# 一、前言

强化学习（Reinforcement Learning, RL），又称再励学习、评价学习或增强学习，是机器学习的范式和方法论之一，用于描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。本文主要介绍平均场多智能体强化学习以及其在订单派发上的应用。

强化学习 (Reinforcement Learning) 本身是对这一个原则的数学化描述。强化学习有一个学习者，经常称为智能体 (agent)。智能体需要和周围的环境和整个世界打交道，智能体所做的事情是做出决策 (decision)，来执行什么行动 (action)，在一些时候也成为控制 (control)；而世界环境则接受智能体所做出的决策，并给出相应的后果。我们主要关注的是两种后果：第一种是智能体所能看到或者感知到的它的行动所产生的结果，称为观测 (observation)：譬如一个机器人想要到某个地方去，它能“看”到它经过一整个过程的周围景观；另一种是智能体能直接通过做出某项行动所受到的奖励 (reward)，譬如狗能够得到食物来让它的胃得到满足。

# 二、强化学习简介

强化学习是机器学习的一个分支，是智能体（Agent）以“试错”的方式进行学习，通过与环境进行交互获得的奖赏指导行为，目标是使智能体获得最大的奖赏，强化学习不同于连接主义学习中的监督学习，主要表现在强化信号上，强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价(通常为标量信号)，而不是告诉强化学习系统RLS(reinforcement learning system)如何去产生正确的动作。由于外部环境提供的信息很少，RLS必须靠自身的经历进行学习。通过这种方式，RLS在行动-评价的环境中获得知识，改进行动方案以适应环境。

强化学习有一个学习者，经常称为智能体 (agent)。智能体需要和周围的环境和整个世界打交道，智能体所做的事情是做出决策 (decision)，来执行什么行动 (action)，在一些时候也成为控制 (control)；而世界环境则接受智能体所做出的决策，并给出相应的后果。我们主要关注的是两种后果：第一种是智能体所能看到或者感知到的它的行动所产生的结果，称为观测 (observation)：譬如一个机器人想要到某个地方去，它能“看”到它经过一整个过程的周围景观；另一种是智能体能直接通过做出某项行动所受到的奖励 (reward)，譬如狗能够得到食物来让它的胃得到满足。

## 2.1单智体强化学习

单智体强化学习理论是强化学习的传统研究领域，发展至今又可分为基于值函数的强化学习、基于直接策略搜索的强化学习以及AC算法，其中基于值函数的强化学习是指学习值函数，最终策略根据值函数贪婪得到。基于直接策略搜索的强化学习算法，一般是将策略参数化，学习实现目标的最优参数。AC算法也就是演员评论家算法，就是将二者结合起来。

强化学习通常把环境建模为马尔科夫决策过程，其中每个智能体都可以用一个五元组表示⟨𝑆,𝐴,𝑅,𝑃，γ⟩，其中𝑆表示环境状态的集合；𝐴表示动作集合；𝑅表示奖励函数；𝑃表示状态转移函数，决定了在动作 𝐴 的作用下，从状态 𝑆 转移到 的概率; γ表示折扣率。

其动作值函数更新规则：

## 2.2多智体强化学习

多智能体系统（MAS）故名思义，是存在多个智能体的系统，是承接40-50年代single agent思想面对不同问题与现实环境的一种扩展。

在多智能体系统中，每个智能体通过与环境进行交互获取奖励值（reward）来学习改善自己的策略，从而获得该环境下最优策略的过程就多智能体强化学习。

n个agent可以用一个多元组表示：

其中S表示环境状态的集合； 表示第i个agent的动作集合; 表示第i个agent的奖励；P表示状态转移函数 ，决定了决定了在所有Agent的联合动作 𝐴 的作用下，从状态𝑆转移到的概率。

其动作值函数表示为：

其中，和是由所有Agent的联合动作计算得到

## 2.3多智能体强化学习的问题

在单智能体强化学习中，智能体所在的环境是稳定不变的，但是在多智能体强化学习中，环境是复杂的、动态的，因此给学习过程带来很大的困难。

1.维度爆炸：在单体强化学习中，需要存储状态值函数或动作-状态值函数。在多体强化学习中，状态空间变大，联结动作空间（联结动作是指每个智能体当前动作组合而成的多智能体系统当前时刻的动作，联结动作，指第i个智能体在时刻t选取的动作。）随智能体数量指数增长，因此多智能体系统维度非常大，计算复杂。

