arXiv:2305.02058v1 [cs.AI] 3 May 2023

基于合作马尔可夫决策过程系统的人机协同适应界面

郭凯瑞 [2]、Adrian Cheng[2]，李雅琪[1]，李军 [2]、Rob Duffield [2]、Steven W. Su [1]*

*†‡

摘要

本文旨在从用户（患者）和机器（机器人）双方的角度出发，通过引入基于模型的强化学习共同适应技术，开发一种新型人机界面，以提高康复效果。以往研究更多关注机器人辅助，即改进控制策略，以实现按需辅助的目标。本研究将机器人辅助康复的完整过程视为协同适应过程或互学习过程，并强调用户对机器的适应。为此，我们在系统的高层抽象层中，基于合作多智能体强化学习（MARL）提出了协同自适应马尔可夫决策过程（CaMDPs）模型，用于量化学习速率。我们在策略迭代框架下提出了多种方法，以协同调整两个智能体之间的策略改进。基于所提出的协同自适应马尔可夫决策过程，仿真研究表明，采用所提出的多种策略改进方法可以缓解非平稳问题。

1 引言

对于机器人康复系统而言，人机交互（HMI）可定义为供人类使用或与人类共同使用的机器人系统 [11],，是实现智能机器与人类用户融合的关键因素之一 [26]。为满足健康、教育、国防和工业领域日益增长的需求，人们开发了配备先进人机交互平台的人机协作系统 [12]。为进一步提高机器与用户之间的学习效率和接受度，人机交互需要采用一种互学习策略，让机器和人类都参与学习 [25]。由于人类与机器之间存在动态交互，因此需要一种协同适应系统，持续监测人类的适应情况，并利用这些信息设计基于强化学习的互学习算法，以提高用户的适应速度。

近期项目已在使用踝关节外骨骼辅助行走时，结合生物力学数据采用协同适应辅助策略 [6]。此外，还开发了用于上肢运动的自适应切换模型[14]。尽管取得了这些成果，但在中风康复过程中，仍需要针对复杂动作的协同适应控制策略，尤其是面向特定且具有明确目的的运动技能。此外，为了进一步将用户适应过程中的高维特征融入人机交互系统，多智能体强化学习是开发人机协作策略的最佳选择 [16]。

** 通讯作者。

†1 山东第一医科大学医学人工智能与大数据学院、山东省医学科学院，中国。

‡2 澳大利亚悉尼科技大学。

第35届神经信息处理系统大会（NeurIPS 2021年）。

近年来，基于强化学习的多智能体协作系统得到了广泛研究 [32][24]。例如，适用于两个智能体之间协作决策学习演化的协作马尔可夫决策过程（CMDP）框架，已得到深入研究 [21]。这促使研究人员建立一种专门监督中风康复过程的模型，以指导患者的动作/行为模式，并改进机器人的控制策略，从而加速用户适应。

在本研究中，我们将提出一种特定的协同自适应马尔可夫决策过程（CaMDP），以构建一个处理用户与机器人之间协同适应问题的框架。通过采用基于模型的强化学习技术 [27]，结合状态抽象和时间抽象 [22] [28] [31]，所提出的框架/策略可以轻松扩展到具有更复杂环境模型的系统。

对于多智能体强化学习，由于智能体会根据各自的利益同时改进策略，在相互学习过程中，从单个智能体的视角来看，环境会变得非平稳且难以解释。

多智能体强化学习面临的其他挑战还包括计算复杂度、部分可观测性和信用分配 [33] [24] [32]。

将多智能体强化学习应用于中风康复时，需要特别考虑策略的调整频率。与其他动作对应治疗方案的医疗应用类似，完全自适应的强化学习算法往往不切实际；相反，通常只能运行经领域专家批准的固定策略来收集数据，并且每次都需要单独的审批流程 [4] [1] [2] [17]。此外，在个性化/个体化推荐 [30], 中，基于瞬时数据在线调整策略在计算上并不实际。更常见的做法是在部署新策略之前，先长时间积累数据。在许多此类应用中，适应能力实际上成为真正的瓶颈[4]。近年来，有限的适应能力被视为设计强化学习算法时的一种约束，这在概念上与受约束马尔可夫决策过程中的约束类似 [7] [35], ，但主要关注的是对策略改进频率设置上限的情形。

In t本文中，我们考虑通过协作马尔可夫决策过程系统实现人类与机器之间的协同适应。 a foc我们聚焦于机器人辅助康复。本文的贡献总结如下：
1。我们为机器人辅助康复系统提出了一种协同自适应马尔可夫决策过程（CAMDP）模型。 m.
2我们分析了所提出的 CaMDP 价值函数的渐近特性。
3. 我们提出了一种修订的策略改进过程，以降低 Agent_0 (患者) 的策略适应率，同时将价值损失限制在一定范围内。

4. 我们提出了相应方法，以平衡所提出的 CaMDPs 模型中两个智能体的策略适应率。

2 面向协同适应的所提出的协作MDP模型

对于机器人辅助康复场景，我们在此定义两个智能体。智能体$_0$ 代表患者，智能体$_1$ 代表机器人。我们将降低智能体$_0$, 的切换成本，也就是说，策略切换越少越好。

