# 三、强化学习自博弈训练

### 3.1 强化学习概述

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。换句话说，强化学习是一种学习如何从状态映射到行为以使得获取的奖励最大的学习机制。这样的一个智能体需要不断地在环境中进行实验，通过环境给予的反馈即奖励来不断优化状态-行为的对应关系。因此，反复实验和延迟奖励是强化学习最重要的两个特征。

强化学习在使用过程中面对复杂任务时存在很多难题：

第一是稀疏奖励，稀疏奖励是强化学习中的一个常见问题，指智能体在执行一系列动作以完成任务时，只在完成整个任务或达到特定阶段后才获得奖励信号。这种情况下，智能体很难从环境中获取有效的反馈，导致学习过程缓慢或者失败。

第二是在实际训练中，我们很难为了每一个特定的复杂技能去搭建复杂的环境，提供不同水平的针对性训练，这样对人工工作量和计算机的算力都提出了过高的要求。

第三是强化学习训练目标本身与竞争有关，如围棋等游戏，需要双方博弈以得到能力判断，获取奖励，无法通过简单的单个智能体表现给予奖励。

### 3.2 强化学习训练方法

针对以上难题，目前有几个较为成熟的训练策略。

奖励塑形：通过人为设计额外的奖励信号，引导智能体逐步接近目标。例如，在机械臂任务中，可以根据机械臂与目标物体的距离给予不同的奖励。

经验回放：将智能体与环境交互的经验存储在经验池中，进行离线训练时从经验池中采样，增加稀疏奖励样本的利用率。

分层强化学习：将任务分解为多个子任务，每个子任务由不同的策略来完成。分层主要解决的是稀疏reward的问题，我们人类在解决一个复杂问题时，往往会将其分解为若干个容易解决的子问题，分而治之，分层的思想正是来源于此。将复杂任务分解为多个子目标，每个子目标都有相对密集的奖励信号；在子目标之间引入内部奖励机制，使得智能体在完成子目标时能够获得奖励，从而加速学习过程。

自博弈训练：通过循环迭代的多次自我博弈，智能体自己逐渐意识到哪些动作能使自己在博弈中取胜的概率最大化，并最终掌握最终的复杂技能。

### 3.3 自博弈训练概述

在实际训练中，很难为了每一个特定的复杂技能去搭建复杂的环境，提供不同水平的针对性训练，这样对人工工作量和计算机的算力都提出了过高的要求，基于这个问题，提出了自博弈(Self-Play)的强化学习训练框架。将这种自我博弈以获得进化的思想应用于现实的对抗性运动中。

自我博弈（Self-Play）是应用于智能体与自己之间处于对抗关系的训练方法，这里的对抗关系指的是一方的奖励上升必然导致另一方的奖励下降。通过轮流训练双方的智能体就能使得双方的策略模型的性能得到显著提升，使得整个对抗系统达到正和博弈的效果。这种训练机制最早运用于跳棋游戏，后来在棋类，MOBA类等对抗性策略游戏中被广泛应用，例如AlphaGo、AlphaStar。

### 3.4 自我博弈主要创新点

第一是探索课程，在稀疏奖励的环境下，将最终目标分割成若干个容易实现的小目标，每个目标都基于一定的奖励，使奖励变得稠密，并对智能体的探索行为给予一定的奖励。在训练的前期，智能体就可以迅速掌握基础动作，这些在自我对抗中常见的基础动作对未来智能体实现最终目标，获得终端奖励有很大帮助。

但随着训练的进行，探索奖励应当逐渐衰减为0。因为由于最终目标相对复杂，因此在强化学习训练过程中实现该动作概率很低，一旦智能体发现只要做出基础动作就能不断获得探索奖励，它可能就不会去尝试更为复杂的目标动作，反而对训练产生了阻碍。因此为了解决探索奖励与竞争奖励的平衡问题，随着自我对抗训练过程的进行，探索奖励会逐渐衰减至0，利于智能体最终顺利完成目标获得稀疏竞争奖励。将整个训练过程前10~15%的时间定为探索阶段是最有效的训练模式。这样的设计既确保智能体一步步循序渐进掌握复杂技巧，又保证了学习过程的收敛速度，同时避免了智能体一味地贪图阶段性获得的探索奖励，只满足于基础动作的掌握，最终还是无法完成目标。

奖励具体设置表达式为：

其中为线性退火因子，令第t个时间单位下的探索奖励、竞争奖励、探索终止时间分别为St，R，T。该公式更加清楚展现了探索课程的训练过程，首先用探索奖励刺激智能体学习基础行为，随着训练程度的加深，智能体逐渐掌握基础动作后，探索奖励衰退为0，所有奖励由取得博弈胜利后的竞争奖励提供，确保了最终学习目标的顺利完成。

第二是对手抽样，智能体在面对过强的对手时，一直无法取胜获得正向的反馈激励，我们就一直无法进步，如果太弱，奖励始终很高，训练都会难以取得成效，因此训练过程中遇到过于强大的对手时，系统会在该对手的早期版本中抽样最为博弈对象，这样让弱者有获胜的机会，在博弈中取得奖励和进步，一旦弱者的水平提升到一定程度，也能反过来训练激励强者使其变得更强。