2.目标奖励确定困难：多智能体系统中每个智能体的任务可能不同，但是彼此之间又相互耦合影响。奖励设计的优劣直接影响学习到的策略的好坏。

3.不稳定性：在多智能体系统中，多个智能体是同时学习的。当同伴的策略改变时，每个智能体自身的最优策略也可能会变化，这将对算法的收敛性带来影响。

4.探索-利用：探索不光要考虑自身对环境的探索，也要对同伴的策略变化进行探索，可能打破同伴策略的平衡状态。每个智能体的探索都可能对同伴智能体的策略产生影响，这将使算法很难稳定，学习速度慢。

# 三、平均场多智体强化学习

为了解决以上问题，伦敦大学学院汪军教授团队提出了《平均场多智体强化学习》，该方法主要致力于极大规模的多智能体强化学习问题，解决大规模智能体之间的交互及计算困难。由于多智能体强化学习问题不仅有环境交互问题，还有智能体之间的动态影响，因此为了得到最优策略，每个智能体都需要考察其他智能体的动作及状态得到联合动作值函数。由于状态空间跟动作空间随着智能体数量的增多而迅速扩大，这给计算以及探索带来了非常大的困难。

该文章引入了物理学中的平均场理论。平均场论（Mean Field Theory，MFT）是一种研究复杂多体问题的方法。在物理学场论和机器学习的变分推断中，平均场论是对大且复杂的随机模型的一种简化。未简化前的模型通常包含巨大数目的含相互作用的小个体。平均场理论对模型做了近似处理：对某个独立的小个体，所有其他个体对它产生的作用可以用一个平均的量给出，这样，简化后的模型对于每个个体就成了一个单体问题。

按照平均场论的思想，本文对多智能体系统给出了一个近似假设：对某个智能体，其他所有智能体对其产生的作用可以用一个均值替代。这样就就将一个智能体与其邻居智能体之间的相互作用简化为两个智能体之间的相互作用（该智能体与其所有邻居的均值）。这样极大地简化了智能体数量带来的模型空间的增大。应用平均场论后，学习在两个智能体之间是相互促进的：单个智能体的最优策略的学习是基于智能体群体的动态；同时，集体的动态也根据个体的策略进行更新。

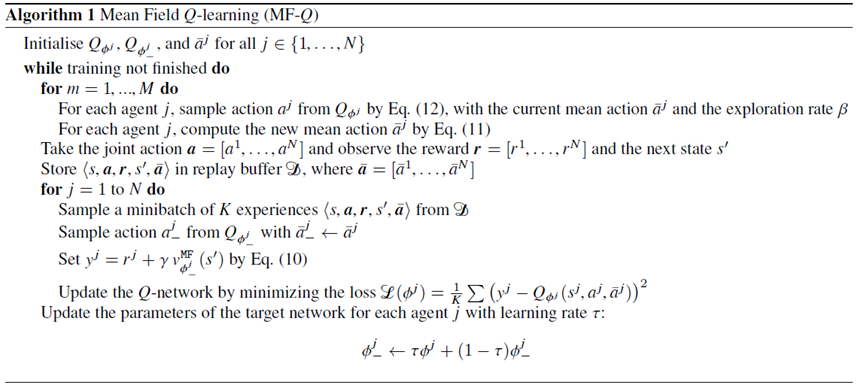
具体而言，每个智能体的动作采用one-hot编码方式， 如智能体j的动作表示共有D个动作的动作空间每个动作的值，若选取动作i，则 ，其余为0。定义为智能体j邻居N(j)的平均动作，其邻居k的one-hot编码动作可以表示为与一个波动的形式