由于人类与机器人之间的通信有限，因此在实时康复训练中，整体结构应采用分散式部分可观测系统。然而，在协同适应开发的早期阶段，分散式输出反馈配置并非我们首要关注的问题。在此，我们假设已采用观测器设计，且两个智能体中的任意一个都可以直接观测所有状态。
针对不同类型的分散式控制协作问题，已有研究提出了多种马尔可夫决策过程模型，以满足各种特殊需求。Goldman [10] 通过引入成本，研究了智能体之间不同的通信方式。作为针对特殊协同适应策略 [21], 的初步研究步骤，我们从简单的马尔可夫决策过程开始，具体如下。

假设双智能体马尔可夫决策过程的模型为 $M =< S, A_0, A_1, P, R >$，其中

1. S 是一个状态有限集合。
2. A_0 和 A_1 是需要考虑的两个智能体的控制动作有限集合。

3. P是转移概率函数。 $P(s'|s, a_0, a_1)$表示智能体 0 和 1 分别执行动作 a_0 和 a_1 时，从状态 $s \in S$ 转移到状态 $s' \in S$ 的概率。我们注意到，转移模型是平稳的，即与时间无关。4. R 是全局奖励函数。 $R(s, a_0, a_1, s')$表示在状态 s 下，智能体 0 执行动作 a_0 、智能体 1 执行动作 a_1 ，并因此转移到状态 s'时，系统整体获得的奖励。

这里，与 [10], 的安排类似，我们考虑三个状态集合：第一组状态是 $S_{0_i}\ i \in \{1, 2, \cdots, Ns_0\}$ 仅由患者控制的状态；第二组状态是 $S_{s_i}\ i \in \{1, 2, \cdots, Ns_s\}$ 同时受机器人和患者影响的状态；第三组状态是 $S_{1_i}\ i \in \{1, 2, \cdots, Ns_1\}$ 仅由机器人控制的状态。需要强调的是，当状态 S_{0_i} 和 S_{1_i} 的维度都较小时，两个智能体的共同适应会更加透明 [34] [5] 、更加容易，因为此时问题更接近单智能体学习问题。为降低这两组状态的维度，我们可以使用更多传感器来生成通信渠道。例如，患者情绪特征可以从生理信号（如脑电图、心电图）和/或基于计算机视觉的信号中提取，并由两个智能体共享。

我们对该系统给出更为形式化的定义如下：

定义 1 如果状态集 S 可以分解为三个部分 S_0、 S_1 和 S_s ，则双智能体马尔可夫决策过程是一个协同自适应马尔可夫决策过程（CAMDP）系统：
$\forall\, s_0, s'_0 \in S_0, \forall\, s_s, s'_s \in S_s, \forall\, s_1, s'_1 \in S_1$, 我们有$P(s'_0|(s_0, s_s, s_1)\ a_0, a_1) = P(s'_0|s_0, a_0)$;$P(s'_s|(s_0, s_s, s_1)\ a_0, a_1) = P(s'_s|s_s, a_0, a_1)$;$P(s'_1|(s_0, s_s, s_1)\ a_0, a_1) = P(s'_1|s_1, a_1)$;
$R(s_0, a_0, a_1, (s'_0, s'_s, s'_1)) = R(s_0, a_0, s'_0)$;$R(s_s, a_0, a_1, (s'_0, s'_s, s'_1)) = R(s_s, a_0, a_1, s'_s)$;$R(s_1, a_0, a_1, (s'_0, s'_s, s'_1)) = R(s_1, a_1, s'_1)$。
换句话说，CAMDP 的转移概率 P和重述函数 R 均可表示为

$P = P_0 \bigotimes P_s \bigotimes P_1$，其中 $P_0 = P(s'_0|s_0, a_0)$、 $P_s = P(s'_s|s_s, a_s)$ 和 $P_1 = P(s'_1|s_1, a_1)$；以及 $R = R_0 \bigotimes R_s \bigotimes R_1$，其中 $R_0 = R(s_0, a_0, s'_0)$、 $R_s = R(s_s, a_0, a_1, s'_s)$ 和 $R_1 = R(s_1, a_1, s'_1)$，其中“$\bigotimes$”表示克罗内克积。

在此，我们未考虑任何智能体$_0$ 或智能体$_1$, 都无法控制的状态，但这些状态可能会影响状态 S_0 、 S_s 和 / 或 S_1。例如，患者或机器人都无法控制天气，但天气可能会影响患者的情绪。未来如有需要，也可以将不可控状态纳入 CAMDP。

如前所述，基于各个子系统的 P 和 R ，我们可以通过克罗内克积为整个系统构造增广的 P 和 R，即 $P = P_0 \bigotimes P_s \bigotimes P_1$。然而，在某些情况下，三个状态集合中的某些集合可能为空，但仍可通过其余状态集合的克罗内克积构造增广的 P 和 R 。

要为不同的康复情境构建 CAMDP 模型，我们首先需要选择合适的配置（去中心化或集中式），然后确定系统的主要组成部分，包括状态、动作和奖励函数。

康复场景有两种：医院康复和居家康复 [18]。对于医院康复，由于各种医疗设备的可用性以及医生监督，通常可以获取两个智能体的状态信息。因此，可以采用集中式分析/训练配置。另一方面，对于居家康复，由于缺乏充足的医疗设备监测、医生监督和快速通信渠道，应考虑采用去中心化配置。这里，我们考虑去中心化的耦合多智能体决策过程模型，其中，奖励函数、状态信息和动作只能分别由患者和机器人获取（参见 [10] 中关于去中心化记录函数设置的讨论）。