第三是鲁棒学习策略，在强化学习与深度学习中，常出现的一个问题是过拟合，为了防止智能体的策略拟合在某一种固定的对手身上，需要给对手引入多样性，给环境引入随机变量。

第一种是世界随机性。为了提高学习策略的鲁棒性，我们可以对环境引入随机效果。智能体的初始位置都可以是不固定的，但这样的随机性给早期训练带来了极大的探索可能性，导致智能体的训练时间过长，收敛速度过慢，因此为了让智能体更好更快地学会普适性博弈技巧，设计了一个可变的随机性算子，其以较小的值开始训练，以便更快掌握基础动作，随后该值随着训练过程的推进逐渐变大，最终通过实验证明了这样的可变随机性给学习带来了鲁棒性上的大幅提升。

其次是对抗整合策略。让智能体同时学习多个策略，引入一个策略池，在每次推出特定策略时，会随机选择其他策略中的一个作为对手(在对称博弈中，同样的策略也可能是对手)，这样的方法类似于多任务学习，在多任务学习中使用同一个网络来建模多个相关的任务，这样可以在任务之间共享统计强度，减少过拟合。这样就同时保证了训练过程的集中性与多样性，我们也可以从作者的实验数据中看到策略的有效性：

### 3.5 自博弈训练应用于空战问题

#### 3.5.1 智能体强弱判断

为了判断智能体的强弱，使自我博弈使用水平相近的智能体进行对抗训练，根据ML-Agents中的做法，较好的解决方式是引入ELO评级系统，这是一种计算零和游戏中两个玩家之间相对水平的方法。每一个玩家都有一个初始ELO分数（通常为1200），通过对战胜负结果更新ELO分数，假定玩家A的初始ELO分数为，玩家A的初始ELO分数为，则玩家A对于B的预期胜利概率为：

比赛结束后，玩家A的分数调整为：

其中为A的预估最大胜利概率。

#### 3.5.2 智能体分阶段训练

控制敌我飞机的智能体身为对抗双方，可以在对手的早期模板中选择自己合适训练环境。也就是说，对方的智能体在己方训练的时候相当于环境的一部分，对手的策略也极大影响了自身的策略。

因此，双方交替训练的时机以及选择对手的策略是影响自我博弈结果的重要因素，当双方训练进度不一致时，会造成双方的竞技水平不平衡，不利于双方策略的学习。实验中智能体每隔一段时间会保存自己的副本，在训练的时候会从敌方的副本中抽取一个当做训练的对手，抽取的策略有很多种，例如每个样本概率平均、最近的样本概率最大并线性变化、多智能体同时协作训练等。如下图中所示，有很多训练思路：

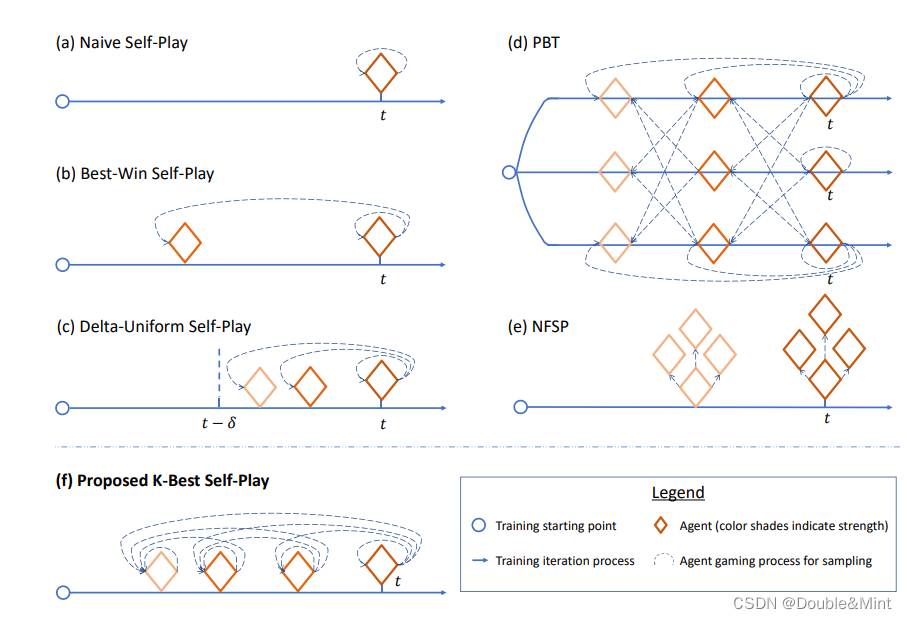


图1 自博弈训练思路图

#### 3.5.3 终端奖励计算

以单局中自博弈的输赢作为终端奖励计算依据，考虑因素包括：

胜负奖励：如果智能体在战斗中获胜，可以给予较高的奖励。例如，击败对手或完成主要任务目标时给予高额奖励。如果智能体在战斗中失败，可以给予负奖励，以鼓励智能体避免失败的策略。

任务完成度奖励：根据智能体完成任务的程度给予不同的奖励。例如，摧毁敌方部分目标或达到某个阶段目标时给予相应的奖励。

资源消耗奖励：根据智能体在战斗中使用资源的效率给予奖励。例如，使用较少的弹药或燃料完成任务可以获得额外奖励。

时间奖励：根据智能体完成任务所用的时间给予奖励。例如，在规定时间内完成任务可以获得额外奖励，超过时间限制则减少奖励。