 设 $\mathcal{M}, t, \pi$ 的定义与上述相同，对于每个玩家 $i \in \mathcal{N}$，称

$$V_{\pi,t}^{i}(\rho) \coloneqq \mathbb{E}_{s\sim\rho_t}\left[V_{\pi,t}^{i}(s)\right] \tag{2.4}$$

为状态分布 ρ 下的平均价值函数。

如果一个策略被表示为一个函数，则通常使用 $V_{\pi,0}^{i}(\rho)$ 来衡量策略的优劣。该策略度量指标可用于量化和比较不同策略的有效性。通过计算不同策略下的平均价值函数，可以比较各策略的效果，从而选择最优策略。在智能体学习过程中，策略度量指标用于评估和改进学习策略，以最大化长期回报。通过计算平均价值函数，玩家能够更好地理解在特定状态分布下策略的表现，从而做出更明智的决策。总的来说，策略度量指标有助于量化和比较策略效果，为玩家提供决策依据，在博弈论和强化学习中都有广泛应用。因此，本文采用该指标来衡量保守策略的效果。当上下文清楚时，将其简写为 $V_{\pi}^{i}(\rho)$。

3. 保守策略下的策略梯度与策略改进

本节将在[11]的基础上，将单人的保守策略推广到 TNCMG 中，并给出保守策略梯度和这种策略下的策略改进的条件。

3.1. 保守策略

我们的目标是基于分布 ρ 优化策略，并使用性能指标 $V_{\pi}(\rho)$。考虑到我们是在双人非合作博弈，我们希望避免对新策略 π' 进行贪婪更新，因为这可能会导致价值的严重高估。因此，我们通过采用更保守的更新规则来抑制该指标。

$$\begin{aligned} \mu_{\text{new}}(a \mid s, \alpha) &= (1-\alpha)\mu(a \mid s) + \alpha\mu'(a \mid s), \\ \nu_{\text{new}}(b \mid s, \beta) &= (1-\beta)\nu(b \mid s) + \beta\nu'(b \mid s), \end{aligned} \tag{3.1}$$

其中 $\pi' = (\mu', \nu')$，$\alpha, \beta \in [0,1]$。

注：类似于[3]中的方法，TNCMG 在每次迭代中也将策略更新为先前策略和贪婪策略之间的混合分布。

3.2. 策略梯度

下面考虑关于参数 α, β 的策略梯度。

引理 1. 在定义 4 的前提下，对于 $i \in \mathcal{N}$，有

$$V_{\pi,t,\mathcal{M}}^{i}(s) = \mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[Q_{\pi,t,\mathcal{M}}^{i}(s,a,b)\right] \tag{3.2}$$

证明：对于 $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$，有

$$
\begin{aligned}
&\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[Q^i_{\pi,t,\mathcal{M}}(s,a,b)\right]\\
&=\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[(1-\gamma)r^i_t(s,a,b)+\gamma\mathbb{E}_{s'\sim P_t(\cdot|s,a,b)}\left[V^i_{\pi,t+1,\mathcal{M}}(s')\right]\right]\\
&=(1-\gamma)\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]+\gamma\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\mathbb{E}_{s'\sim P_t(\cdot|s,a,b)}\left[V^i_{\pi,t+1,\mathcal{M}}(s')\right]\\
&=\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]+\gamma\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\mathbb{E}_{s'\sim P_t(\cdot|s,a,b)}\left[V^i_{\pi,t+1,\mathcal{M}}(s')\right]-\gamma\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]\\
&=\gamma\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\mathbb{E}_{s'\sim P_t(\cdot|s,a,b)}\left[(1-\gamma)\mathbb{E}_{(s_{t+1},a_{t+1},b_{t+1},\cdots,s_T)\sim\Pr(\cdot|\pi,s_{t+1}=s')}\left(\sum_{i=t+1}^{T-1}\gamma^{i-(t+1)}r^i_t\right)\right]\\
&\quad+\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]-\gamma\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]\\
&=\mathbb{E}_{(s_t,a_t,b_t,\cdots,s_T)\sim\Pr(\cdot|\pi,s_t=s)}\left(\sum_{i=t+1}^{T-1}\gamma^{i-t}r^i_t\right)-\mathbb{E}_{(s_t,a_t,b_t,\cdots,s_T)\sim\Pr(\cdot|\pi,s_t=s)}\left(\sum_{i=t+1}^{T-1}\gamma^{i-t+1}r^i_t\right)\\
&\quad+\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]-\gamma\mathbb{E}_{(a,b)\sim\pi_t(\cdot|s)}\left[r^i_t(s,a,b)\right]\\
&=\mathbb{E}_{(s_t,a_t,b_t,\cdots,s_T)\sim\Pr(\cdot|\pi,s_t=s)}\left(\sum_{i=t}^{T-1}\gamma^{i-t}r^i_t\right)-\gamma\mathbb{E}_{(s_t,a_t,b_t,\cdots,s_T)\sim\Pr(\cdot|\pi,s_t=s)}\left(\sum_{i=t}^{T-1}\gamma^{i-t}r^i_t\right)\\
&=V^i_{\pi,t,\mathcal{M}}(s)
\end{aligned}
$$

