过热蒸汽温度PID控制器参数的RL整定方法

**摘要：**由于负荷频繁而广泛的变化带来的未知干扰以及对效率和安全的严格控制要求，过热蒸汽温度(SST)的控制越来越具有挑战性。随着可再生能源的普及以及其输出过程的不确定性，这种情况变得更加严重。为此，本文描述了SST的工作原理，并根据开环阶跃响应数据识别了SST的模型。针对系统对高阶动力学扰动响应缓慢的问题，我们采用了深度强化学习算法来学习有效的自适应增益调整策略，因为它不需要任何关于电厂热力模型的先验知识，提出了一种改进的自抗扰控制器，以提高控制性能。从理论上推导了修正自抗扰控制器的稳定性分析，完善了修正自抗扰控制器的理论。然后，总结了一种实用的自抗扰控制器调整程序，使现场工程师可以很容易地理解。仿真实例表明，改进的自抗扰控制器在保持良好鲁棒性的同时，能够提高跟踪性能和抗干扰性能，并且针对SST控制系统和基于所辨识模型进行了仿真，初步验证了控制性能。本文提出的自抗扰控制系统方法的可行性，预示着随着可再生能源并网需求的不断增加，自抗扰控制可在电力行业中得到广泛应用。

# 引言

随着电力需求的不断增长和可再生能源在电力市场占比的不断扩大，电厂的过程变量控制一直是重要且具有挑战性的问题。根据预测，全球电力装机总容量将以约2.2%的速度增长至2040年。为了缓解电力行业对化石能源的依赖，使用清洁能源如太阳能、风能和潮汐能成为了人们的选择。预计到2040年可再生能源发电占比将达到25%。然而可再生能源的发电特性受环境影响因素较多，给并网带来了一定的困难。解决这个问题的一种方法是提高火电厂对自发电控制（AGC）系统的响应速度。为了配合新能源系统的发电特性，火电厂的负荷需要频繁变化，这就对过热蒸汽温度（sst）在安全范围内的有效控制提出挑战。SST是电厂机组运行过程中安全和经济性的重要指标。当过热蒸汽超过一定温度时会对机组造成不可逆的损伤甚至停机，低于一定温度则会降低机组的热效能。一般认为其温度不应超过设定值的正负5摄氏度。一个性能强大的控制器可以使电网融入更多的可再生能源。

面对SST的高阶动态、非线性和多源干扰等带来的控制困难，传统的固定参数的PID控制策略不能保证较好的控制性能，特别是在大范围负载调节的情况下。许多改进和先进的控制策略已经被开发来处理上述困难。例如神经网络PID，内模控制，自适应PID，分数阶PID。高级控制策略例如多变量约束，模糊逻辑，动态矩阵和预测前馈也被用于提高控制性能。但是这些方法只在模拟过程中有作用，很少用于实际的发电过程。主要原因有：

1. 这些控制方法需要复杂的计算，一般的分布式控制系统难以实现。
2. 一些控制策略对过热蒸汽模型的精确性依赖程度较高，但是精确的模型往往难以实现，且会随着环境和工况而变化。

自抗扰控制有着良好的抗干扰性和非线性适应性，且具有不依赖被控对象的数学模型和易于在DCS中实现的特点，故受到很多关注。且已经成功应用于动作系统，直流转换系统以及化学和温度控制过程。本文实现了应用于复杂对象模型的属于自适应控制的强化学习方法，对控制器的PID参数进行了动态控制。

然而，自抗扰控制系统处理高阶系统带来的困难的能力是有限的。具体而言，其局限性是自抗扰控制策略虽然与传统控制策略相比具有抗干扰优势，但对高阶系统的设定点响应较慢，这一点将在第4节的数值模拟中得到证实。针对这些控制难题，提出了一种改进的自抗扰控制(MADRC)，并提出了MADRC- pi级联控制策略，以提高在役循环流化床机组SST的控制性能。此外，本研究还以创新的方式获得了以下有趣的结果:

1. 针对高阶动态所带来的控制困难，提出了一种自适应自适应控制算法，提高了系统的跟踪能力和抗干扰能力，具有良好的鲁棒性。
2. 对MADRC进行了稳定性分析。此外，本文还给出了一个参数的理论范围。
3. 通过分析自适应控制参数对控制性能的影响，总结了一种实用的自适应控制方法，并通过数值仿真验证了自适应控制的跟踪和抗干扰能力。
4. 基于MADRC的串级控制方法已成功应用于某循环流化床机组，运行数据证实了该方法降低温度偏差、扩大负荷调节范围的可行性。