康复用 CaMDP 的构建是一项复杂的工作，需要医疗专业人员与系统工程师协作选择模型参数（例如状态、

控制动作和奖励函数）。康复训练的低层模型会非常复杂，因为机器人和患者都属于该系统的一部分。根据 [9] [29], 中讨论的参数敏感性分析技术，在医疗专业人员的咨询指导下，可以简化模型的低层部分。

所提出的 CaMDPs 并非为机器人辅助康复系统的低层设计。实际上，通过使用抽象技术，可以构建一种分层强化学习 [15] 配置，其中 CaMDPs 作为康复系统的高层。

为了进一步简化讨论，本研究仅考虑 $finite$ 马尔可夫决策过程 [3]。原因之一是，在康复工程领域，患者的评级通常仅采用近似整数值。例如，NIHSS（美国国立卫生研究院卒中量表）作为临床卒中评估与测量的金标准，通过整数值来评定卒中患者的严重程度（例如，对手臂运动功能的评估采用 0 至 4 的整数评分）。尽管一些生理信号（例如肌电图 EMG 和脑电图 EEG）已被用于个性化卒中评估，但患者的严重程度仍然可以通过有限的离散值进行评估。

此外，所提出的 CAMDP 的动作空间也是有限维的。一种可能的动作空间离散化方法是构建分层控制结构。在双层配置中，控制器的高层负责切换预先设计的低层控制器，详见 [14]。人类专家的知识可以通过子系统（例如子控制器）的设计集成到系统 [23] [19] 中。

总之，CaMDP 旨在通过有限状态和控制动作处理机器人辅助康复的高层。

3 初步知识

基于增强后的 P 和 R，我们提出以下引理，以探究价值函数的渐近特征，从而便于后续分析。

引理 1 假设转移矩阵 P 是一个不可约非周期随机矩阵（即拟正或遍历的 *[8]*）。那么，下列矩阵的所有列向量都】【。

$$\lim_{n\to\infty}\lim_{\gamma\to 1^-}(\frac{1}{n}\sum_{i=0}^{n}(\gamma^i P^i))$$

将趋向于相同的值。

引理 2 对于三个可能性转移矩阵 A、B 和 C，其增广可能性转移矩阵：

$$A \bigotimes B \bigotimes C$$

当且仅当 A、B 和 C 都是不可约非周期随机矩阵时，它才是不可约非周期随机矩阵。

对于康复训练，由于很难专门选择特定的初始状态，我们提出以下引理，以分析价值函数与状态初始条件之间的关系。

引理 3 对于 CaMDP 模型，假设三个子系统的所有概率转移矩阵均为拟正矩阵。那么，对于任意两个控制策略 π_0 和 π_1 ($\pi = \{\pi_0, \pi_1\}$)，价值函数 $V(s_i)$ 将收敛至：

$$[\mathbf{I} - \gamma P]^{-1}(diag(PR')) = \lim_{n\to\infty}(\sum_{i=0}^{n}(\gamma^i P^i))(diag(PR')),$$

其中，$0 < \gamma < 1$ 是折扣因子。此外，当 γ 趋近于 1 时，$V(s_i)$（对于 $\forall s_i$）将收敛到相同的值：

$$g^\pi = \lim_{n\to\infty}\lim_{\gamma\to 1^-}(\frac{1}{n}\sum_{i=0}^{n}(\gamma^i P^i))(diag(PR')).$$

4 CAMDP策略改进的调整

如前所述，非平稳性是多智能体强化学习（MARL）面临的最大挑战之一。对于零和多智能体强化学习，Mazumdar 等人 [20] 证明，在简单的线性二次博弈中，单智能体策略梯度方法无法收敛。原因在于，各智能体会根据自身利益同时改进各自的策略。单个智能体无法判断状态转移或奖励变化究竟是自身动作导致的实际结果，还是其他智能体探索所致。

对于本研究所探讨的 CAMDP 合作强化学习，如果存在 $cheap$ 通信渠道 [10]，则可以协调两个智能体调整各自的策略。此外，为了妥善处理非平稳问题，我们可以设计一个双层分层学习框架，并在高层实现各种切换策略。特别是对于所提出的 CAMDP 模型，我们认为可以在高层开发一种“智能”切换算法，以显著减轻非平稳性带来的不利影响。例如，我们可以设计一种特殊的切换律，使两个智能体不再同时执行策略改进过程，而是交替改进各自的策略。在这种情况下，由策略调整引起的相互影响将会减小，从而有望缓解非平稳现象。

然而，在某些情况下，在高层构建最优切换律可能耗时较长或没有必要。本研究不再预先设计切换律，而是引入新的方法，使两个智能体能够“自动”调整其策略改进过程的更新频率，并确保互学习过程的收敛。

