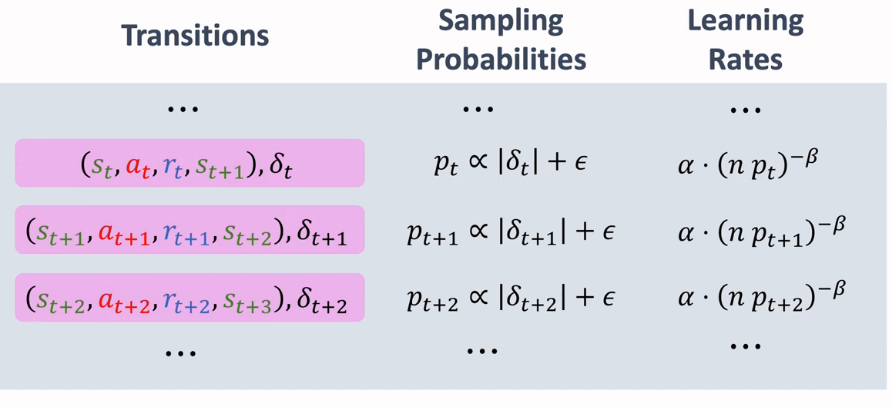
### 优先经验回放（Prioritized Experience Replay）

在经验回放的基础上，优先经验回放通过为经验分配优先级来决定其被采样的概率。通常使用TD误差值计算优先级，即预测值与真实值之间的差值，TD误差越大则优先级越高。这样可以使模型更多地关注那些对改进模型有贡献的重要经验，提高学习效率和表现。（与经验回放的区别在于用非均匀抽样代替均匀抽样）

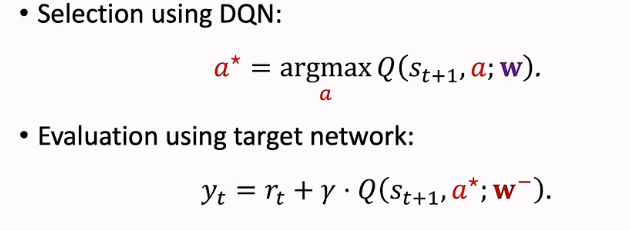
#### 1.1抽样方式

pt∝|t|+（让抽样概率pt正比于t的绝对值，t为TD error，为一个很小的数，防止概率pt等于0）原理即为TD error越大，transition被抽到的概率就越大。

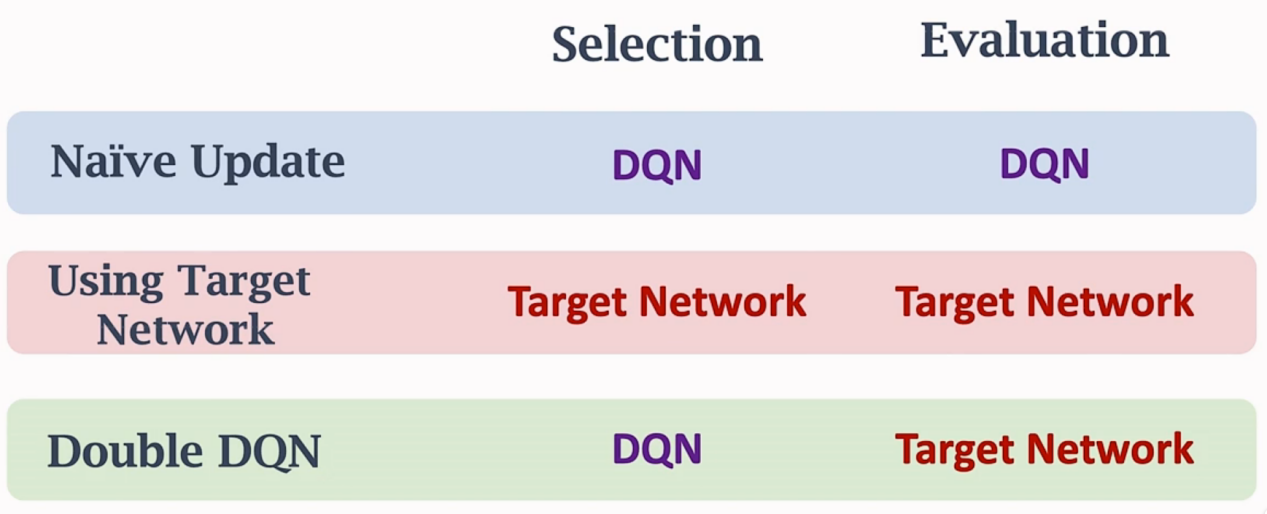
由于抽样是非均匀的，不同的transition有不同的概率，会导致DQN的预测产生偏差，需相应调整学习率，抵消掉不同概率造成的偏差。如果一条transition有较大的抽样概率，那应该减少它的学习率。通过计算(n\*pt)-β，将其乘到学习率上，对于大的pt会使(n\*pt)-β更小，β（0，1）为超参数。如果一条transition刚被抽到还未用来训练DQN，直接将t设置为最大值。

