[[1]](#footnote-0)

基于混合整数规划和深度强化学习解决机组组合优化问题

*Abstract*— 机组组合问题是电力市场中的经典问题，旨在一定的调度周期内，通过合理控制不同机组的运行情况，实现以最小成本安排发电计划并满足给定负荷的平衡。机组组合问题常使用混合整数线性规划进行建模与求解，在本次项目中，我们通过建立机组组合问题的目标函数与约束条件的数学模型，并使用工具进行求解，最终求得在给定发电计划下的机组运行情况与最低成本。同时，我们也使用了深度强化学习与遗传算法对机组组合问题进行求解并分析了其结果。

# INTRODUCTION

对于一个火力发电系统而言，其发电资源由众多各类型的发电机组组成，不同类型的机组有不同的运行成本与约束条件。随着人们生产、生活活动的规律性变化以及其他因素的影响，电力负荷呈现出周期性起伏波动的特性。如果仅靠改变机组出力而不改变启停机组的数目，在已开机机组的容量范围内往往难以通过调节满足负荷的这种变化。

在这种情况下，由于负载负荷变化的周期性，可以预先对不同机组在下一个调度周期内的运行情况进行优化，以实现成本的最小化，即机组组合问题。

在本次《优化方法》课程设计中，我们实现了多种约束方程定义最短启停时间约束，且得到在给定的负载需求下24小时最低运行成本474390。同时我们使用了多种强化学习算法与遗传算法实现了机组组合优化问题的求解，包括PPO,DDPG,TD3,A2C以及混合动作空间H-PPO,并据此分析了相关的算法和问题。

 ： Binary variable that is equal to 1 if unit j is online in period k and 0 otherwise

：Power output of unit j in period k.

：Maximum available power output of unit j in period k.

：Shutdown cost of unit j in period k.

：Production cost of unit j in period k.

：Startup cost of unit j in period k.

：Minimum up time of unit j.

：Minimum down time of unit j.

: Load demand in period k.

: Initial commitment state of unit j (1 if it is

online, 0 otherwise).

：the cost for a cold start

：the cost for a hot start

:the duration threshold in periods for considering a hot start or a cold start

: Set of indexes of the generating units.

: Set of indexes of the time periods.

# MILP-UC FORMULATION

## A. 组合优化的目标和约束方程建立

机组组合优化问题在电力系统运营中具有独特的特色，主要涉及到如何合理选择和配置发电机组，以满足系统负荷需求并优化运营成本。这类问题通常采用混合整数规划（MIP）方法进行建模和求解。

机组组合问题涉及到决策变量既包含连续型变量（如发电机的发电功率）又包含离散型变量（如机组的启停状态），这使得问题的复杂性得以体现。其次，由于每台机组的启停过程具有一定的时间和成本，问题中必须考虑时间约束、启停成本、机组的最小启停时间等实际运营中的约束条件，这增加了问题的难度。最后，机组组合优化问题需要在系统运行的短期内进行决策，以适应实时的负荷波动和市场变化，因此对解决方案的高效性和实时性要求较高。

MIP方法在解决机组组合优化问题中显示出优势。由于MIP方法能够同时处理连续型和离散型变量，因此能够更全面地刻画机组组合问题的特性。此外，MIP方法提供了灵活和准确的建模框架，能够考虑到诸如机组启停状态、时间约束、成本最小化等多方面的实际运营因素。虽然MIP方法可能在处理大规模问题时面临计算挑战，但通过优化算法和高性能计算技术的不断进步，MIP方法在解决机组组合优化问题中仍然是一个强大而可行的工具。

机组组合优化问题的目标是最小化成本，对于发电系统，除了机组的运行成本外，还包括机组的开启与停止成本。因此，机组组合优化的目标函数可以写成

 (1)

在机组运行的过程中，其输出功率需要满足负载所需功率的要求，因此约束方程之一为，

 (2)

为此我们进行了以下分析过程，逐步建立约束条件。

## B. MILP方程建立

**1）目标方程**：运行成本和开停成本可以被表示为

 (3)

 (4)

 (5)

**2）功率约束：**机组运行过程中存在最大最小的功率约束条件，可以据此建立数学方程。

 (6)

3）**最短开停时间约束**：最短开停时间约束最初被制定为混合整数线性表达式，依赖于与发电机组的启动、关闭和开/关状态相关联的二进制变量。接下来，提出了一种仅基于二进制变量的等效混合整数线性公式。对于最小上线时间约束的新表达式如下：

 (7)

 (8)

  (9)



这里的表示机组必须在线的初始周期数，数学表达式为。

约束（7）和机组的初始状态有关，（8）式的约束用于后续时段，以满足在所有可能的连续时段组合中的最小开启时间约束。（9）式的约束对最终时段进行建模，其中如果单位 j 启动，则其保持在线直至时间跨度结束。

类似地，最小停机时间约束的制定如下：









这里的表示机组必须在线的初始周期数，数学表达式为。

4）不考虑上一个周期末状态的最短开停时间约束：

上问介绍了如果考虑上一个周期末机组状态对优化目标周期影响的约束方程建立，但是在某些情况下不需要考虑这些因素，那最短开停时间限制就可以被简化为。



上式表明，如果机组在时刻开启，那么机组必须保持开启状态至少个周期，同理最短停机时间可表示为：



5）求解

CBC 是一个混合整数线性规划求解器，是由Coin-OR项目提供的开源线性规划求解器。主要基于分支定界法和割平面法。

分支定界法是一种递归的、分而治之的方法。它将整个问题划分为较小的子问题，然后逐步求解这些子问题。在每个阶段，通过界定目标函数值的上限和下限，来排除一些不必要的搜索空间。如果在某个子问题中找到了可行解，就更新全局最优解。这个过程一直进行，直到找到整个问题的最优解或者确定整个问题无解。

割平面法是一种在不断迭代中添加约束的方法。在每次迭代中，通过添加新的约束来割掉当前可行区域内的一部分，从而逐渐逼近最优解。这些新的约束通常基于线性规划松弛问题的解，用来修正整数规划问题中的不连续性。

在所给题目中，机组数目为8，时间周期是24小时，我们在python中定义以上优化目标函数和变量，使用CBC求解器求解，最终得到结果如下，运行24小时一个周期的总成本为475,560：