本文的其余部分组织如下:下一节简要介绍SST控制系统，并根据开环实验数据识别出串级模型。第三节介绍了常规的自抗扰控制器设计，随后介绍了提出的自抗扰控制器，其中的稳定性分析从理论上推导。第四节给出了MADRC的实际调谐过程，并进行了仿真验证其控制性能。在第5节中，将基于MADRC的SST控制系统串级控制应用于运行中的CFB电厂，结果表明在降低温度偏差方面有显著改善。结论在第6节中得出。

# 用于本文中的SST模型

## 2.1．SST模型综述

电厂过热器蒸汽系统如图1所示。该系统包含两级减温装置:一级减温器和二级减温器。水冷壁中的水变为蒸汽后经过减温器和过热器并且从从烟气中获得热量。火电厂主要的汽温控制方法为水雾喷洒。冷却水来源是锅炉给水泵的中间部分。一级和二级减温器由串级回路控制，反馈回路为它们的进口温度和出口温度。将分布参数过热系统近似化为两个线性传递函数模型和。内环传递函数表示过热器到二级减温器的过程。外环传递函数表示二级减温器到出口温度的过程。γ是二级减温器出口水温的设定温度。操作变量为水雾阀门开度，即内环控制器的输出。外环控制器的输出是内环的输入。

内环扰动主要来自减温水的温度和压力变化，将其归结为阶跃型扰动。外环扰动主要有负荷调节、煤质变化、燃烧波动等，并且通过二级过热器影响一次汽温，这里用阶跃扰动表示。基于这些条件考虑，SST的控制难度有以下几点：

* 1. 二级减温器的动力学模型阶次较高，对扰动的响应缓慢。
  2. SST的扰动是多源的。负荷需求、主汽流量、烟气流量等诸多因素都会影响主汽温度。
  3. 系统复杂性较高，难以建立精确数学模型。
  4. 由于火力发电厂机组通常运行在半负载到满负载范围内且变化频繁,SST系统与操作条件急剧变化,需要控制系统具有良好的鲁棒性。另外机组负载产生的强非线性也会对控制性能产生不利影响。

基于上述条件，SST控制系统应保证具有较强的抗干扰能力，即保持良好的鲁棒性，同时对数学模型的准确性不严格依赖。

## SST模型辨识

系统理论中SST属于参数分布式系统，在稳定工况中通过开环实验参数辨识，近似得到SST的线性模型，其工艺过程如图1，方块图如图2。在某个燃煤电厂中得到300MW机组的SST模型为：

|  |  |  |
| --- | --- | --- |
|  |  | （1） |
|  |  | （2） |

上式中，，分别为SST内环和外环模型的传递函数，和是内环传函的动态参数，和是外环传函的动态参数。这些参数都是通过辨识来确认的。

采用多目标遗传算法优化这些动态参数，同时使SST模型的误差与实际降到最少，最后得到模型为：

|  |  |  |
| --- | --- | --- |
|  |  | （3） |
|  |  | （4） |

由模型识别得到的传递函数的响应与实际过程中得到的相比误差较小，由此可见该SST传递函数可用于控制器设计。

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 图（1）过热汽温控制过程 |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 图（2）SST控制方块图 |  |

# 应用强化学习的控制器设计

过热蒸汽温度调控主要采用传统的PID控制器。然而，其方法有两个主要的局限性。首先，PID控制器的结构中有P、I、D增益，需要将它们作为参数进行调整。这些参数通常通过经验丰富的人员手动进行调整，但往往不能达到最佳效果。为了实现PID参数整定的智能化，人们进行了大量研究，典型方法有齐格尔-尼格尔斯（ZN）方法、戴维森（Davison）方法、Penttinen方法和Maciejowski方法。然而，这些方法都具有局限性，首先都会使调节增益保持固定。其次当负载发生时变，传统PID控制器的性能会受到限制。而SST处于多变的运行环境，因此增益固定的PID控制器不能有效的执行调节任务。处理负载时变的有效方法是自适应，本文采用自适应PID控制器微调系统增益，其特性应有以下几点：

1. 不应该依赖人类基于经验的增益调整策略。
2. 对SST控制系统的执行机构不做出频繁而剧烈的调整，即良好的控制效率。

为实现第一个概念,采用深度强化学习(DRL)算法以及深度策略梯度(DDPG),因为它不需要任何先验知识，而是通过与环境的交互获得增益的调整策略。为实现第二点，控制器有效增益范围应为，其中为基本增益，是通过PID增益调整方法如手动调整方法，ZN方法，或任何PID增益调整方法预调整增益。在这两条要求约束下，控制器必须在有限范围内输出，保证了其稳定性。该系统中的“微调”一词来自深度学习中的卷积神经网络。

