# 重要概念

## QN基本概念

在强化学习中，四个最基本的元素是

s : 当前状态state

a : 当前状态下能够采取的行动action

r : 在当前状态s下采取行动a之后所获得的即时奖励reward，可以是正激励，也可以是负激励。

: 当前状态s下采取行动a之后达到的下一状态

在DQN中还有另外一个重要概念：经验q

: 在状态s下执行a行动所获得的即时奖励

: 在状态s下执行a行动所获得的经验分数

奖励矩阵是固定的，经验矩阵则是根据的值一步步计算出来的：

* 初始时都是0，通过上式不断迭代最终计算出一个不变的经验矩阵
* 是学习率，是折损率，都是人工设置的
* 代表下一状态采取的所有行动中经验最高的

当学习率时，

在计算出经验矩阵之后，对于任意状态都可以根据选择最优的行动。

## DQN算法设计

准备工作：奖励矩阵，提前设计好每种状态下采取某种行动后得到的奖励分数的计算方式。

模型输入：当前状态s

模型输出：当且状态s下能采取的所有行动，对应的所有经验分数，记为

模型损失函数：

* Q是eval\_net模型的输出
* target\_Q类似于数据的标签，

target\_Q的计算方式：

即

其中可以通过将输入target\_net模型得到

训练数据来源：

构造一个简单的智能体去探索环境。即随机初始化状态，再随机选取行动，不断的循环执行，并记下每一步的，这样就构造了一个记忆库。每次从记忆库中随机抽取一部分

## 算法流程

1. 初始化参数，随机设置一个状态s
2. 循环遍历以下操作：
3. 根据策略选择一个行动a
4. 执行该行动a，获得奖励r，达到下一状态，并判断游戏是否结束
5. 将存入记忆池，记忆池满了之后，就用新记忆替换掉旧记忆，继续往里存。
6. 记忆池数量或者行动步数达到一定条件，进行一次学习。
7. 判断游戏是否结束：
8. 结束：随机设置一个状态s，重新循环
9. 未结束：用更新s

## 模型学习流程

假设batch\_size=32，action\_num=4(可选择的行动个数)

1. 从记忆池中随机抽取batch\_size组数据()，将训练数据有序的拆分成S, A,R,，分别对应着所有的。
2. 将S输入eval\_net计算出，维度为(32,4)

将输入target\_net计算出，维度为(32,4)

1. 计算
   1. 对中的每个元素进行更新，只更新这次状态s下所选择行动a的q值：
2. 计算损失函数
3. 更新eval\_net参数
   1. eval\_net学习一定次数之后，更新target\_net参数。即将eval\_net模型参数复制给target\_net模型。

说明：损失函数的计算其实是，而且这组数据action对应的q值是有意义的，其他3个q值都应该是0，这样更新模型才有作用。所以我们要先令，然后只更新中每一元素对应的action那一项。