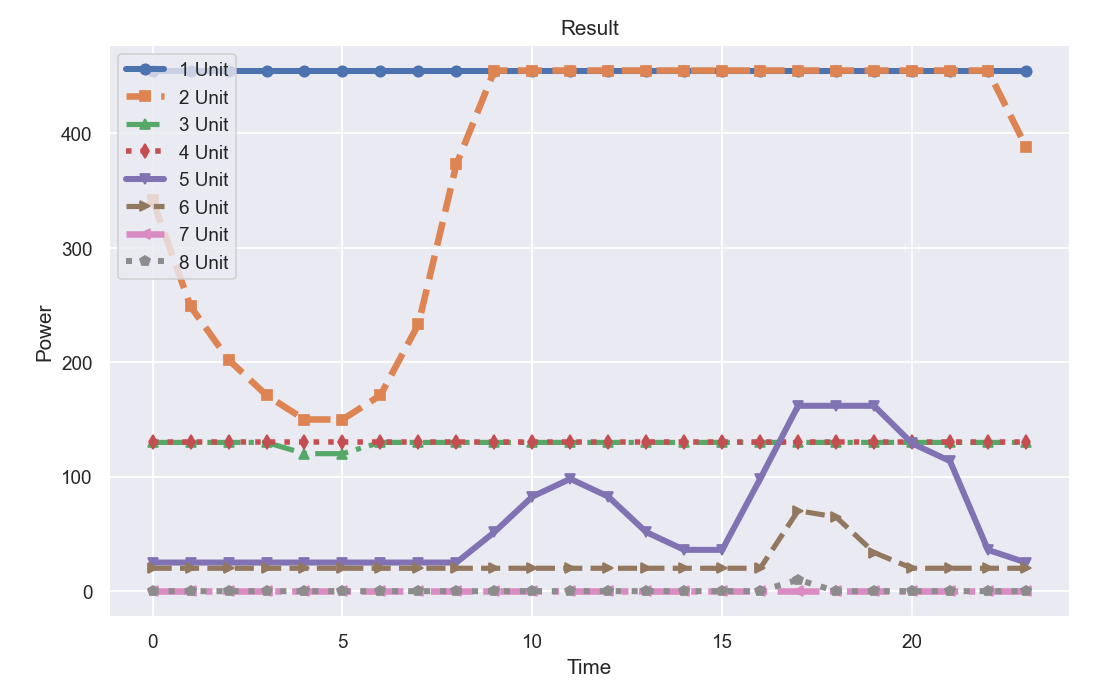


图 8个机组的优化分配

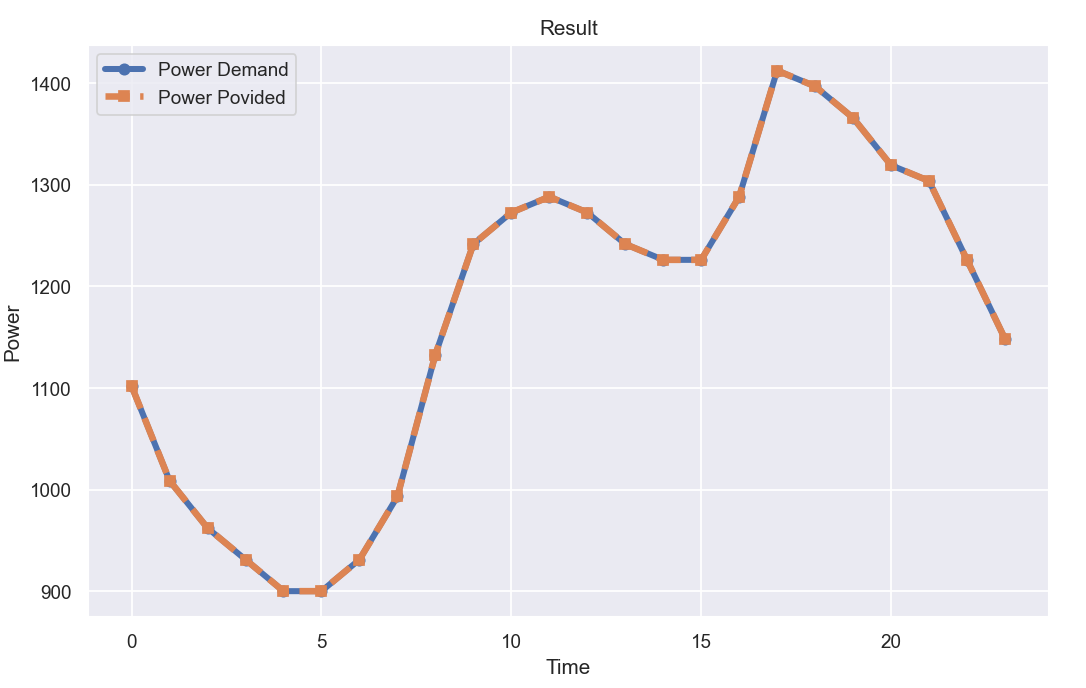
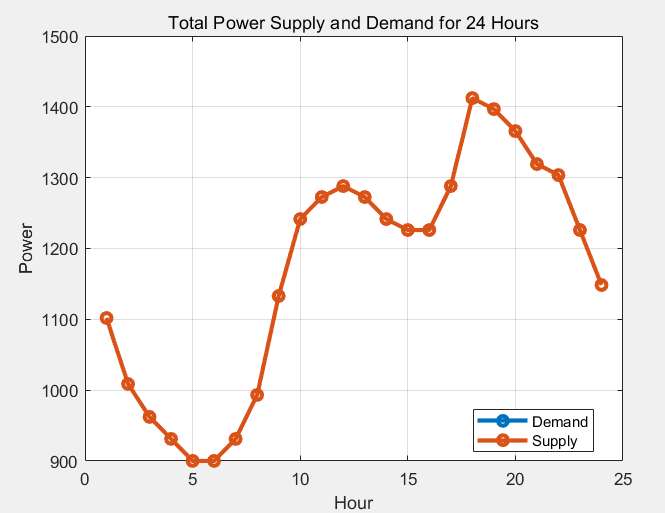
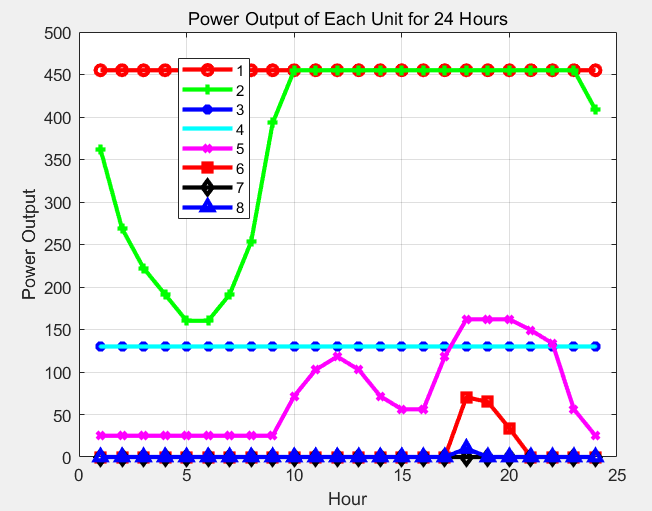


图 功率输出与功率要求

我们也在matlab里面重新建立混合整数规划模型，调用额外安装的gurobi求解器，该求解器目前全球综合排名第一，计算出的结果如下：





利用gurobi求解器求解的最终总成本为$474,390

# Genetic Algorithm for Units Unit Commitment Problem

遗传算法是一种模仿生物进化过程的优化方法，用于解决问题的算法。它模拟了自然选择、交叉和变异的过程，通过不断迭代地演化种群来寻找问题的最优解。遗传算法基于种群中个体的基因编码，以及这些个体之间的选择、交叉和变异操作，逐步优化搜索空间，以发现最优解或近似最优解。这些个体经过评估和选择后，通过遗传操作（如交叉和变异）产生新的后代个体，进而逐步优化问题的解。遗传算法在优化问题、搜索空间广泛且复杂的情况下具有较好的适用性，例如在机器学习、调度问题、工程优化和自动化设计等领域中被广泛应用。

机组启停优化问题涉及混合0-1整数规划，其特点包括离散且复杂的搜索空间以及伴随复杂约束条件。对于这类问题，遗传算法是一种合适的求解方法。GA能有效应对离散的决策空间，灵活地探索不同的机组启停状态组合，并且能处理问题中的多种复杂约束条件。其迭代优化策略和全局搜索能力有助于避免陷入局部最优解，更有可能找到全局最优解。由于其并行性和鲁棒性，GA能够在复杂环境中进行搜索，逐步优化启停状态组合，因此在解决机组启停优化问题时具备良好的适用性。