有关深度强化学习（DRL）控制器的研究，其已经在机器人控制中得到应用，在Carlucho et al.(2020, 2019)和Gloye et al.(2005)的研究中，具有DRL的机器人通过自适应调整增益，成功地学会了通过快速收敛实现快速参考跟踪。接下来的部分主要讨论强化学习理论以及提出的自适应微调系统，然后给出了仿真结果和讨论。

## 3.1. 强化学习理论

在标准强化学习(RL)中，有一个代理和一个环境，代理与离散时间步相互作用。在每个时间步，代理接收一个状态，采取一个行为，并且产生一个奖励，在每个周期产生顺序列的状态转换。采取的行为由π策略（即将状态映射到每个动作的概率）决定，采用概率是因为需要确保在状态和动作空间中探索。代理的学习过程可用马尔科夫决策过程(Markov Decision process, MDP)数学定义。该过程的一般流程图如图3所示。在图中，随着代理与环境的持续交互，策略π可以通过最大化所获得的报酬来学习到越来越好的策略。

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 图（3）代理使用MDP的学习过程 |  |

许多RL算法都采用动作值函数（即Q函数）。该函数返回预期的奖励值如式5。

|  |  |  |
| --- | --- | --- |
|  |  | （5） |

其中γ是一个折现因子，用于考虑在系统侧重于未来的高回报还是当下的利益。Q函数可以用Bellman方程(Bellman, 1954)写成递归的形式：

|  |  |  |
| --- | --- | --- |
|  |  | （6） |
|  |  | （7） |

如果用确定的策略μ将状态映射到每个动作，则式6可写为式8。

|  |  |  |
| --- | --- | --- |
|  |  | （8） |

其中可以通过Watkins和Dayan(1992)提出的一种称为Q- learning的非策略学习算法来学习，即由潜在不同的随机行为策略β生成其转换，其策略可使用贪婪算法。Q学习的目标是学习一种策略，使期望Q函数最大化。但是Q学习的难度会随着状态-奖励空间的增大而提高。因此，通常使用神经网络作为Q函数的模拟器。该模拟器由参数化，可以通过最小化下面的损失函数来优化。

|  |  |  |
| --- | --- | --- |
|  |  | （9） |
|  |  | （10） |

表示由构成的状态访问分布。

存在的问题是将Q-learning应用于连续动作空间是不可能的，因为寻找贪婪策略需要在每个时间步上优化。策略可以通过使用DPG算法(Silver，2014)在Q的梯度方向上进行更新，采用“动作-评分”方法，其中动作由称为actor的神经网络逼近，评分（Q）由另一个称为critical的神经网络逼近。在动作-评分方法中，动作和评分用和参数化，策略由动作表示。

actor是通过应用式11使J最大化来学习的，而critical是通过Bellman方程学习的，就像Q-learning中一样。

|  |  |  |
| --- | --- | --- |
|  |  | （11） |

然而，如果观测空间是高维的，这种行为-批评方法在收敛上存在困难，并且学习不稳定。Lillicrap为了解决这一问题提出了DDPG，它是基于“动作-评分”和深度Q网络 (DQN)的结合。DQN用于DDPG的两个主要特性是重置缓冲区和目标网络。重置缓冲是一个存储的转换（），目标网络是动作-评分的复制，它的权重随着动作-评分的权重而更新。结果表明，利用重置缓冲和目标网络可以显著地稳定和改善学习过程。

## SST系统中PID增益自适应微调系统的提出

所提出的自适应微调系统的总体结构如图4所示。

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 图（4）自适应微调系统的总体结构 |  |

在该图中，状态是动作和评分的输入向量，是自适应PID增益的向量，是误差积分的更新。的构型如式(12)所示。在中使用前面动作的历史记录来处理记忆效应（即当前动作由前面任何时刻的动作决定）。

|  |  |  |
| --- | --- | --- |
|  |  | （12） |

误差的积分也包括在中，因为它影响输出的平均位置的保持，并且它需要足够的信息去更新自身。 在对动作和评分进行训练时，需要将归一化来保证训练的稳定性（式13）。

|  |  |  |
| --- | --- | --- |
|  |  | （13） |

其中为存储在缓冲区中的所有状态，表示标准差，中的误差积分项归一化的方法是将其除以100，使其足够小以稳定的神经网络训练过程。

动作在每一步输出和，包括一级过热器至二级减温器，二级过热器至出口的自适应PID增益。的具体内容如式14所示，其中每个由式15所示的各个环节的自适应增益P、I和D组成。