在接下来的两个小节中，我们将介绍所提出的适用于 CAMDP 的自调整切换方法，这些方法也可扩展到其他多智能体强化学习问题。

4.1 Agent_0的修订策略改进过程

正如引言部分所讨论的，在医疗应用中，策略对应于治疗方案 [7] [35], ，因此无法切换 Agent_0 （即患者）的策略。为降低切换频率，我们首先为 Agent_0介绍一种新的策略改进过程。我们将证明，该过程能够在预先设定的有界价值损失下，显著降低切换频率。

为了更直观地介绍所提出的切换频率调节方法，我们借鉴了二阶系统的调节方法，用“阻尼比”来表示切换频率的可调节程度。为了调节系统的“阻尼比”，即调整切换频率，我们将介绍一些新方法。这些方法的关键是修订的策略改进过程（如下文的程序1所示）。与经典强化学习相比，修订的策略改进过程通过预先设定的阈值 (η)，能够降低切换频率，同时将价值损失限制在与 η 成比例的有界范围内。

程序1（修订策略改进程序）策略改进
$policy - stable \longleftarrow true$
对于每个 $s \in S$、 k 以及给定的 η:

$temp \longleftarrow \pi_k(s)$

Under policy π_k *calculate*

$J_k = \sum_{s_0',r} p(s', r|s, a)[r^a_{s,s'} + \gamma V(s')]$.

If $\max_a(\sum_{s_0',r} p(s', r|s, a)[r^a_{s,s'} + \gamma V(s')]) - J_k \geq \eta$,

$\pi(s) \longleftarrow \arg\max_a(\sum_{s_0',r} p(s', r|s, a)[r^a_{s,s'} + \gamma V(s')])$.

如果 $temp \neq \pi(s)$，则 $policy - stable \longleftarrow false$
如果 $policy - stable$，则停止并返回 V和 π；否则转到策略评估步骤。

定理1 对于一个 CAMDP，我们假设它在所有策略下都是遍历的 $\pi \in \Pi$。如果 Agent_0 根据修订策略改进（程序1）进行调整，并且 Agent_1 正在执行

策略 π_1^j，那么价值损失 δV 将小于 $\eta[\mathbf{I}-\gamma\mathbf{P}^{\pi^*}]^{-1}\mathbf{1}$，其中 $\mathbf{P}^{\pi^*}$ 是最优策略 π^*下的状态概率转移矩阵，而 1 是全一向量。

证明 1 如果将整体 CAMDP 视为单智能体马尔可夫决策过程，则其在策略 π_0 和 π_1 下的概率转移矩阵和奖励矩阵可根据状态集合 S_0、S_s 和 S_1构造。接下来，考虑 π_1 固定采用其第 j个策略的情况，即 π_1^j。此时需要对智能体$_0$进行策略改进。假设智能体$_0$ 的当前策略为π_0^m，并且通过经典策略改进过程选出的策略为 π_0^n。我们将整体系统的两个增广策略表示如下：

$$\pi^m=\{\pi_0^m,\pi_1^j\}\quad \pi^n=\{\pi_0^n,\pi_1^j\}$$

根据经典策略改进例程 *[13],*，由于选择了 π_0^n而非 π_0^m，因此有

$$r_i^{\pi^n}+\gamma\sum_{j=1}^{N}p_{ij}^{\pi^n}v_j^{\pi^m}\geq r_i^{\pi^m}+\gamma\sum_{j=1}^{N}p_{ij}^{\pi^m}v_j^{\pi^m}.$$

其中，$i\in\{1,2,\cdots,N\}$和 N是增广系统的状态总数。
对于两个组合策略 π^m 和 π^n 以及每个状态 s_i，有如下等式 *s:*

$$\begin{aligned}v_i^{\pi^m}&=r_i^{\pi^m}+\gamma\textstyle\sum_{j=1}^{N}p_{ij}^{\pi^m}v_j^{\pi^m},\\ v_i^{\pi^n}&=r_i^{\pi^n}+\gamma\textstyle\sum_{j=1}^{N}p_{ij}^{\pi^n}v_j^{\pi^n}.\end{aligned}\tag{1}$$

考虑到策略改进过程，我们针对每个特定状态 s_i 定义：

$$g_i=r_i^{\pi^n}+\gamma\sum_{j=1}^{N}p_{ij}^{\pi^n}v_j^{\pi^m}-r_i^{\pi^m}-\gamma\sum_{j=1}^{N}p_{ij}^{\pi^m}v_j^{\pi^m}\tag{2}$$

显然，$\forall i,g_i\geq 0$。此外，在策略 π^m 和 π^n下，根据方程 (1) 和 (2)，我们有

$$v_i^{\pi^n}-v_i^{\pi^m}=g_i+\gamma\sum_{j=1}^{N}p_{ij}^{\pi^n}(v_j^{\pi^n}-v_j^{\pi^m}).\tag{3}$$

定义 $\delta v_i=v_i^{\pi^n}-v_i^{\pi^m}$ 后，我们有

$$\delta v_i=g_i+\gamma\sum_{j=1}^{N}p_{ij}^{\pi^n}\delta v_j\tag{4}$$

由于假设整体 CAMDP 在所有策略下都是遍历的，因此我们知道上述方程的向量解如下：

$$\delta\mathbf{V}=[\mathbf{I}-\gamma\mathbf{P}^{\pi^n}]^{-1}\mathbf{g}.\tag{5}$$