遗传算法基于受到自然系统中遗传和演化机制启发的原理，模仿生物群体的演化过程。其基本原则是以编码信息个体的形式维护解决问题（基因型）的一个群体。这些个体在时间中不断进化。进化基于自然选择的法则（适者生存）以及种群内的基因信息重组。进化的种群对搜索空间进行采样，并积累有关良好和糟糕质量区域的知识，通过重新组合这些知识形成对特定问题性能最优的解决方案。

首先，随机生成一个包含M个解决方案的群体，以类似自然染色体的符号串（最好是二进制串）进行编码。然后，将种群中的每个成员解码为实际的问题解决方案，并通过一个质量函数为其分配一个“适应度”值，该值表示解决方案的质量。

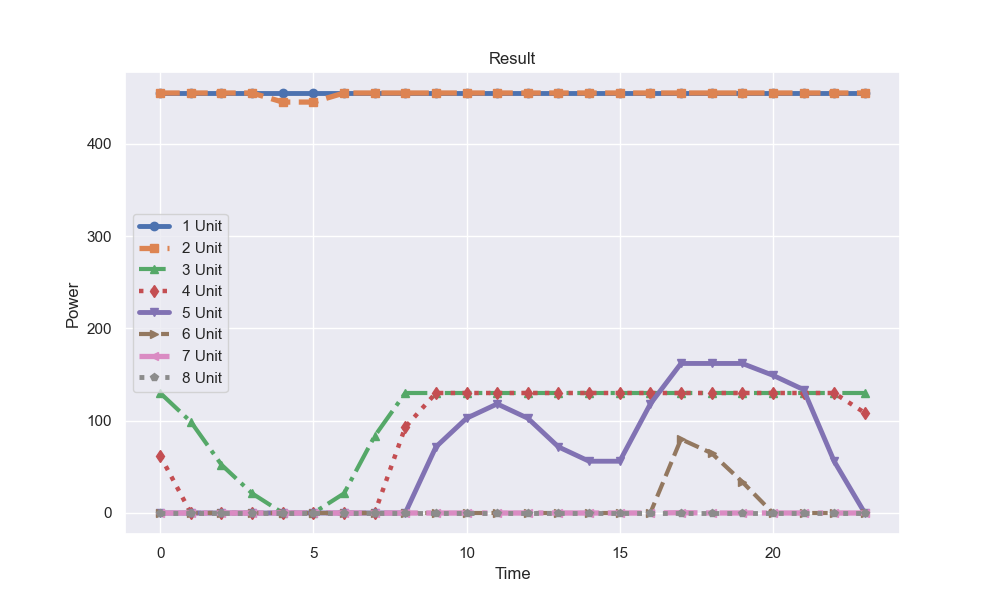
评估完成后，从种群中以成对方式选择个体进行复制，形成“后代”个体（即新的问题解决方案）。选择是以概率方式进行的，使得个体被选中的概率与其适应度成比例。这确保了高质量的解决方案被多次选择，成为许多新解决方案的“父母”，而低质量的解决方案则在新种群中的贡献较少，甚至有可能根本不被选中进行繁殖。

当选择了两个父母后，它们的符号串将使用类似遗传的操作符合并，生成一个“后代”解决方案。主要使用的操作符是“交叉”和“变异”。交叉简单地将父母的符号串组合成一个新的染色体串，该串继承了两个父母的解决方案特性。尽管交叉是主要的搜索操作，但它无法产生种群中尚不存在的信息。变异通过随机改变新染色体的符号来补充这一缺陷。一般来说，变异被视为次要但不是无用的操作符，它给予每个解决方案被考虑和评估的非零概率。

当生成了M个新的解决方案串时，它们被视为一个新的代，并完全取代了“父代”，以便进化继续进行。许多代的迭代是必要的，使得种群收敛到最优解或接近最优解，而迭代次数则根据问题的难度而增加。

在本次使用遗传算法求解机组组合优化问题过程中，我们需要着重介绍一下的是我们设计过程中定义的功率分配原则，因为遗传算法本质上是二进制码在不停的迭代更新，所以十分适合求解机组组合开停策略的学习，但是本题目要求不仅需要求解开停策略，还需要求解最小化成本的功率约束同时满足负载功率，于是在搜索过程中，我们将这一条件添加进个体适应度的计算中，分3种情况

1. 若某一时刻所有开启的机组的最大输出功率之和仍不满足负载需求，则该机组组合各取其对应的最大输出功率，计算功率差值，并入个体适应度
2. 若某一时刻所有开启的机组的最小输出功率之和都大于负载需求，毫无疑问，这会增加额外成本，我们令所有机组取最小输出功率，计算多余裕量，并入适应度
3. 若上面两个情况不成立，则说明存在一组分配情况使得所有开启机组的输出功率之和等于负载需求，则我们定义了依次分配，优先选择单位能量价格较低的机组，使其提供最大功率，依次分配下去，我们一开始想把一个很简单的线性规划放进去求解最佳功率分配，但在遗传算法里，这并不是求解问题，而是搜索问题，这个过程只是在评价某一庞大种群中不同个体的适应度，在这个庞大种群中，总会有一个符合各种约束且成本最小的。事实证明我们是对的，遗传算法的搜索结果和混合整数规划的求解结果十分接近，运行总成本大约$479,439。



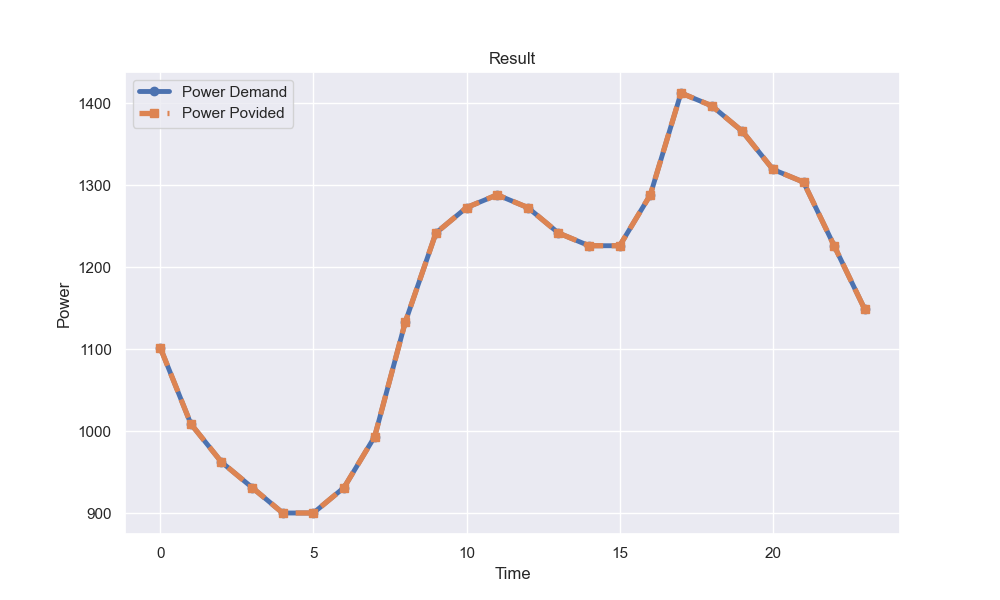
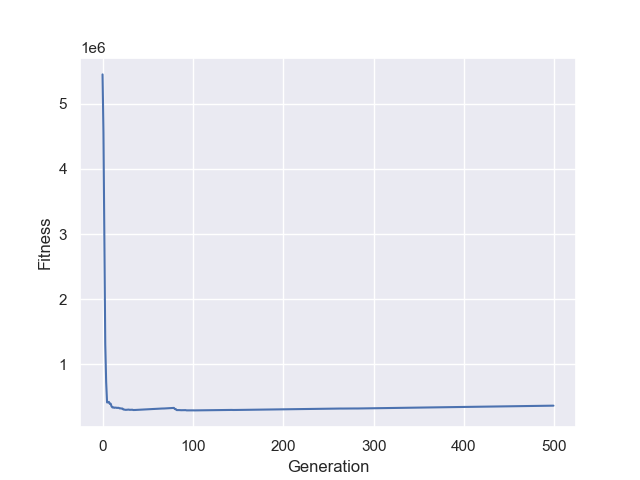