定理 1. 设 $\mathcal{M}$ 是一个视野长度为 T 的 TNCMG，t 是 $\mathcal{M}$ 中的时间步，对于与 $\mathcal{M}$ 相关的策略 $\pi=(\mu,\nu)$ 和 $\pi'=(\mu',\nu')$ 是与 $\mathcal{M}$ 有关的策略。对于 $\alpha,\beta\in[0,1]$，在更新策略(3.1)下，有

$$\nabla_\alpha V^1_{\pi_{\text{new}}}(\rho)\big|_{\alpha=0}=\mathbb{A}^1_{(\mu,\nu_{\text{new}})}(\rho,(\mu',\nu_{\text{new}})) \tag{3.3}$$

$$\nabla_\beta V^2_{\pi_{\text{new}}}(\rho)\big|_{\beta=0}=\mathbb{A}^2_{(\mu_{\text{new}},\nu)}(\rho,(\mu_{\text{new}},\nu')) \tag{3.4}$$

其中

$$\mathbb{A}^1_{(\mu,\nu_{\text{new}})}(\rho,(\mu',\nu_{\text{new}}))=\mathbb{E}_{s_0\sim\rho_0}\left\{\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|(\mu,\nu_{\text{new}}),s_0)}\mathbb{E}_{(a,b)\sim(\mu'_t,\nu_{\text{new},t})}\left[A^1_{(\mu,\nu_{\text{new}}),t}(s,a,b)\right]\right\} \tag{3.5}$$

$$\mathbb{A}^2_{(\mu_{\text{new}},\nu)}(\rho,(\mu_{\text{new}},\nu'))=\mathbb{E}_{s_0\sim\rho_0}\left\{\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|(\mu_{\text{new}},\nu),s_0)}\mathbb{E}_{(a,b)\sim(\mu_{\text{new},t},\nu'_t)}\left[A^2_{(\mu_{\text{new}},\nu),t}(s,a,b)\right]\right\} \tag{3.6}$$

证明：首先，根据引理 1，对于任意的 $s\in\mathcal{S},t<T$

$$
\begin{aligned}
&\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\nabla_\alpha V^1_{\pi_{\text{new}},t}(s)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\nabla_\alpha\mathbb{E}_{(a,b)\sim\pi_{\text{new},t}(\cdot|s,\alpha,\beta)}\left[Q^1_{\pi_{\text{new}},t}(s,a,b)\right]\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\left(\nabla_\alpha\pi_{\text{new},t}(a,b|s,\alpha,\beta)\right)Q^1_{\pi_{\text{new}},t}(s,a,b)+\sum_{a,b}\pi_{\text{new},t}(a,b|s,\alpha,\beta)\nabla_\alpha Q^1_{\pi_{\text{new}},t}(s,a,b)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\left(\nabla_\alpha\mu_{\text{new},t}(a|s,\alpha)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{\pi_{\text{new}},t}(s,a,b)\right]\\
&\quad+\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\nu_{\text{new},t}(a|s,\alpha)\nabla_\alpha Q^1_{\pi_{\text{new}},t}(s,a,b)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\left(\nabla_\alpha\mu_{\text{new},t}(a|s,\alpha)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{\pi_{\text{new}},t}(s,a,b)\right]\\
&\quad+\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\mathbb{E}_{(a,b)\sim\pi_{\text{new},t}(\cdot|s,\alpha,\beta)}\left[\nabla_\alpha\left((1-\gamma)r^1_t(s,a,b)+\gamma\mathbb{E}_{s'\sim P_t(\cdot|s,a,b)}V^1_{\pi_{\text{new}},t+1}(s')\right)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\left(\nabla_\alpha\mu_{\text{new},t}(a|s,\alpha)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{\pi_{\text{new}},t}(s,a,b)\right]\\
&\quad+\gamma\mathbb{E}_{s\sim\Pr(s_{t+1}=s|\pi_{\text{new}},s_0)}\left[\nabla_\alpha V^1_{\pi_{\text{new}},t+1}(s)\right]
\end{aligned}
$$