### **Double DQN**

Double DQN（Deep Q-Network）是一种用于解决增强学习中Q-Learning存在的过估计问题（overestimation），从而提高学习效果的方法。该算法主要通过利用目标网络和当前网络的不同参数，来减小Q值函数过分估计值（over-estimation）的问题。



Double DQN把计算TD target分成两步，第一步选择最优动作时用DQN本身，第二步用target network计算TD target，该方法相比于原始的DQN以及target network拥有更好的效果。下图是三者的对比。



基于DDQN及优先经验回放的模块化机械臂研究是应用深度强化学习技术实现机械臂动作规划、控制和优化的一种新方法。该方法采用DDQN算法和优先经验回放技术，将机械臂视为智能体，通过与环境交互获得经验，训练神经网络，从而学习出高效的机械臂动作策略。

在这个方法中，机械臂由多个模块组成，每个模块控制机械臂的一个关节或者连杆。通过设计合适的状态表示、动作表示和奖励函数，可以让机械臂完成特定任务，如拾取物体、平移、旋转等。

具体地，DDQN算法可以减小Q值过高估计的风险，优先经验回放可以提高采样效率和数据多样性，使得机械臂在与环境交互时更加高效和稳定。通过这种方法，我们可以为机械臂提供更好的自主决策和执行能力，使其可以适应复杂多变的生产环境和任务要求。

总之，基于DDQN及优先经验回放的模块化机械臂研究是一个有前途的研究方向，可以为智能制造、物流等领域的自动化生产提供更加高效和可靠的解决方案

在强化学习中，一个Epoch（时期）指的是一组连续的交互和学习周期，在该周期内智能体与环境互动并尝试学会执行任务。一个Epoch通常包含多个回合（Episodes），每个回合表示从初始状态开始到最终状态结束的一个完整过程。

在每个回合中，智能体会根据当前状态选择一个动作，然后将该动作应用于环境中，环境会反馈给智能体一个奖励信号，同时智能体会从环境中观察到新的状态。智能体通过不断地进行这些交互和观察来提高自己的策略和价值函数，以便更好地完成任务。

在一个Epoch完成之后，智能体可以对当前的策略进行评估和更新，并且重新开始下一个Epoch，以进一步提高智能体的性能。Epoch的长度通常由具体的强化学习算法和任务所决定，是可调参数之一。

在强化学习中，Episode（回合）指的是从初始状态开始到达目标状态或结束的一次完整互动过程。每个Episode由多个时间步组成，在每个时间步中智能体会观察到当前的环境状态，然后选择一个动作来影响环境，并且从环境中获得奖励和下一个状态。

在一些任务中，每个Episode都对应着相同的起始状态，因此可以进行多个Episode的批量学习，其中每个Episode都用于独立地收集大量的数据并更新策略和价值函数以最小化总的累积奖励。这种批量学习的方法在深度强化学习中得到了广泛的应用，并且已经被证明是可行的有效的。

在其他任务中，初始状态和目标状态是随机生成的或者根据先前的状态生成的，在这种情况下，每个Episode对应着不同的互动轨迹，智能体需要通过探索来学习最佳策略来实现特定的目标。通过多次的Episode互动，智能体可以逐渐优化自己的策略和价值函数，进而使得长期的累积奖励最大化。

在强化学习中，Transition（转换）指的是智能体从当前状态采取一个动作后，进入到下一个状态的过程。每个转换都包含四个元素，即当前状态（state）、采取的动作（action）、环境给出的奖励信号（reward）、以及产生的新的状态（next state）。这些元素可以表示为(s, a, r, s')的形式。

智能体在交互中不断进行转换操作来学习最优策略和价值函数。通过观察不同的转换数据，智能体可以利用各种强化学习算法，例如Q-learning、SARSA等，来更新自己的策略和价值函数，并且希望最大化累积奖励的总量来实现任务目标。

转换是强化学习中重要的概念之一，也是一组数据的基本单位。通过特定的采样方法，可以从各种强制学习问题中收集大量的转换数据，在训练周期内使用它们来逐渐提高智能体的性能。

在强化学习中，学习率（learning rate）是指在每一次模型参数更新时所使用的步长或者衰减系数。它控制着每次参数更新的数量以及在学习过程中对新数据和旧数据进行平衡的权重。

学习率是一个超参数，需要根据特定问题和算法去选择适合的值。如果学习率太高，可能会导致参数跳过最优解并最终发散；如果学习率太低，则可能会让学习过程变得过于缓慢，甚至无法收敛到最佳策略。

通常情况下，学习率随着训练的进行而逐渐降低，以使得智能体在解决任务时更加稳定和可靠。同时，在一些复杂的问题中，学习率还可以通过自适应的方式来调整，如Adam、Adagrad、RMSProp等自适应优化器的方法，以更好地平衡新旧数据的权重，并且在搜索空间中寻找更好的局部最优解