图 遗传算法求解结果

以及适应度（代价）随着迭代不断更新的过程：



最后的多次迭代中适应度趋于平缓，这表明遗传算法搜索到的是某个较大区间的最优解。

# 深度强化学习求解算法

深度强化学习相对于混合整数规划在解决机组组合优化问题上具有独特的优势。深度强化学习能够处理大规模、高维度的问题，并在复杂的决策空间中找到近似最优解。通过神经网络和强化学习算法的结合，它能够自动学习并优化模型，适应不断变化的环境和条件，从而更好地解决实际工程中的机组组合问题。

深度强化学习具有灵活性和泛化能力，能够处理具有不确定性和动态变化的系统。相比之下，混合整数规划通常需要事先对问题进行数学建模，并且对于复杂的大规模问题，求解过程可能面临指数级的计算复杂性，难以有效求解。

已有相关的文献使用深度强化学习求解机组组合优化问题，文献[1]提出使用强化学习辅助树搜索算法求解机组优化问题，文献[2]将指针网络和演员-评论家算法结合，从而将机组在一个周期内的优化问题转变为时间周期上的规律学习。文献[3]在多步骤的深度 Q 学习的集成框架下进行求解机组组合优化问题。该算法通过解决定制化的优化问题建立候选动作集，以确保相对高的性能和满足运行约束条件。

## A. 马尔可夫决策过程建立

作为最广泛使用的数学框架来构建顺序决策问题，马尔可夫决策过程可以被定义为一个元组，这包含了了状态S,动作A,状态转移概率P,奖励(回报)R,折扣因子.这种设置将允许在强化学习中进行高效的利用。在每一个时间步智能体根据观测到的环境状态选择动作,并且得到相应的奖励，为了实现这个功能，智能体必须学习到一个策略使其得到最大的期望奖励回报



然后智能体根据学到的策略沿着状态s选择最佳动作，环境的策略转移概率取决于。

为了描述状态和状态-动作对在策略下的价值，我们引入了状态价值函数和状态动作价值函数









我们定义最佳策略为或者。

通过使用马尔可夫决策过程（MDP）构建机组组合问题，同时给出以下的情节（episode）、状态（state）、动作（action）和奖励函数（reward functions）的定义。

### EPISODE AND TIME STEPS

Episode被定义为强化学习智能体与机组组合环境进行一次完整互动。每个操作周期 t 被定义为一个时间步。由于机组组合问题在日前市场每天解决一次，我们将其构建为持续任务，这意味着一个情节只有在找不到可行的动作时才会结束。

### STATES

为了将环境构建成马尔可夫决策过程，时间t的状态被定义为，其中t表示时间步（0-23），表示在t时刻各个机组的运行状态，这时一个有J维的向量，其中若=-5表示第个机组在当前时刻已经累计停运5小时，这个变量约束可以帮助机组学习最短开停时间的限制，表示t+1时刻的能量需求，这是因为在时间t，智能体需要选择t+1时刻的开停决策与功率输出。

### ACTIONS

关于机组组合优化问题下的动作空间，我们需要特别的关注的是这是一个混合动作空间，既有多维离散二进制变量的机组开停决策过程，也有每个时刻对应的多维连续变量功率输出，由于离散-连续混合动作空间的复杂性和较难收敛，论文大多只关注多维离散输出控制启停作，而不关心其实际的具体功率输出连续空间。如文献[4]提出使用多层Q-learning算法实现控制多阶段机组启停动作，因为Q-learning在不进行离散动作连续化之前只能实现离散动作空间。

同样的，深度强化学习求解机组组合优化问题没有固定的范式，模型算法五花八门，甚至有很多额外的技巧，因为想让一个智能体学习机组组合优化的策略是很复杂的

### REWARD AND PUNISHMENT

强化学习机组组合优化问题的奖励项和惩罚项设置十分重要，因为这个组合优化问题存在最短开停时间限制，这造成了不同时刻的状态和动作在相互约束，同时又有不同时刻的负载能量需求，因此我们将时间周期的变换内嵌在环境中随着智能体的探索实现周期性变化，我们的奖励分成5项



其中 表示t时刻的运行成本， 表示t时间的开启成本， 表示t时间的关闭成本， 表示表示如果动作违反了最短开停时间约束，我们会用一个掩码机制将不符合开停约束的动作进行矫正，同时根据汉明距离计算智能体输出动作不符合最短开停时间约束的程度并计算约束惩罚项。最重要的是负载不同时间周期的能量需求约束，被定义为如果所提供功率小于负载功率要求，则根据提供的能量和负载要求的能量计算负载需求惩罚项。于是Reward的定义为



### MASK MECHANISM

为了实现最短启停时间约束，在上文我们提到过在状态中引入表示变量的过去开停状态对现在的影响，的迭代过程如下：



我们据此实现了一个MASK掩码机制用来监督智能体的动作是否违反了的最短开停时间限制，并将违反规则的动作进行矫正并纳入惩罚项，具体计算过程如下





上式和表示经过对最短启停时间分析，下一个状态各个机组若在受影响的时间范围内所必须保持的状态，若不在受影响的时间范围内则不受约束。MASK的掩码机制需要对原始智能体动作输出进行过滤约束。



其中表示异或逻辑。表示点积，表示取反。

# DEEP Q-learning for Units Unit Commitment Problem

本学期学习了动态规划算法。动态规划和 Q-learning 是两种解决问题和优化过程的不同算法。动态规划是一种解决多阶段决策问题的方法，将复杂问题分解为更小的子问题，并通过解决子问题来获得原问题的最优解。其特点包括具有最优子结构和重叠子问题性质，通常适用于这类问题。

动态规划基于贝尔曼方程，考虑一个最优问题，其状态表示为S,动作表示为A,转移概率表示为P,奖励为R,折扣因子为，动态规划的核心过程可以写成如下形式：



其中，表示状态s处的最优值函数，表示所有可能动作a求最大值，表示状态s执行动作a之后转移到的概率，状态s执行动作a之后转移到所获得的即时奖励，折扣因子，用于平衡当前和未来奖励的重要性。

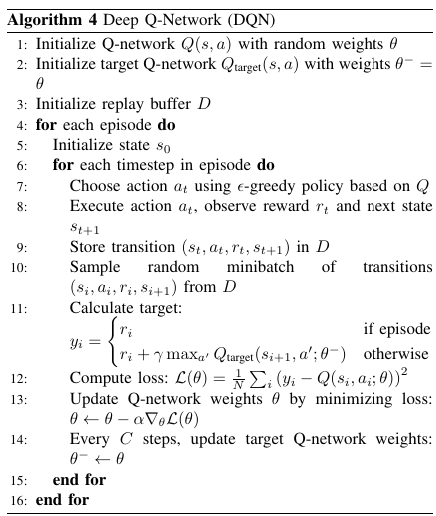
Q-learning是基于值迭代的强化学习算法，其核心更新规则如下所示：



表示在状态s执行动作a的Q值，是学习率，用于控制新获得信息的重要程度，在状态s执行动作a的即时奖励，折扣因子，平衡未来奖励的重要性，表示在下一个状态选择具有最大Q值的动作。