其次，由于 $\nabla_\alpha V^1_{\pi_{\text{new}}}(s_0)=\mathbb{E}_{s\sim\Pr(s_0=s|\pi_{\text{new}},s_0)}\left[\nabla_\alpha V^1_{\pi_{\text{new}}}(s)\right]$，对前面的等式进行递归，并利用 $V_{\pi,T}=0$ 对于所有 π，可以得到

$$\nabla_\alpha V^1_{\pi_{\text{new}}}(s_0)=\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\left(\nabla_\alpha\mu_{\text{new},t}(a|s,\alpha)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{\pi_{\text{new},t}}(s,a,b)\right]$$

根据(2.4)，有

$$\begin{aligned}\nabla_\alpha V^1_{\pi_{\text{new}}}(\rho)&=\mathbb{E}_{s_0\sim\rho_0}\left[\nabla_\alpha V^1_{\text{new}}(s_0)\right]\\&=\mathbb{E}_{s_o\sim\rho_0}\left[\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\sum_{a,b}\left(\nabla_\alpha\mu_{\text{new},t}(a|s,\alpha)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{\pi_{\text{new},t}}(s,a,b)\right]\end{aligned}\tag{3.7}$$

根据更新规则(3.1)，有

$$\begin{aligned}&\sum_{a,b}\left[\left(\nabla_\alpha\mu_{\text{new},t}(a|s,\alpha)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{\pi_{\text{new},t}}(s,a,b)\right]\Big|_{\alpha=0}\\&=\sum_{a,b}\left(\mu_t'(a|s)-\mu_t(a|s)\right)\nu_{\text{new},t}(b|s,\beta)Q^1_{(\mu,\nu_{\text{new}}),t}(s,a,b)\\&=\sum_{a,b}\left(\mu_t'(a|s)-\mu_t(a|s)\right)\nu_{\text{new},t}(b|s,\beta)A^1_{(\mu,\nu_{\text{new}}),t}(s,a,b)\\&=\mathbb{E}_{(a,b)\sim(\mu_t',\nu_{\text{new},t})}\left[A^1_{(\mu,\nu_{\text{new}}),t}(s,a,b)\right]\end{aligned}\tag{3.8}$$

最后，将(3.8)式代入到(3.7)式有

$$\begin{aligned}\nabla_\alpha V^1_{\pi_{\text{new}}}(\rho)\Big|_{\alpha=0}&=\mathbb{E}_{s_0\sim\rho_0}\left\{\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|(\mu,\nu_{\text{new}}),s_0)}\mathbb{E}_{(a,b)\sim(\mu_t',\nu_{\text{new},t})}\left[A^1_{(\mu,\nu_{\text{new}}),t}(s,a,b)\right]\right\}\\&=\mathbb{E}_{s_0\sim\rho_0}\left\{\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|(\mu,\nu_{\text{new}}),s_0)}\mathbb{E}_{(a,b)\sim(\mu_t',\nu_{\text{new},t})}\left[A^1_{(\mu,\nu_{\text{new}}),t}(s,a,b)\right]\right\}\\&=\mathbb{A}^1_{(\mu,\nu_{\text{new}})}\left(\rho,(\mu',\nu_{\text{new}})\right)\end{aligned}\tag{3.9}$$

类似地，可以证明 $\nabla_\beta V^2_{\pi_{\text{new}}}(\rho)\Big|_{\beta=0}=\mathbb{A}^2_{(\mu_{\text{new}},\nu)}\left(\rho,(\mu_{\text{new}},\nu')\right)$，为了篇幅限制，证明放在附录。

注：该定理说明，对于玩家 1 的策略梯度，状态空间上的期望是关于 (μ,ν_{new}) 诱导的，而动作空间上的期望是关于策略 (μ',ν_{new})，对于玩家 2 的策略梯度，状态空间上的期望是关于 (μ_{new},ν) 诱导的，而动作空间上的期望是关于策略 (μ_{new},ν')。

3.3. 策略改进

定理 2. 在定理 1 的设定和保守更新策略(3.1)下，如果满足条件：

$$\mathbb{A}^1_{(\mu,\nu_{\text{new}})}\left(\rho,(\mu',\nu_{\text{new}})\right)>0\quad\text{and}\quad\mathbb{A}^2_{(\mu_{\text{new}},\nu)}\left(\rho,(\mu_{\text{new}},\nu')\right)>0\tag{3.10}$$