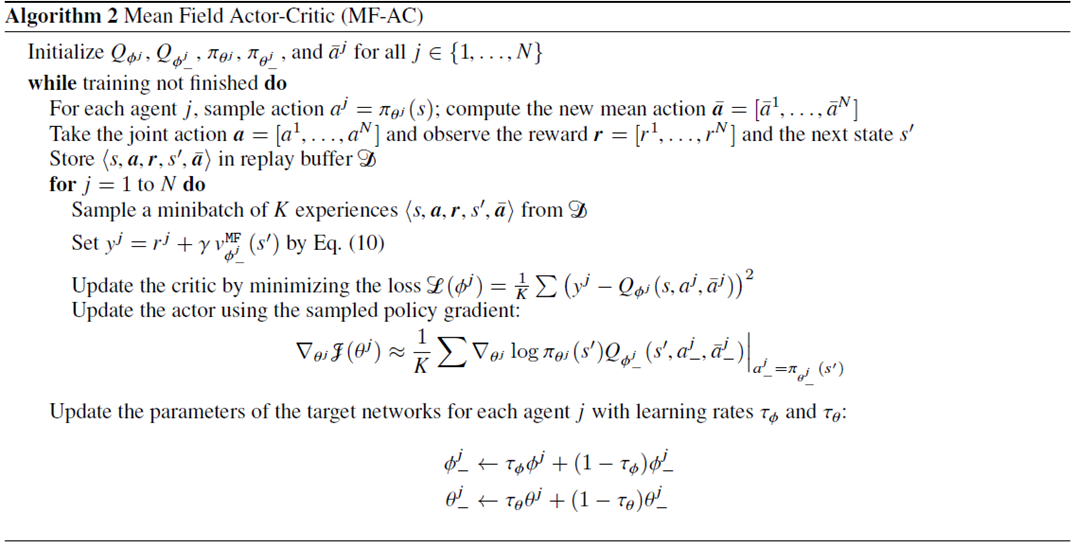
agent 𝑗周围的agent的平均动作由如下公式计算：

agent 𝑗 的Q值函数可以由如下公式计算：

其中均场状态值函数：

基于平均场理论，该文章开发了两种算法：MF-Q算法和MF-AC算法，由此可看出算法主体仍然是强化学习算法（DQN和DDPG），但是在群体的处理上，主要是是群体动作，运用了平均场理论，将核心智体周围智体的动作等效成一个智体的动作。





# 四、使用平均场多智体强化学习的高效共乘订单派发

滴滴出行与伦敦大学学院合作研究的基于平均场多智体强化学习的高效共乘订单派发主要关注于公共出行方面。

## 4.1论文要解决的问题

该文指出，对于大规模实时共乘系统，其关键性问题是如何去派发订单，即如何将订单实时地分配给一组活跃的司机。因为订单派发的质量直接影响交通资源的利用，服务收入的数量，客户满意度等级。而对于订单派发，关键是去寻找短途和长途收益的最佳平衡。因此根据现实需要，订单派发需要同时考虑：（1）派发高价值订单给司机（2）预测预测未来的供需缺口并分散不同目的地的不平衡（3）司机到订单的接载距离最小化

现有的订单派发系统主要有基于规则的方法和免模型的强化学习方法。其中基于规则的方法，就是将手动设置的特征应用到中央调度机构（即组合优化算法）或者分布式多智体调度系统，其中包含一组共享相同环境互相交互的自动智体。但是其本身拥有很多缺陷：1.这两种算法高度依赖特殊设计的加权方案；2.集中调度方法有一个致命的缺陷，即潜在的“单点故障”，即中央机构控制的失败将导致整个系统的失败；3.多智体方案在执行期间需要智体之间多路通信，因此仅限于拥有少量智体的区域。而集中机构控制和免模型的强化学习方法是最近开始兴起的，但仍然有不少缺陷：1.仅通过单智体设置制定订单调度问题，无法对司机和订单之间的复杂交互建模，这样的做法过度简化了大规模共享场景中的随机供需状态；2.集中式执行订单调度系统仍然存在上述可靠性问题（即“单点故障”）。

## 4.2论文使用的方法

如何优化订单调度并避免上述方法的缺陷呢？该文提出了使用平均场多智体强化学习来解决订单调度问题，即将问题建模为多智能体强化学习（MARL），其中智体共享集中式的裁判（critic）为他们的决定（action）打分，以更新他们的策略（policies）。这样的设置有以下好处：1. 当智体能够独立的服从他们学到的策略时，就不再需要集中式的评论家，这样使得订单派发系统对潜在的硬件或连接故障问题具有更好的鲁棒性；2. 通过允许每个司机通过最大化累积奖励来学习，基于强化学习的方法可以不用为匹配算法设计加权值；3. 多智体设置遵循对等共享问题的分布式特性，使调度系统能够捕获大规模共乘场景中的随机需求 - 供应动态；4. 完全分布式的执行还使我们能够通过更多智体扩展到更大的场景，即具有数百万司机的共乘服务的可扩展实时订单调度系统。

然而这样的建模仍然有一定的问题：由于将采取有效订单并且新订单继续到达，因此该集合的大小和内容将不断变化；当智体移动到另一个位置并到达新邻域时，操作集也会发生变化；在真实场景中，司机也可以在线和离线的切换，因此订单派发任务的数量也会因此变化，即动作集和订单派发数量动态变化

那么如何解决上述问题呢？该文主要使用了MFAC算法（基于演员-评论家算法的平均场算法），在分散执行的集中训练框架下解决订单派发问题；其中，评论家从其他智体获取信息来合并对等信息，每个演员仅使用本地信息进行独立操作；为了解决订单派发任务的变量大小问题，我们采用平均场近似来将智体之间的相互作用转换为智体与邻域中的子群体的平均响应之间的成对相互作用；为了解决改变动作集的问题，我们使用每个顺序的矢量化特征作为网络输入来生成一组排序值，这些排序值被馈送到玻尔兹曼softmax选择器以选择动作。