在能源调度的相关优秀论文中，有很多提出使用动态规划+策略价值评估对机组组合，能源调度等问题进行求解的，如文献[x] 利用了近似动态规划和深度递归神经网络学习算法，以解决微电网的动态能源管理问题。其重点在于实时对微电网进行经济运行，开发了一种新颖的动态能源管理系统，将能量存储系统的有效管理融入到微电网的实时调度中，同时考虑了负载、可再生能源发电和实时电价的不确定性以及功率流约束。该动态能源管理机制无需长期预测和优化，也无需了解不确定性的分布知识，但仍能优化微电网的长期运行成本。首先，将实时调度问题建模为一种在一天内的有限时间段的马尔可夫决策过程。然后，采用近似动态规划和深度递归神经网络学习来推导出近似最优的实时调度策略。

在此次课程设计中，我们按照近似动态规划与价值评估的思想，设计了深度Q-learning网络实现了机组组合优化的控制问题，伪代码如下：



DQN只能实现离散动作控制，所以我们只能使用DQN控制24小时机组的开停动作，也即类似[00011101]这样的8位二进制数组，在环境约束过程中，我们只要求某一时刻所有开启的机组的最大输出功率之和大于等于负载需求即可，同时最小化运行成本和各种约束惩罚项。之后一旦机组开停策略确认，我们只需建立一个很简单的线性规划求解成本最小的实际输出功率策略，这个线性规划我们使用python里面的CBC求解器实现，约束条件为开启机组的输出功率区间，开启机组输出功率和满足负载需求，目标是最小化运行成本，于是，我们按照上面的思路编写相关的代码，运行测试结果如下：

DQN结果如下，计算总代价为$487,183：

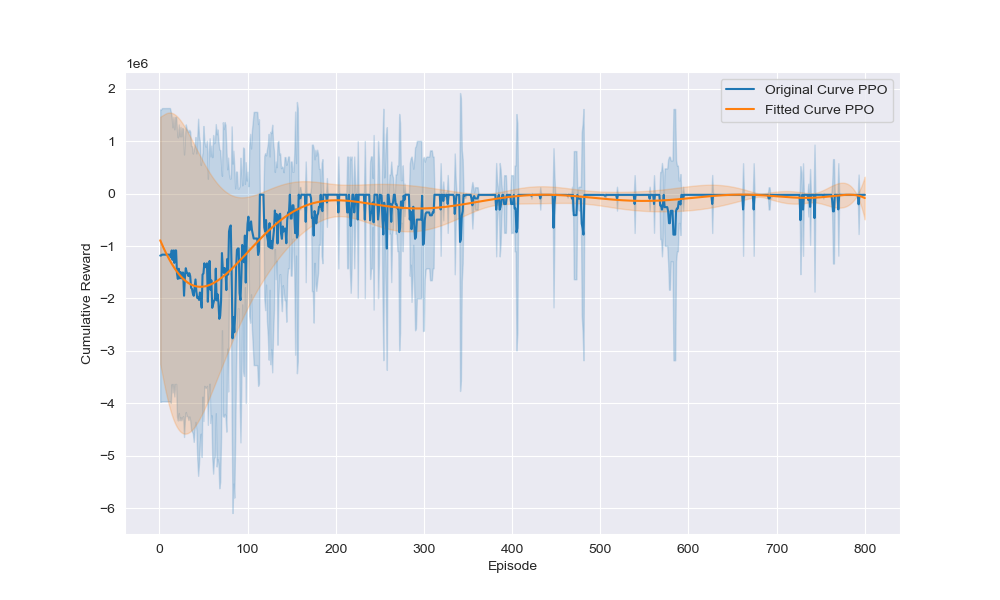


图 DQN 奖励收敛曲线

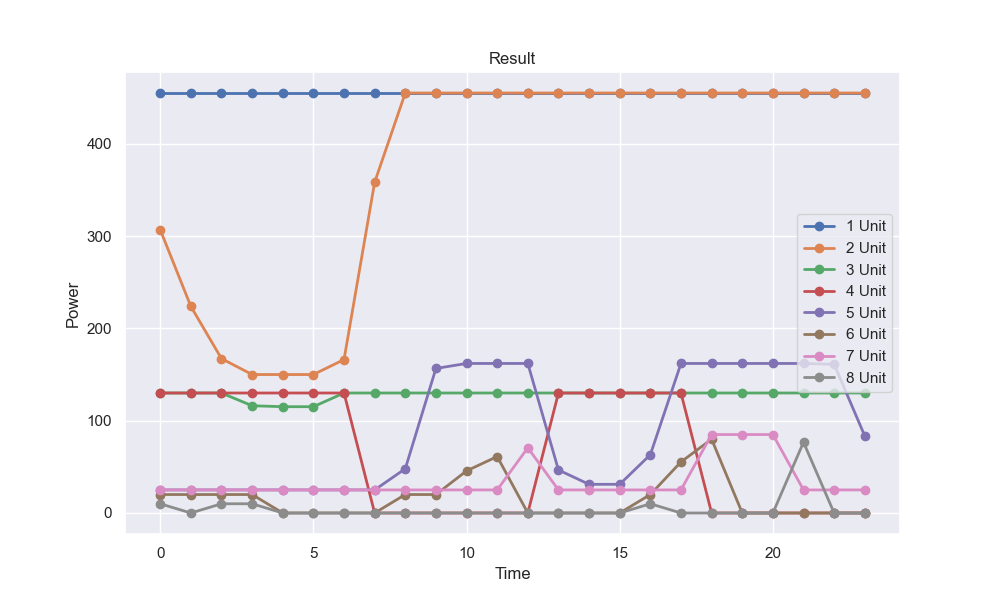


图 DQN 机组控制结果

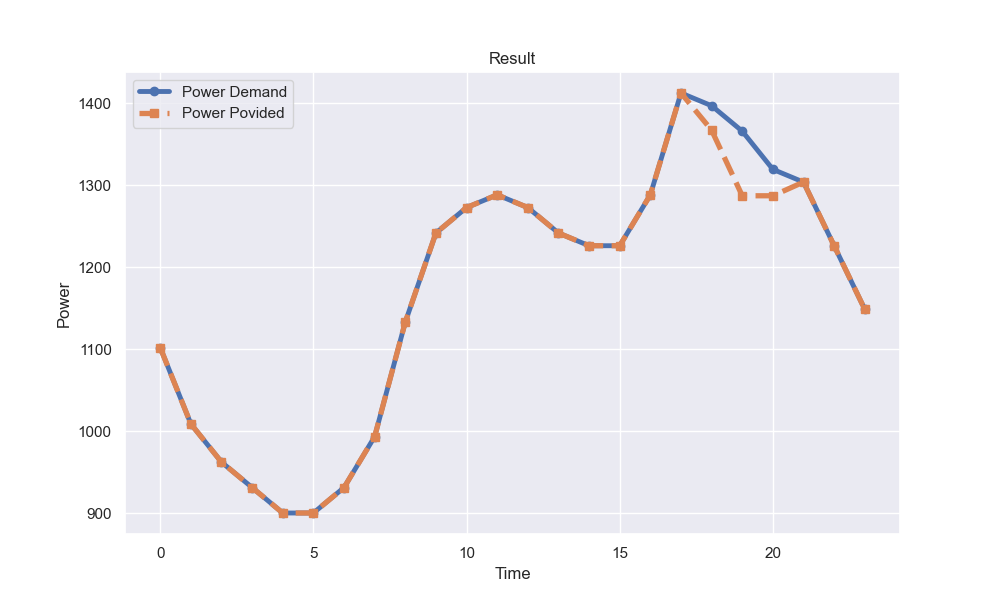


图 负载功率需求和提供的功率

由DQN的结果图可以看出，结合了线性规划和DQN共同求解机组组合优化问题，其策略和混合整数规划大致是相同的(各个机组的功率状态折线图近似)。在本次课程实践中，由于时间上的限制，我们并没有深入挖掘关于Q-learning算法求解机组组合优化问题的背景和研究现状，但从参考的论文来看，这方面研究内容还是挺丰富的，有很多学者提出了非常新奇有效的算法，这有待我们在接下来的时间中深度学习理解并掌握。