则策略 π_{new} 可以在足够小的 α 和 β 下得到改进。

证明：该定理可以由定理 1 直接得出。

注：该定理说明智能体在足够小的 α 和 β 下可以使得自身的价值函数得到提升，在实际应用中我们一般设置一个比较小的平滑值 ε，使得

$$\mathbb{A}^1_{(\mu,\nu_{\text{new}})}\left(\rho,(\mu',\nu_{\text{new}})\right),\mathbb{A}^2_{(\mu_{\text{new}},\nu)}\left(\rho,(\mu_{\text{new}},\nu')\right)\le\varepsilon$$

时停止更新策略。

4. 结论与展望

本文将单智能体下的保守策略方式推广到双智能体，在双人非合作马尔科夫博弈的模型下，引入了策略度量指标，给出了一种保守策略梯度和策略改进条件，给出了双人非合作马尔科夫博弈的一种策略改进方式和方向，为之后寻找纳什均衡奠定了一定的基础和方向，其中关键定理指出，玩家 1 的策略更新在状态空间上取决于策略(μ,ν_{new})，在动作空间上取决于(μ',ν_{new})；玩家 2 的策略更新在状态空间上取决于策略(μ_{new},ν)，在动作空间上取决于(μ_{new},ν')。另外，由于保守策略不易导致价值函数高估，因此未来的工作将考虑在离线强化学习中的应用，探索在离线强化学习中是否保守策略能有效降低价值函数高估问题，并尝试找出达到纳什均衡所需的样本复杂度。

参考文献

[1] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., *et al*. (2016) Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, **529**, 484-489. https://doi.org/10.1038/nature16961

[2] Brown, N. and Sandholm, T. (2017) Libratus: The Superhuman AI for No-Limit Poker. *Proceedings of the* 26*th International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 5226-5228. https://doi.org/10.24963/ijcai.2017/772

[3] Vinyals, O., Babuschkin, I., Chung, J., *et al*. (2019) Alphastar: Mastering the Real-Time Strategy Game Starcraft II. DeepMind Blog, 2.

[4] Kober, J., Bagnell, J.A. and Peters, J. (2013) Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research*, **32**, 1238-1274. https://doi.org/10.1177/0278364913495721

[5] Wei, E. and Luke, S. (2016) Lenient Learning in Independent-Learner Stochastic Cooperative Games. *Journal of Machine Learning Research*, **17**, 1-42.

[6] Cui, Q.W. and Du, S.S. (2022) When Are Offline Two-Player Zero-Sum Markov Games Solvable? 36*th Conference on Neural Information Processing Systems* (*NeurIPS* 2022), New Orleans, 28 November-9 December 2022, 25779-25791.

[7] Yan, Y., Li, G., Chen, Y. and Fan, J. (2024) Model-Based Reinforcement Learning for Offline Zero-Sum Markov Games. *Operations Research*, **72**, 2430-2445. https://doi.org/10.1287/opre.2022.0342

[8] Sayin, M., *et al*. (2021) Decentralized Q-Learning in Zero-Sum Markov Games. 35*th Conference on Neural Information Processing Systems* (*NeurIPS* 2021), 6-14 December 2021, 18320-18334.

[9] Yang, Y.D. and Wang, J. (2020) An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective.

[10] Puterman, M.L. (2014) Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons.

[11] Kakade, S.M. (2003) On the Sample Complexity of Reinforcement Learning. University of London, University College London.

附　录

下面证明 $\nabla_\beta V^2_{\pi_{\text{new}}}(\rho)\big|_{\beta=0} = \mathbb{A}^2_{(\mu_{\text{new}},\nu)}\left(\rho,(\mu_{\text{new}},\nu')\right)$

首先，对于任意的 $s\in\mathcal{S}, t<T$

$$
\begin{aligned}
&\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\nabla_\beta V^2_{\pi_{\text{new}},t}(s)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\nabla_\beta \mathbb{E}_{(a,b)\sim\pi_{\text{new},t}(\cdot|s,\alpha,\beta)}\left[Q^2_{\pi_{\text{new}},t}(s,a,b)\right]\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\left(\nabla_\beta\pi_{\text{new},t}(a,b|s,\alpha,\beta)\right)Q^2_{\pi_{\text{new}},t}(s,a,b)+\sum_{a,b}\pi_{\text{new},t}(a,b|s,\alpha,\beta)\nabla_\beta Q^2_{\pi_{\text{new}},t}(s,a,b)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\left(\nabla_\beta\nu_{\text{new},t}(b|s,\beta)\right)Q^2_{\pi_{\text{new}},t}(s,a,b)\right]\\
&\quad+\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\nu_{\text{new},t}(a|s,\alpha)\nabla_\beta Q^2_{\pi_{\text{new}},t}(s,a,b)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\left(\nabla_\beta\nu_{\text{new},t}(b|s,\beta)\right)Q^2_{\pi_{\text{new}},t}(s,a,b)\right]\\
&\quad+\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\mathbb{E}_{(a,b)\sim\pi_{\text{new},t}(\cdot|s,\alpha,\beta)}\left[\nabla_\beta\left((1-\gamma)r^2_t(s,a,b)+\gamma\mathbb{E}_{s'\sim P_t(\cdot|s,a,b)}V^2_{\pi_{\text{new}},t+1}(s')\right)\right]\\
&=\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\left(\nabla_\beta\nu_{\text{new},t}(b|s,\beta)\right)Q^2_{\pi_{\text{new}},t}(s,a,b)\right]\\
&\quad+\gamma\mathbb{E}_{s\sim\Pr(s_{t+1}=s|\pi_{\text{new}},s_0)}\left[\nabla_\beta V^2_{\pi_{\text{new}},t+1}(s)\right]
\end{aligned}
$$