可以看出，δV ≥ 0 作为 $[\mathbf{I}-\gamma\mathbf{P}^{\pi^n}]^{-1}=\sum_{k=0}^{+\infty}\gamma^k(\mathbf{P}^{\pi^n})^k$ 的元素，均为非负数。此外，若 $g_i\leq\eta$，则有

$$\delta\mathbf{V}\leq\eta[\mathbf{I}-\gamma\mathbf{P}^{\pi^n}]^{-1}\mathbf{1}.$$

如果 策略 π^n 是最优的，我们可以看出，价值损失不会大于 $\eta[\mathbf{I}-\gamma\mathbf{P}^{\pi^*}]^{-1}\mathbf{1}=\frac{\eta}{1-\gamma}$.
th

注1 此处，我们仅考虑智能体$_1$ 执行控制策略 π_1^j ($j\in\{1,\cdots,N_{action_1}\}$ (N_{action_1}是智能体$_1$的动作选项数）的情况。为确保总体损失小于 ϵ，我们需要考虑情况 $\forall\pi_1^j,j\in\{1,\cdots,N_{action_1}\}$。

4.2 面向 CaMDP 的自调整切换适应策略

为解决所提出的 CAMDP 的非平稳性问题，我们引入一种基于修订的策略改进过程的实用方法，以提高 CAMDP 的收敛性。我们通过经典二阶系统的阶跃响应来阐释这一思路。对于二阶系统，若其阻尼比“ζ”减小，系统响应会更快，但稳定裕度会降低。另一方面，若阻尼比增大，系统响应会更慢，但稳定裕度会更大。对于 CAMDP 的两个智能体，若其中一个智能体的响应速度远快于另一个（通常，我们预期 Agent_1 (机器人) 在策略优化方面的响应速度快于 Agent_0 (患者)），那么出现非平稳性的可能性就会降低。基于这一思路，我们提出以下两种方法来调整两个智能体的策略改进速率（即 $damping\,ratio$），使它们能够协调工作。

基于修订的策略改进过程（程序1），根据可容忍的价值损失预先设定阈值 η。这个固定阈值 η 类似于价值损失的 $proportional$ 度量。由于在大多数情况下，学习过程开始时无法获得完整模型和渐近价值，因此预先选定合适的 η 并不现实。另一种选择是选取相对较大的 η，并定义一个 $integral-alike$ 参数。结合预先选定的$proportional-alike$ 参数，我们便可构建比例积分型策略改进方案，调整策略改进频率，从而按照期望的节奏调整系统的 $damping\,ratio$。我们可以在学习过程中对这两个参数进行微调。我们将这种方法称为“类 PI”调整方案。

Spec具体而言，假设类积分参数为“κ_I”，我们可以修改切换条件
在程序中
过程1如下：
过程2（类PI策略改进过程）策略改进$policy - stable \longleftarrow true$
对于每个 $s \in S$、 k 以及预先选定的参数 M、 η 和 κ_I：

$temp \longleftarrow \pi_k(s)$

Under policy π_k *calculate*

$J_k = \sum_{s'_0,r} p(s', r|s, a)[r + \gamma V(s')]$.

Calculate $I_k = \max_a(\sum_{s'_0,r} p(s', r|s, a)[r + \gamma V(s')]) - J_k$.

if $\kappa_I \sum_{j=0}^{M} I_{k-j} \geq \eta$,

$\pi(s) \longleftarrow \arg\max_a(\sum_{s'_0,r} p(s', r|s, a)[r + \gamma V(s')])$.

如果 $temp \neq \pi(s)$，则 $policy - stable \longleftarrow false$
如果 $policy - stable$，则停止并返回 V 和 π；否则转到策略评估步骤。

类PI调节方法的实现需要修改经典策略改进过程。我们将在第5节通过数值仿真展示这种方法。

5 数值分析

本节考虑康复机器人与人类用户之间共同适应的一个简单示例。

考虑一个 CaMDPs系统 $M =< S$、 A_0、 A_1、 P_0、 P_1、 R_0、 $R_1 >$，具体如下：
状态是三个状态集合的克罗内克积： $S_0 = \{s_{00}, s_{01}\}$、 $S_s = \{s_{s0}, s_{s1}\}$ 和 $S_1 = \{s_{10}, s_{11}\}$。

动作集合是两个集合的克罗内克积： $A_0 = \{a_{0_0}, a_{0_1}\}$ 和 $A_1 = \{a_{1_0}, a_{1_1}\}$。
每个子系统在不同控制动作下的概率转移矩阵如下：

$$P_0(S_0, a_{0_0}, S_0) = \begin{bmatrix} 0.8229 & 0.1771 \\ 0.7826 & 0.2174 \end{bmatrix};$$

$$P_0(S_0, a_{0_1}, S_0) = \begin{bmatrix} 0.6406 & 0.3594 \\ 0.4919 & 0.5081 \end{bmatrix};$$

$$P_s(S_s, a_{0_0}, a_{1_0}, S_s) = \begin{bmatrix} 0.5821 & 0.4179 \\ 0.3839 & 0.6161 \end{bmatrix};$$

$$P_s(S_s, a_{0_0}, a_{1_1}, S_s) = \begin{bmatrix} 0.1838 & 0.8162 \\ 0.5686 & 0.4314 \end{bmatrix};$$