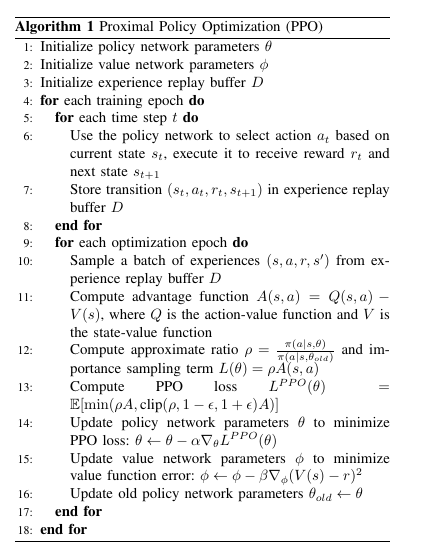
# Proximal Policy Optimization实现强化学习机组组合优化

除了值函数估计的算法，强化学习还有一个重要的算法方向就是策略梯度，在本次课程实践中，我们使用了策略梯度算法中的PPO算法实现机组组合优化问题的求解并得出相关结果，本节我们将介绍PPO算法的原理，介绍怎么利用PPO算法实现多维二进制动作输出并更新模型参数，介绍使用PPO算法得出的机组组合优化策略结果，包括分析奖励曲线，机组控制结果以及负载需求和实际输出需求对比图。

### PPO(Proximal Policy Optimization)

PPO(Proximal Policy Optimization)是强化学习中的一个重要算法,主要思想是每次只对策略做小幅度的更新,以保证新的策略不会偏离老策略太多。其具体步骤如下:

（1）收集数据:与环境交互,收集轨迹数据。（2）计算优势函数:基于收集的数据,计算每条轨迹的优势函数。优势函数反映某个动作相对平均动作的优势。（3）裁剪优势函数:将优势函数的值限制在一个范围内,避免极大值对学习带来不稳定性。（4）最大化策略优势:通过最小化旧策略和新策略的KL散度,最大化策略的优势函数。这可确保策略只作较小更新。（5）更新策略:结合裁剪后的优势函数,通过梯度上升来更新策略网络。PPO简单高效,既确保样本效率,又不易发生剧烈的策略变化,在多种任务中表现优异。其算法的具体流程如下：

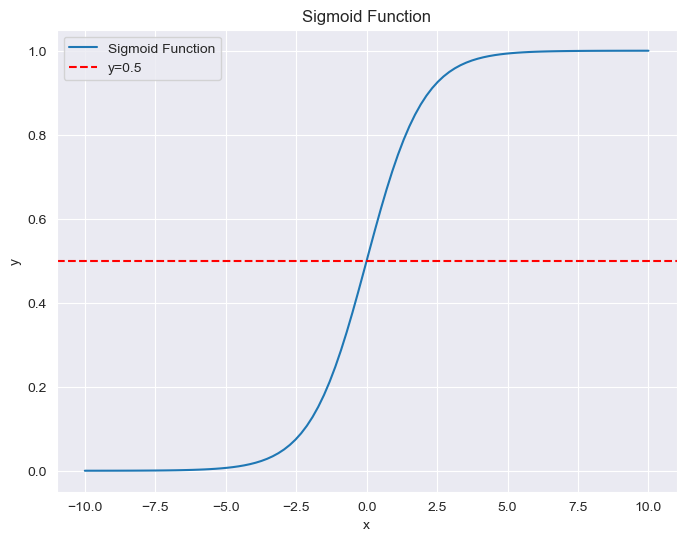


在本次课程设计中，有关多维二进制动作的的PPO算法部署实现并没找到相关参考文献或开源代码，但是我们独自设计出两种算法实现PPO策略网络多维二进制动作的输出并更新模型参数，最终我们分析其计算结果。这个过程中存在很多探索，下面将一一介绍。

### Sigmod函数输出及其对应的参数更新策略

我们选择让策略网络每次输出一个8维的向量，这些向量最后经过一层sigmod函数。其数学表达式为





如图中红线所示，这是我们的动作分割点，相当于我们将联系的8维输出离散成一个8维二进制变量，如[0.8,0.4,0.9…]离散成[1,0,1…]分割点是0.5。至于参数更新过程，原始PPO算法只是在一个N维的动作选择某一个动作，这N维的动作空间对应N个概率向量，其总和是1，我们只需要最大化其价值最大的那个动作对应的概率值即可，如下，这种策略网络最后一层都是Softmax函数归一化操作。



在我们设计的Sigmod函数策略网络输出过程中，由于我们设置的是大于0.5取开动作，于是我们在原始的PPO算法基础上进行了一些小改进。



其中表示原始的8维动作概率分布，事实证明我们的算法是可行的，模型确实在一开始表现为奖励曲线提高趋势，但是这样的操作会带来梯度消失的现象，使其不能深入训练，于是我们放弃这个方案转而实现另一种PPO多维二进制动作输出的方法。

### 二维动作概率分布输出

前面我们提到PPO算法原始论文只针对一维动作概率分布实现而非一次选择多个离散动作，但是我们可以让策略网络输出一个大小为（8，2）的数组并在第二为进行softmax归一化，并对其采样生成一个长度为8的二进制向量，概率数组的第二维表示对应每个机组有两种概率，一个开一个停。这样我们仍需要修改原始的PPO参数更新过程，原始PPO算法只需要计算N维动作中某1维的概率并利用其概率对数值代入公式更新，而我们需要计算(N，M)维中某M维动作的概率组合，这里N=2,M=8，表示8个机组的两种概率分布状态，于是我们采用贝叶斯网络中的极大似然函数的思想，每个开启机组对应的概率相乘作为该动作的整体概率分布。之后沿用PPO离散动作的策略更新方式。我们利用这个算法在我们的机组组合优化环境上测试，奖励收敛曲线如下：