其次，由于 $\nabla_\beta V^2_{\pi_{\text{new}}}(s_0)=\mathbb{E}_{s\sim\Pr(s_0=s|\pi_{\text{new}},s_0)}\left[\nabla_\beta V^2_{\pi_{\text{new}}}(s)\right]$，对前面的等式进行递归，并利用 $V_{\pi,T}=0$ 对于所有 π，可以得到

$$
\nabla_\beta V^2_{\pi_{\text{new}}}(s_0)=\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\left[\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\left(\nabla_\beta\nu_{\text{new},t}(b|s,\beta)\right)Q^2_{\pi_{\text{new}},t}(s,a,b)\right]
$$

根据(2.4)，有

$$
\begin{aligned}
\nabla_\beta V^2_{\pi_{\text{new}}}(\rho)&=\mathbb{E}_{s_0\sim\rho_0}\left[\nabla_\beta V^2_{\text{new}}(s_0)\right]\\
&=\mathbb{E}_{s_0\sim\rho_0}\left[\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_{s\sim\Pr(s_t=s|\pi_{\text{new}},s_0)}\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\nabla_\beta\nu_{\text{new},t}(b|s,\beta)Q^2_{\pi_{\text{new},t}}(s,a,b)\right]
\end{aligned}
$$

根据更新规则(3.1)，有

$$
\begin{aligned}
&\sum_{a,b}\left[\mu_{\text{new},t}(a|s,\alpha)\nabla_\beta\nu_{\text{new},t}(b|s,\beta)Q^2_{\pi_{\text{new},t}}(s,a,b)\right]\Big|_{\beta=0}\\
&=\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\left(\nu'_t(b|s)-\nu_t(b|s)\right)Q^2_{(\mu_{\text{new}},\nu),t}(s,a,b)\\
&=\sum_{a,b}\mu_{\text{new},t}(a|s,\alpha)\left(\nu'_t(b|s)-\nu_t(b|s)\right)A^2_{(\mu_{\text{new}},\nu),t}(s,a,b)\\
&=\mathbb{E}_{(a,b)\sim(\mu_{\text{new},t},\nu'_t)}\left[A^2_{(\mu_{\text{new}},\nu),t}(s,a,b)\right]
\end{aligned}
$$

最后，结合上面两个式子可以得到：

$$\begin{aligned}\nabla_{\beta} V_{\pi_{\text{new}}}^{2}(\rho)\Big|_{\beta=0} &= \mathbb{E}_{s_o\sim\rho_0}\left\{\sum_{t=0}^{T-1}\gamma^{t}\mathbb{E}_{s\sim\Pr\left(s_t=s\middle|\left(\mu_{\text{new}},\nu\right),s_0\right)}\mathbb{E}_{(a,b)\sim\left(\mu_{\text{new}},t,\nu_t'\right)}\left[A_{\left(\mu_{\text{new}},\nu\right),t}^{2}(s,a,b)\right]\right\}\\ &= \mathbb{E}_{s_o\sim\rho_0}\left\{\sum_{t=0}^{T-1}\gamma^{t}\mathbb{E}_{s\sim\Pr\left(s_t=s\middle|\left(\mu_{\text{new}},\nu\right),s_0\right)}\mathbb{E}_{(a,b)\sim\left(\mu_{\text{new}},t,\nu_t'\right)}\left[A_{\left(\mu_{\text{new}},\nu\right),t}^{2}(s,a,b)\right]\right\}\\ &= \mathbb{A}_{\left(\mu_{\text{new}},\nu\right)}^{2}\left(\rho,\left(\mu_{\text{new}},\nu'\right)\right)\end{aligned}$$