$$P_s(S_s, a_{0_1}, a_{1_0}, S_s) = \begin{bmatrix} 0.6990 & 0.3010 \\ 0.6169 & 0.3831 \end{bmatrix};$$

$$P_s(S_s, a_{0_1}, a_{1_1}, S_s) = \begin{bmatrix} 0.3448 & 0.6552 \\ 0.6432 & 0.3568 \end{bmatrix};$$

$$P_1(S_1, a_{1_0}, S_1) = \begin{bmatrix} 0.8022 & 0.1978 \\ 0.5396 & 0.4604 \end{bmatrix};$$

$$P_1(S_1, a_{1_1}, S_1) = \begin{bmatrix} 0.4083 & 0.5917 \\ 0.5815 & 0.4185 \end{bmatrix}.$$

奖励函数如下:

$$R_0(S_0, a_{0_0}, S_0) = \begin{bmatrix} 0.1565 & 0.1769 \\ 0.1909 & 0.1425 \end{bmatrix};$$

$$R_0(S_0, a_{0_1}, S_0) = \begin{bmatrix} 0.0520 & 0.2813 \\ 0.1530 & 0.1803 \end{bmatrix};$$

$$R_s(S_s, a_{0_0}, a_{1_0}, S_s) = \begin{bmatrix} 0.2136 & 0.1197 \\ 0.1533 & 0.1800 \end{bmatrix};$$

$$R_s(S_s, a_{0_0}, a_{1_1}, S_s) = \begin{bmatrix} 0.3047 & 0.0286 \\ 0.0895 & 0.2438 \end{bmatrix};$$

$$R_s(S_s, a_{0_1}, a_{1_0}, S_s) = \begin{bmatrix} 0.0077 & 0.3257 \\ 0.1378 & 0.1955 \end{bmatrix};$$

$$R_s(S_s, a_{0_1}, a_{1_1}, S_s) = \begin{bmatrix} 0.2806 & 0.0527 \\ 0.1625 & 0.1708 \end{bmatrix};$$

$$R_1(S_1, a_{1_0}, S_1) = \begin{bmatrix} 0.0190 & 0.3144 \\ 0.3120 & 0.0213 \end{bmatrix};$$

$$R_1(S_1, a_{1_1}, S_1) = \begin{bmatrix} 0.1878 & 0.1455 \\ 0.0450 & 0.2883 \end{bmatrix}.$$

5.1 不同折扣因子下价值的渐近特征

当折扣因子为 $\gamma = 0.5$ 时，见图1下半部分。各状态的渐近价值彼此差异较大。然而，当折扣因子接近 1 时，见图1上半部分（$\gamma = 0.98$）。各状态的渐近价值彼此十分接近。这与引理2一致。基于这一性质，我们将不考虑状态间的差异，提供一种更好的策略。

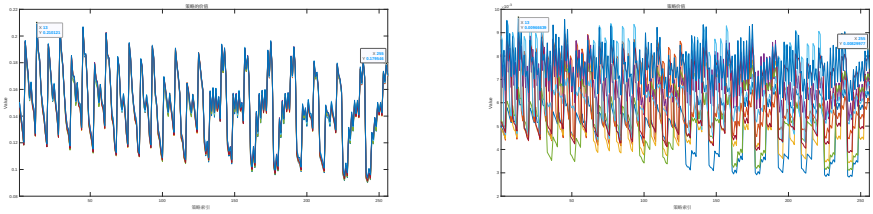
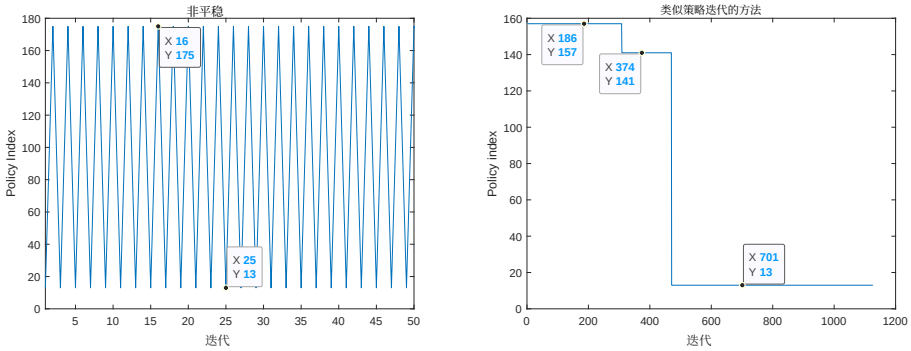


图1：价值与状态　　在不同折扣因子下：a. $\gamma = 0.98$ (l　　左）。b. $\gamma = 0.50$（右).