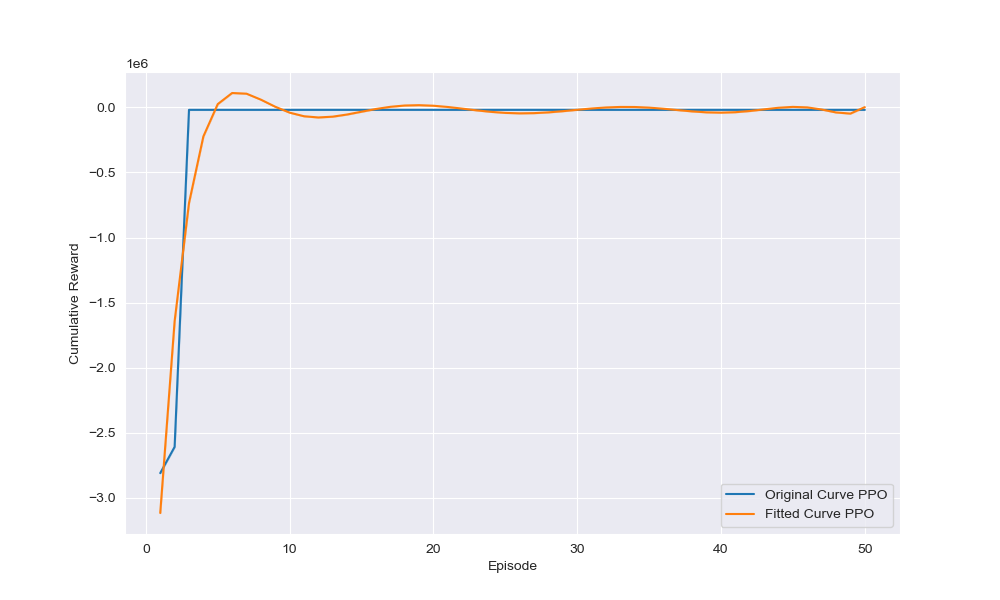


图 PPO算法对于机组组合优化问题的奖励曲线

从图中可以看出所设计的算法在本题的环境背景中有着极强的收敛趋势，但之后就停止更新了。为了简便起见，我们直接调用Stabel-baseline3强化学习算法库，自定义机组组合优化问题环境，调用算法库中的PPO算法，该算法支持多维二进制动作，仍然采样强化学习和线性规划结合的方式求解，经过一段时间训练后，所得出的结果如下图。

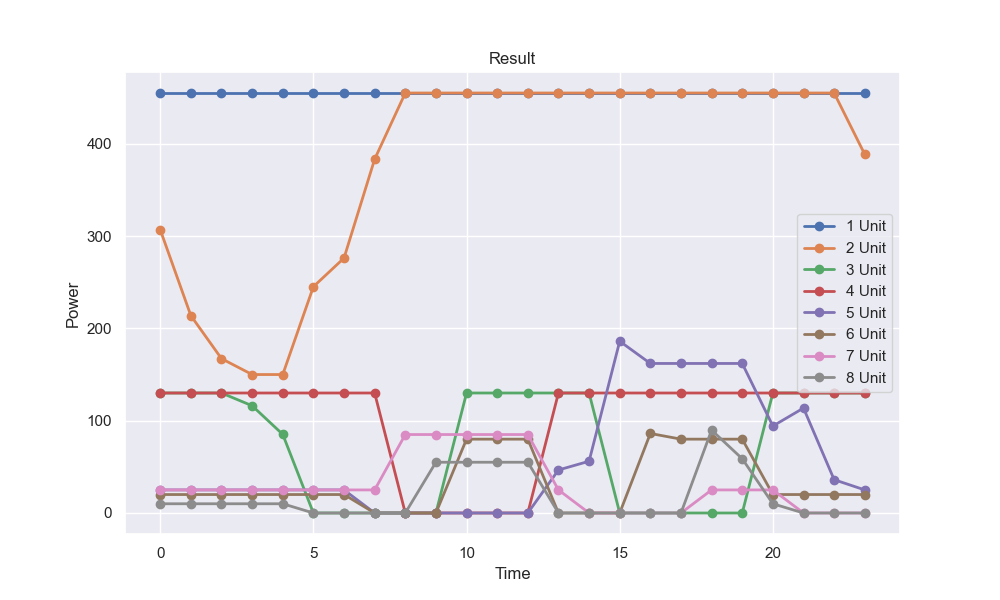


图 PPO算法控制机组策略

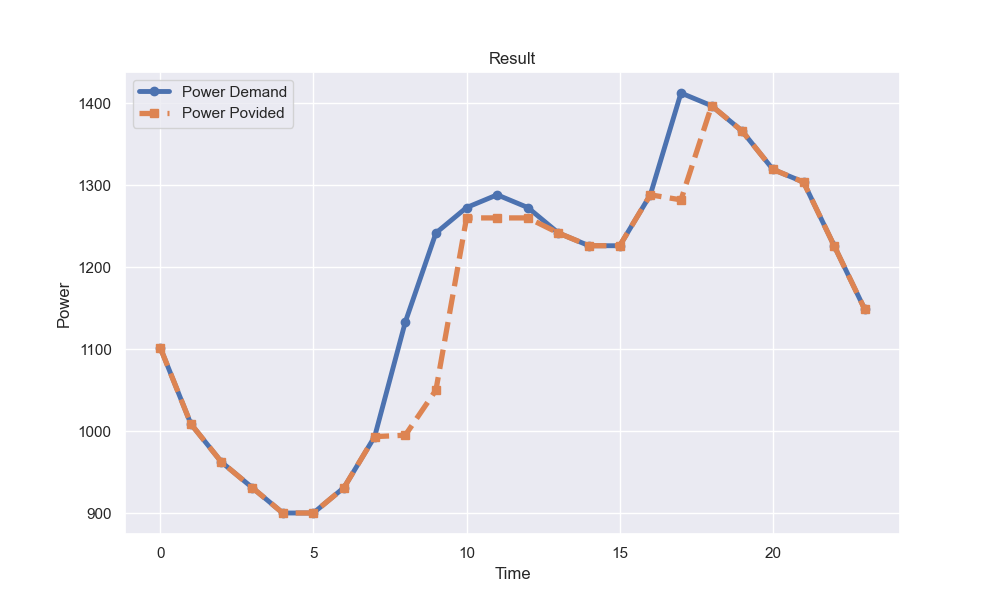


图 PPO算法控制机组输出功率和负载需求

经计算，PPO算法运行的策略成本为：$485,001

# H-PPO强化学习算法实现机组组合优化

前面我们提到了，无论是DQN还是PPO,我们都是只控制离散开停动作，并结合解析解的方式求出该开停动作对应的输出功率。最近的参考文献中提出一种将离散连续动作结合的强化学习算法H-PPO。H-PPO算法分成两部分策略，一个离散部分，一个连续部分进行优化。这个混合架构包括多个并行的子演员网络，将结构化的动作空间分解为更简单的动作空间，并且还有一个评论家网络来引导所有子演员网络的训练。虽然主要关注参数化动作空间，但提出的架构可以扩展到具有分层结构的更一般的动作空间。

## A. 强化学习中的离散动作和连续动作

Q-learning是一种基于值函数的更新方法，它基于贝尔曼方程更新模型参数

 在离散动作空间领域中，深度Q网络（DQN）采用了一个框架，并利用深度神经网络来近似Q函数。在离散动作空间中，也广泛使用了一些DQN的变体，包括异步DQN、双重DQN和交错DQN。这些变体都是基于DQN框架的改进版本，在解决离散动作空间问题上发挥着重要作用。

策略梯度是强化学习的另一种算法，通过优化策略网络参数来实现最大化策略价值，策略梯度算法的公式可以被定义为



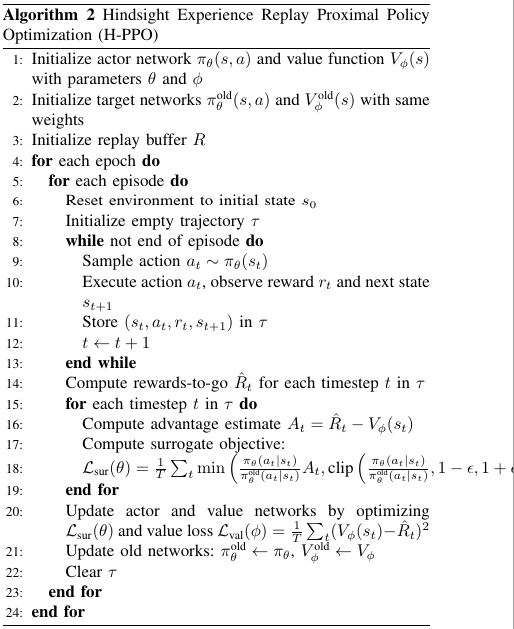
作为一种替代方法，策略梯度也可以利用优势函数来计算。



## B. H-PPO

首先，考虑到参数化动作空间，其中离散动作集合，每个都对应一组连续参数一个完整的动作由表示，整个动作空间 A 则是每个离散动作与其所有可能参数的组合。这种参数化动作空间可被称为参数化动作马尔可夫决策过程（PAMDPs）。