## 4.3模型的设置

在此，该文将订单派发过程建模为部分可观测马尔科夫决策过程（POMDP），通过一个7位元组定义模型：

其中：

N：由标识的N个同构智体定义为环境中的活动司机。（司机可以在在线和离线之间转换，因此N是不断变化）

S，O：每个时间步t，智体i根据观测函数产生与真实环境状态关联的观测值

A：在订单调度任务中，每个司机都只在他的附近选择活跃订单。每个候选动作通过订单的起点和终点的标准化向量表示来参数化表达，即。

P：每个时间步，每个智体采取动作，形成一组联合司机订单对集 ，根据状态转移函数在环境中产生转移

R：每个智能体通过奖励函数计算奖励

要通过考虑每个订单的价值和目的地的潜力来最大化总体收入，就要给奖励值加上限定条件，因此奖励就由以下元素组成。

（1）当前选择订单的全部收益

（2）订单目的地的潜力收益

比较订单起点和终点的供需状态，并鼓励司机选择具有较大供需缺口的订单目的地；

订单目的地潜力（DP）：DP=#DD-#DS，其中#DD：目的地需求，#DS：目的地供应；

只有当起始点订单数量大于司机数量时考虑DP；

如果订单终点司机比订单多，就根据目的地的供需缺口惩罚该订单；

（3）接载距离的收益

最终奖励被定义为：

其中为DP与订单价格的比率，为接载距离与订单价格的比率

：衰减因子，控制着MDP对未来的展望程度，一般取值范围为[0,1]。

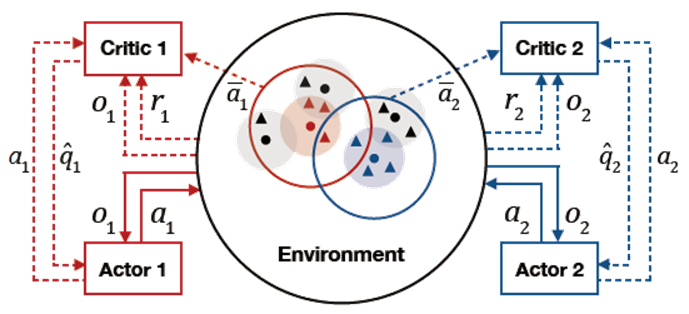
## 4.4具体的算法设计

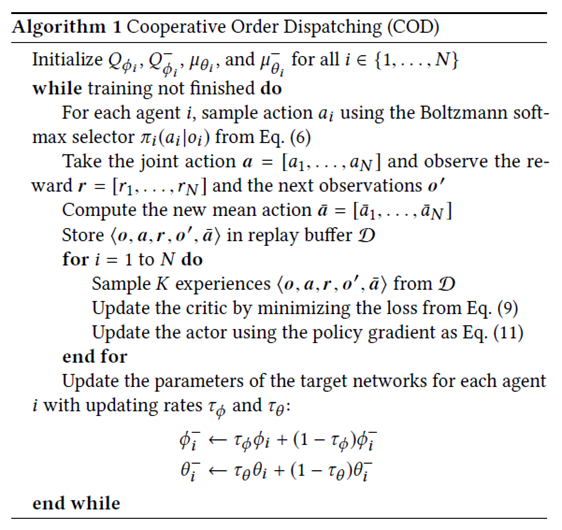
该文设计了两个算法，有一个是Independent Order Dispatching（IOD）算法，核心是单智体AC算法（DDPG），通过扩展单智体到多智体，该论文提出了另一个算法Cooperative Order Dispatching with Mean Field Approximation（COD）算法。

在IOD算法中，AC算法包含四个神经网络，分为critic和actor。其中，Critic网络是DQN网络，包含Q网络和目标网络，其将观测值作为输入，观测值包含智体i的位置和时间步t。通过通过最小化loss函数使用神经网络函数逼近学习动作值函数。

使用确定性动作生成每个智体i的动作集的每一个Mi候选的观测动作对的排序值，然后输入玻尔兹曼选择器来选择最佳动作。

其COD算法示意图及伪代码如下：





其中Q网络loss函数：

均场值函数：

Actor策略梯度：

在COD算法中，在订单调度任务中，智体通过选择需求较高的订单目的地来相互交互，以优化供需缺口；智能体的交互半径为其接收订单的2倍，这样智能体能与邻域智能体的动作集交互同时也能与相邻智能体交互；平均场近似使得智体可以用相互作用方的意识学习，因此有助于训练之后提高智体的训练稳定性和鲁棒性；平均响应只作用在模型更新阶段，因此中心调度只在训练时需要；执行阶段，智体完全以分布式执行，以此避免了单点失败。

该文最后通过实验证明了所提算法的有效性。