Figu图 2：策略改进的收敛性：a. 非平稳。b. 对 t 采用类似策略迭代的方法收敛性的改进。

5.2 修订策略改进过程中的价值损失

根据程序1（修订的策略改进过程）和定理1，我们进行数值分析。

根据引理的渐近分析，最优策略（$value = 0.2101$）可计算如下：

对于智能体$_0$，有

$$\begin{bmatrix} State: & \{s_{0_0}, s_{s_0}\} & \{s_{0_0}, s_{s_1}\} & \{s_{0_1}, s_{s_0}\} & \{s_{0_1}, s_{s_1}\} \\ Action_0: & 0 & 0 & 0 & 0 \end{bmatrix}.$$

对于智能体$_1$，有

$$\begin{bmatrix} State: & \{s_{1_0}, s_{s_0}\} & \{s_{1_0}, s_{s_1}\} & \{s_{1_1}, s_{s_0}\} & \{s_{1_1}, s_{s_1}\} \\ Action_1: & 1 & 1 & 0 & 0 \end{bmatrix}.$$

为简便起见，在不会造成混淆时，我们简称为 $\pi_0 = [0\,0\,0\,0]$ 和 $\pi_1 = [1\,1\,0\,0]$，或 $[0\,0\,0\,0][1\,1\,0\,0]$。

如果我们将初始控制策略设为 $\pi_0 = [1\,1\,1\,1]$ 和 $\pi_1 = [1\,1\,0\,0]$（$value = 0.1736$），且智能体 $_0$ 采用程序1，而智能体 $_1$ 仍采用常规的策略改进策略。

当我们选择不同的 η 价值时，Agent$_0$ 可以保持相同的策略（以避免切换控制策略），也可以针对某些状态部分切换策略，或者完全切换到最优策略。详细结果见表 1。

从表1可以看出，为避免出现非平稳现象，应合理选择 η 价值。然而，在一开始，由于缺乏对该过程的了解，预先选定合适的 η颇具挑战。在此，我们采用过程2所述的类似策略迭代的方法。我们简单地选择 $\kappa_I = 1$ 和 $\eta = 0.1$；策略改进的非平稳问题已得到解决，并且二者收敛至最优策略 $\pi_0^* = [0\,0\,0\,0]$ 和 $\pi_1^* = [1\,1\,0\,0]$（$value = 0.210$），如图2所示。

策略	策略编号	策略值	η 值
[1111] [1111]	编号 256	0.180	$> 1.0625 \times 10^{-4}$
[0111] [1100]	编号 125	0.187	1.0000×10^{-4}
[0001] [1100]	第29号	0.207	8.1250×10^{-5}
[0000] [1100]	第13期	0.210	6.2500×10^{-5}
$\begin{matrix}[0000]\,[1100]\\ [1010]\,[1110]\end{matrix}$	13号对阵175号	0.210 对比 0.196	$< 3.1250 \times 10^{-5}$

表 1：修订的策略改进过程的数值分析。

6 结论

本文旨在构建一个基于强化学习的中风康复框架，并提出了一种协作马尔可夫决策过程模型，以改进患者和智能康复设备双方的协同适应学习过程。我们在多智能体强化学习框架下研究了该模型，其中最关键的问题是两个智能体在相互学习过程中协同适应所具有的非平稳性。我们提出了若干自动切换策略，以确保策略改进过程收敛。此外，为了降低患者策略改进的频率，我们提出了一种修订的策略改进过程，以平衡策略改进与总体价值损失。

基于所提出的策略改进方法，对患者与智能康复设备之间的协同适应进行了数值研究。我们的最终目标是在真实实验和临床环境中，从患者和智能康复机器人双方的角度实现最优自适应学习。

参考文献

[1] D. Almirall、S. N. Compton、M. Gunlicks-Stoessel、N. Duan 和 S. A. Murphy。为开发自适应治疗策略而设计的试点序贯多重分配随机试验。《医学统计》，31(17)：1887–1902，2012。[2] D. Almirall、I. Nahum-Shani、N. E. Sherwood 和 S. A. Murphy。自适应干预开发的智能设计简介：应用于减重研究。《转化行为医学》，4(3)：260–274，2014。[3] F. D. M. P. M. Arias。医学中的马尔可夫决策过程：机遇与挑战。[4] Y. Bai、T. Xie、N. Jiang 和 Y.-X. Wang。具有低切换成本且可证明高效的Q学习。arXiv预印本 arXiv:1905.12849，2019。[5] D. Boyd。在自适应数字系统中实现透明度。《新探索：文化与传播研究》，2(3)，2022年7月。网址 https://jps.library.utoronto.ca/index.php/nexj/article/view/39030。[6] M. Collins、M. Sutherland、L. Bouwer、S.-M. Cheong、T. Frolicher、H. J. DesCombes、M. K. Roxy、I. Losada、K. McInnes、B. Ratter 等。极端事件、突变与风险管理。2019。[7] C. Dann、L. Li、W. Wei 和 E. Brunskill。策略证书：迈向可问责强化学习。载于《国际机器学习会议》，第1507–1516页。PMLR，2019。[8] F. Garcia 和 E. Rachelson。马尔可夫决策过程。《人工智能中的马尔可夫决策过程》，第1–38页，2013。[9] D. Gold、P. Reed、A. Hadjimichael、K. Malek、T. Karimi、V. Srikrishnan、K. Keller、R. Gupta、C. Vernon 和 J. Rice。应对多部门动力学研究中的不确定性：面向新手和资深建模者的电子书指南。载于《AGU秋季会议摘要》，第2021卷，第GC15E–0740页，2021年。[10] C. V. Goldman 和 S. Zilberstein。协作系统的分散式控制：分类与复杂性分析。《人工智能研究杂志》，22：143–174，2004。[11] M. A. Goodrich、A. C. Schultz 等。人机交互：综述。《人机交互基础与趋势®》，1(3)：203–275，2008。[12] M. A. Gull、S. Bai 和 T. Bak。上肢外骨骼设计综述。《机器人学》，9(1)：16，2020。

[13] R . A. 霍华德. 动态规划与马尔可夫过程 . 1960.