为了设计适用于 PAMDPs 的算法，采用了基于演员-评论家架构的方法。通常的演员-评论家算法包括一个演员网络和一个评论家网络，其中评论家网络用于计算演员网络参数的梯度。然而，对于参数化动作空间，我们提出的架构包含两个并行的演员网络。这两个演员网络分别用于执行动作选择和参数选择：一个离散演员网络学习一个随机策略和离散动作，另一个连续演员网络学习一个随机策略和连续动作，两个演员网络共享前几层以编码状态信息。我们将这种架构称为混合演员-评论家架构，因为在这个架构中同时存在离散演员和连续演员。



于是我们定义了H-PPO算法混合动作网络，在每一个状态下输出启停二进制控制变量，也输出实际功率控制变量，我们通过将离散动作连续动作分开计算loss，并最终合并到Actor-Loss里面实现策略网络的共同控制。

## C. H-PPO

我们根据H-PPO建立相关算法，并在机组组合优化环境上训练，在此问题背景中，H-PPO网络是输出既有控制8个机组的开停概率，也有两个8维的连续向量，这8维连续向量分别对应这8台机组实际输出功率的均值和方差，我们只需要利用高斯分布采样就可以得出其控制功率输出大小，值得注意的是我们设置的8个连续输出变量大小范围约束在了[0,1]故需分别对其反归一化才能得到实际输出功率。但是，结果显示我们的算法根本收敛仍是只收敛一部分，之后就停止了，且训练时间一长就会梯度消失

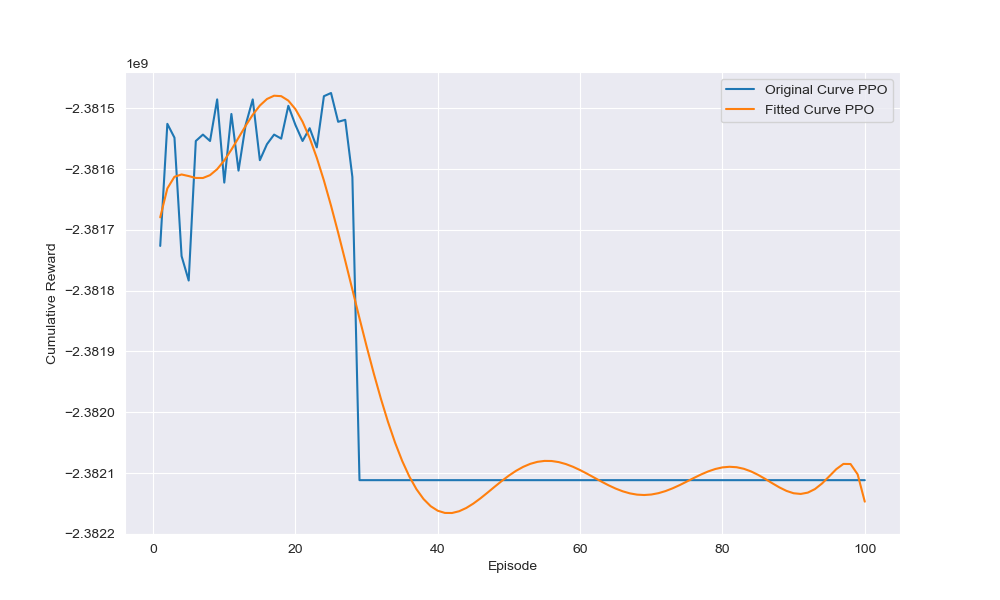


图 H-PPO奖励曲线图

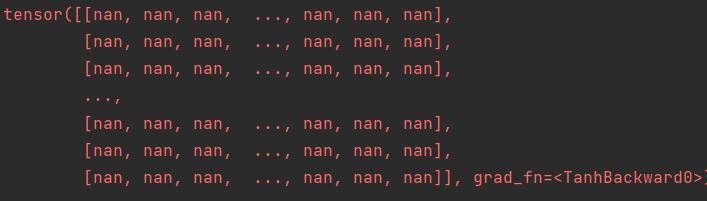


图 H-PPO的梯度消失现象

# VIV Conclution

经过本次课程实践大作业收获良多。我们掌握了在matlab和python里面求解线性规划，混合线性规划以及相关极值约束问题的具体做法。在使用遗传算法求解机组组合优化问题时，我们自学遗传算法的原理，在不调用外接库的条件下自定义了整个遗传算法的流程，得出的运行成本结果和混合整数规划极为接近。

我要强调的是强化学习求解机组组合优化问题，其实经过调研，我们本次实现的机组组合优化问题条件十分简单，真实的机组组合优化问题可能包含代价的功率是二次函数，翻转率限制，功率上升曲线斜率，分段斜率限制，冷启动和热启动成本以及最长运行时间等等，然而加入这些条件后利用混合整数规划求解不仅耗时较长，求出的也不是最优解，这时很多论文就提出了强化学习求解机组组合问题的算法，无论是求解效率还是运行成本上都有证明强化学习的效果可以好于gurobi等优化器计算的结果[3]。这也是我们一开始选择强化学习算法求解的原因。然而强化学习算法较为复杂，在短暂的学习之后很难熟练的运用，求出的结果也并没有想象中的那样合理，无论如何，这都是一次极为宝贵的学习和探索经历，

References

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems* (Book style)*.* Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.
5. E. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” *IEEE Trans. Antennas Propagat.*, to be published.
6. J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” *IEEE J. Quantum Electron.*, submitted for publication.
7. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
8. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style),” *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].
9. M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
10. J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility (Periodical style),” *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34–39, Jan. 1959.
11. S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique for digital communications channel equalization using radial basis function networks,” *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.
12. R. W. Lucky, “Automatic equalization for digital communication,” *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.
13. S. P. Bingulac, “On the compatibility of adaptive controllers (Published Conference Proceedings style),” in *Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory*, New York, 1994, pp. 8–16.
14. G. R. Faulhaber, “Design of service systems with priority reservation,” in *Conf. Rec. 1995 IEEE Int. Conf. Communications,* pp. 3–8.
15. W. D. Doyle, “Magnetization reversal in films with biaxial anisotropy,” in *1987 Proc. INTERMAG Conf.*, pp. 2.2-1–2.2-6.
16. G. W. Juette and L. E. Zeffanella, “Radio noise currents n short sections on bundle conductors (Presented Conference Paper style),” presented at the IEEE Summer power Meeting, Dallas, TX, June 22–27, 1990, Paper 90 SM 690-0 PWRS.
17. J. G. Kreifeldt, “An analysis of surface-detected EMG as an amplitude-modulated noise,” presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.
18. J. Williams, “Narrow-band analyzer (Thesis or Dissertation style),” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
19. N. Kawasaki, “Parametric study of thermal and chemical nonequilibrium nozzle flow,” M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
20. J. P. Wilkinson, “Nonlinear resonant circuit devices (Patent style),” U.S. Patent 3 624 12, July 16, 1990.

1. [↑](#footnote-ref-0)