|  |  |  |
| --- | --- | --- |
|  |  | （14） |
|  |  | （15） |
|  |  | （16） |

其中是基于ZN方法确定的基增益的向量，是动作部分输出层的输出向量，其激活函数sigmoid见式16。因此，和的范围分别是[0,1]和[0,]。用于自适应PID控制器的函数为式17，其中上括号用于标注方向。

|  |  |  |
| --- | --- | --- |
|  |  | （17） |

用来确定PID控制器的积分项有多少误差要更新。更新方程见式18，是为了防止SST调节动作过大而设计的。

|  |  |  |
| --- | --- | --- |
|  |  | （18） |
|  |  | （19） |

|  |  |  |
| --- | --- | --- |
|  |  | （20） |
|  |  | （21） |

动作和评分部分的架构如图5所示，HL代表隐含层，下方数字代表神经元个数，BN代表批量归一化。合理的奖励函数是agent学习有效策略的关键，决定了RL的收敛速度和稳定性。训练动作和评分的奖励部分定义为式20。为了进一步稳定RL算法，将奖励分为[0,2]，并且用式21进行归一化。表示当前存储的所有奖励。使用标准差为0.1的高斯噪声模拟动作的噪声。其它超参数则使用原型中的定义。

## 3.3具体化系统构建

首先构建DQN（深度Q网络），他的作用是给定（输入）一种状态，它将输出对每一个可能的行动和在执行该行动后(但在看到结果之前)所期望的未来奖励的总和进行估算。要使用这个DQN选择一个动作，我们只需选择预测Q值最大的动作。然而，为了确保代理能够探索环境，选择一个概率为epsilon的随机行为，即产生一个随机数，如果它小于epsilon则返回输出范围内的随机值，否则返回Q网络的输出值。另外还需要一个记忆重播区，它存储了代理的经验并且以元组的形式存放，本文采用DeepMind的开源deque类(双链表）以实现鲁棒性较强的经验重播，创建采样函数来从记忆重播区中取样，它将返回几类数组包括观察值、动作、奖励、预估观察值和结束标志。5个NumPy数组:[obs, actions, rewards, next\_obs, doones]。

奖励是指agent根据任务目标衡量其绩效的奖励信号。合理的奖励函数是agent学习有效策略的关键，决定了RL的收敛速度和稳定性。奖励功能可以分为三种类型:连续奖励功能、离散奖励功能和混合奖励功能。连续奖励功能随着观察和行动而不断变化。一般来说，连续的奖励信号可以提高训练过程中的收敛性，简化网络结构。离散的奖励函数随观察和行动而不连续地变化。这种类型的奖励信号会减缓收敛速度，需要更复杂的网络结构。常用的奖励结构为混合奖励，系统对糟糕的状态进行远离而通过连续平滑奖励接近目标状态加速收敛。

|  |  |  |
| --- | --- | --- |
|  |  |  |

r1是一个连续的奖励，可以稳定系统状态，减少减温水消耗。误差越小，r1的回报就越大。r1中的第二项表示减温水消耗。当过热蒸汽的温度非常接近设定值时，误差变化很小，r1的变化很小。因此，在各误差分量绝对值小于0.1的情况下，引入连续奖赏r2来增加奖赏梯度，从而快速、准确地接近目标值。r3是一个离散奖励，可以控制SST温度不超出范围，提高训练速度。

代理接收来自环境的观察和奖励，并向环境发送操作。本文采用经验重播（创建一个体验回放缓冲区来存储历史体验，然后随机抽样更新动作和评价网络）作为代理的学习方法。经验重放缓冲区的存在帮助代理能够学习以前的经验，提高样本利用效率。随机抽样可以打破样本之间的相关性，使agent的学习过程更加稳定。

# 4.仿真结果

在仿真中，利用python求解SST方程。对于温度保持，其参考位置(内环xʹ，外环yʹ)设为(0,0)。

### 3.3.1 自适应微调系统的有效性

首先，使用自适应P, D增益的过热蒸汽温度和增益在时间维度变化如图12中，并用一个具有基本增益(固定)的PID控制器与自适应微调系统的PID控制器进行了比较。在图12中又分别分为只对I微调和全部可调。