[14] Y. Huang、R. Song、A. Argha、B. G. Celler、A. V. Savkin 和 S. W. Su。基于无冲击切换机制的康复机器人人体运动意图描述。《IEEE 神经系统与康复工程汇刊》，29:673–682，2021年。[15] M. Hutsebaut-Buysse、K. Mets 和 S. Latré。分层强化学习：综述与开放的研究挑战。《机器学习与知识提取》，4(1):172–221，2022年。[16] J. Kim 和 H. Lee。面向动态竞争网络的随机梯度下降强化学习自适应人机评估框架。《应用科学》，10(7):2558，2020年。[17] H. Lei、I. Nahum-Shani、K. Lynch、D. Oslin 和 S. A. Murphy。用于构建个体化治疗序列的“智能”设计。《临床心理学年度综述》，8:21–48，2012年。[18] R. López-Liria、D. Padilla-Góngora、D. Catalan-Matamoros、P. Rocamora-Pérez、S. Pérez-de la Cruz 和 M. Fernández-Sánchez。全膝关节置换术后的居家康复计划与医院康复计划。《生物医学研究国际》，2015，2015年。[19] M. Maadi、H. Akbarzadeh Khorshidi 和 U. Aickelin。机器学习中的人机交互综述及其对医疗应用的启示。《国际环境研究与公共卫生杂志》，18(4):2121，2021年。[20] E. Mazumdar、L. J. Ratliff、S. Sastry 和 M. I. Jordan。线性二次动态博弈中的策略梯度不保证收敛。载于《平滑博弈优化与机器学习研讨会：连接博弈》，2019年。[21] X. Mo、D. Xu 和 Z. Fu。合作马尔可夫决策过程系统的收敛性。《熵》，22(9):955，2020年。[22] T. M. Moerland、J. Broekens 和 C. M. Jonker。基于模型的强化学习：综述。arXiv 预印本 arXiv:2006.16712，2020年。[23] N. D. Nguyen、S. Nahavandi 和 T. Nguyen。深度强化学习中的人类混合策略方法。载于《2018年 IEEE 系统、人类与控制论国际会议（SMC）》，第4023–4028页。IEEE，2018年。[24] T. T. Nguyen、N. D. Nguyen 和 S. Nahavandi。多智能体系统的深度强化学习：挑战、解决方案与应用综述。《IEEE 控制论汇刊》，50 3839，2020 9 3826( ): – 。[25] S. Perdikis 和 J. d. R. Millan。脑机接口：两个学习者的故事。《IEEE 系统、人类与控制论杂志》，6(3):12–19，2020年。[26] L. Peternel、N. Tsagarakis、D. Caldwell 和 A. Ajoudani。机器人在人机协同操作中适应人体身体疲劳。《自主机器人》，42(5):1011–1021，2018年。[27] A. Plaat、W. Kosters 和 M. Preuss。高精度基于模型的强化学习：综述。arXiv 预印本 arXiv:2107.08241，2021年。[28] G. Qingji、W. Kai 和 L. Haijuan。基于 PAD 情绪空间的机器人情绪生成机制。载于《智能信息处理国际会议》，第138–147页。Springer，2008年。[29] V. Srikrishnan、D. C. Lafferty、T. E. Wong、J. R. Lamontagne、J. D. Quinn、S. Sharma、N. J. Molla、J. D. Her man、R. L. Sriver、J. F. Morris 等。多部门系统中的不确定性分析：复杂系统风险分析、预测与规划的考量。《地球的未来》，10 002644，2022 8 2021( ):e EF 。[30] G. Theocharous、P. S. Thomas 和 M. Ghavamzadeh。用于优化终身价值且带有保证的个性化广告推荐系统。载于第二十四届国际人工智能联合会议，2015年。[31] G. Wang、Z. Wang、S. Teng、Y. Xie 和 Y. Wang。基于马尔可夫决策过程的交互式虚拟人情绪模型。《中国电气与电子工程前沿》，2(2):156–160，2007年。[32] A. Wong、T. Bäck、A. V. Kononova 和 A. Plaat。多智能体深度强化学习：挑战及迈向类人方法的方向。arXiv 预印本 arXiv:2106.15691，2021年。[33] Y. Yang 和 J. Wang。从博弈论视角概述多智能体强化学习。arXiv 预印本 arXiv:2011.00583，2020年。

[34] E. Yigitbas、K. Karakaya、I. Jovanovikj 和 G. Engels。通过数字孪生和虚拟现实接口增强人在回路自适应系统。载于《2021年自适应与自管理系统软件工程国际研讨会（SEAMS）》，第 30–40 页，2021年。doi：10.1109/SEAMS51251.2021. 00015。

[35] M . Yu, Z . Yang, M . Kolar 和 Z . Wang. 面向安全强化学习的收敛策略优化 . 神经信息处理系统进展，32:3127–3139，2019。