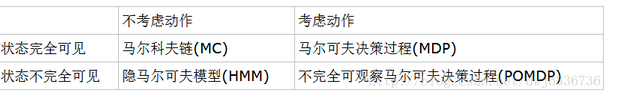
Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless

关联信道符合未知马尔可夫模型，用户选择信道传输信号，使用DQN方法进行动作选择。



假设有N个信道，每个信道有两个可能状态good（1），bad（0）联合分布服从2N的马尔可夫模型，用户（无线网络节点）在每个时隙（time slot）可以选择一个信道进行数据包传送，信道good得到奖赏+1，bad得到奖赏-1，目标是得到最大累计奖赏。每次只有选取信道才知道信道状态，所以是一个状态不能完全观测的情况（POMDP），是一个PSPACE-hard问题，最优解需要大量计算复杂度，且马尔可夫模型事前未知。

现有传统解决方法有两个：

贪心算法（Myopic policy）：在信道间独立同分布情况下，被证明是全局最优解，但在相关信道情况下或信道分布不确定情况下效果不能保证。

信道独立时，问题可以被视为Restless Multi-armed bandit problem（(RMAB），

Whittle Index policy：第二种算法，用每个信道的价值索引（index）进行算法降维。对于半马尔可夫问题，解于贪心算法差不多。且对于不全同分布信道，方法依旧有效，但是对于相关信道不太行。

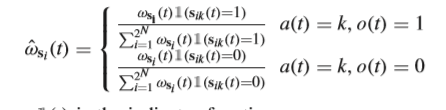
问题建模：

对于动态多信道接入问题，首先要创建一个相关的接入信道系统。马尔科夫链状态空间为：



其中每个Si是一个长度为N的状态向量，代表信道好坏（1/0），马尔可夫转移记作P。用户每次在一个时隙中只能得到单个信道状态，整体状态不能得知，用户通过观测状态和动作对系统分布进行推测。

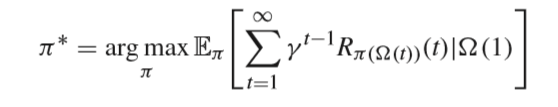
代表belife victor，代表系统在si状态概率，依据之前观测和决策。给出动作决策向量a（t），代表t时刻选取哪个信道感知和传送，得到观测值o（t）：（0/1），对状态概率进行更新：



在得到每个时隙观测更新后状态概率矩阵，乘上状态转移矩阵，就可以得到下一时隙的状态概率矩阵。

。

然后生成一个策略Π，去依据状态矩阵获取动作a（t），以此获得长期累积最大奖赏。



Gamma为未来折扣系数。

解决算法：

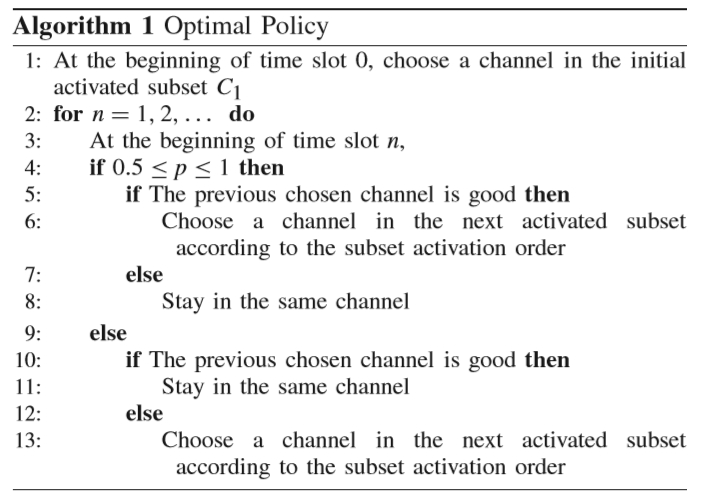
1. 贪心算法：

轮询方式，选取当前最优。

1. Whittle Index Based Heuristic Policy

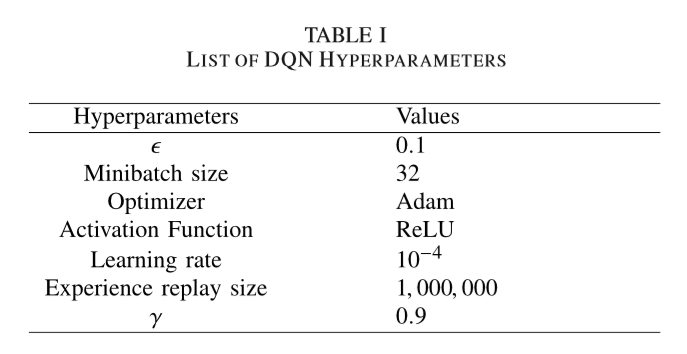
因为信道间存在相关性，index无法定义，无法直接使用，进行改进，忽略信道间关联，依据贝叶斯规则和概率转移矩阵P推测单独信道模型，然后计算index。

对于信道转换模式固定的情况：（方便得到DQN效果对）：

首先要建立一个相关信道系统，信道转移方式：所有N个信道分成几个独立的子集，这些子集依据固定方式轮流转变为活跃模式。假设所有在活跃子集中的信道都是好的。在每个时隙（time slot），在已知概率P下，下一个集合是活跃的，1-p概率原子集保持活跃。假设激活顺序已知，固定。算法1可以提供一个基础对比线。

然后是构建DQN方法进行问题求解：

全连接，2层，每层200隐藏神经元



输入为前M=N次时隙的动作和观测值，输出为选取信道Q值向量。

代码阅读心得：

智能体选取动作不会影响环境，环境按照之前规定的相关生成形式以csv形式储存。交互时间为查询时间，较快。

输入设置为*(a\_t-1, o\_t-1 ,......, a\_t-M,o\_t-M)* 多对过去选择动作与观测对。

输出正常，为信道选择的one-hot形式。

经验回放正常，记忆体队列形式，正常队列为记忆体。

同样加入伊普西隆动作探索策略。

训练时数据从记忆体中随机抽取，组织。

不使用双网络，没有Q—target。