> 以下公式都仅针对于**马尔科夫决策过程**

状态价值与行为价值的关系

状态序列

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...$$

状态价值

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t \mid S_t = s] && (1) \\ &= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] && (2) \end{aligned}$$

如果用矩阵的形式来描述这个式子，那么就有：

$$v = R + \gamma P v$$

其中P就代表了对于后续状态价值求期望

进行矩阵运算即可得到：

$$v = (I - \gamma P)^{-1} R$$

行为价值

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}[G_t \mid S_t = s, A_t = a] && (3) \\ &= \mathbb{E}[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a] && (4) \end{aligned}$$

状态价值与行为价值

一个状态的状态价值可以用该状态下所有行为价值来表达：

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$$

一个行为的行为价值可以用该行为能达到的后续状态价值来表达：

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s^{'})$$

那么，一个状态的状态价值可以表示为：

$$v_\pi(s) = \sum_{a \in A} (R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s^{'}))$$

一个行为的行为价值可以表示为：

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s^{'} \in S} P_{ss^{'}}^a (\sum_{a \in A} \pi(a^{'}|s^{'}) q_\pi(s^{'}, a^{'}))$$

具体的描述一下第三个公式，首先它需要几个条件:

1. 这个状态的状态值基于策略π
2. 该策略下基于状态s有一到多个行为a，采取这些行为后到达的状态$s^{'}$是具有概率性的，也就是说采取行为后到达的状态是不确定的。

那么，最后这个公式可以描述为:
一个状态的状态价值基于给定的策略π，它表示该策略下所有可能的动作获得的即时奖励，加上采取该动作后到达下一步状态的概率乘以下一步状态的状态价值乘以折现率。
再换句话说，一个状态的状态价值等于它当前所有可能行为的即时奖励，加上采取某一行为后后续状态价值的期望。

> 总结:
> 一个状态的状态价值，等于该状态采取某个行为获得的即时奖励加上以该状态为起点的后续所有可能的状态的价值的期望。
> 一个行为的行为价值，同样可以看作是行为获得的即时奖励，加上该行为到达某状态的概率乘以该状态的状态价值。

最优状态价值函数与最优行为价值函数

最优状态价值函数

最优状态价值函数指的是在状态s下，所有的策略中产生的状态价值中最大者:

$$v_* = \max_\pi v_\pi(s)$$

换句话说，如果某个策略π使得对于任意状态s，都有$v_\pi(s) >=$ 其他任意策略下的状态价值，那么就称这个策略是最优策略，使用了最优策略的状态价值函数称为最优状态价值函数。

最优行为价值函数

最优行为价值函数指的是在状态s下，所有策略产生的行为中行为价值最大者。

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

> 再提一次，行为价值依赖于状态s下采取的行为获得的即时奖励，以及该行为可能达到的所有后续状态的概率(由状态转移概率矩阵确定)乘以该后续状态的状态价值。

从公式中可以看出，最优行为价值函数也和策略的选择息息相关。最优行为价值来自于最优的策略，最优策略使得在任何一个状态s下选择某个行动a获得的行为价值最大。
同时，最优策略可以通过最大化行为价值函数来获得。也就是说，假如我们处于任意状态s，此时采取行动a获得的行为价值$q(s,a)$均是最大的。那么我们的最优策略π就应该在这个状态下总是选择这个行为，即$\pi(s,a)=1.$
解决强化学习问题意味着需要寻找一个最优策略，使得该策略下的return最大。而我们最优策略可以通过最大化行为价值函数来获得，因此解决强化学习问题就变成了求解最优行为价值函数的问题。

最优行为价值函数可以换一种表示形式，表示为：

$$q_*(s,a)=\mathbb{E}[R_{t+1}+\gamma v_*(S_{t+1})\mid S_t=s,A_t=a]$$

也就是说，最优行为价值可以用该行为获得的即时奖励加上后续达到的状态价值的期望得到。

动态规划寻找最优策略

预测是对给定策略的评估过程，控制是寻找一个最优策略的过程。

预测($prediction$**)**:已知一个马尔科夫决策过程$<S,A,P,R,\gamma>$和一个策略π，或者是给定一个马尔科夫奖励过程$<S,P_\pi,R_\pi,\gamma>$，求解基于该策略的价值函数v_π
控制($control$**)**:已知一个马尔科夫决策过程$<S,A,P,R,\gamma>$，求解最优价值函数v_*和最优策略π_*。

策略评估

策略评估指的是计算给定策略下状态价值函数的过程。
使用上一个迭代周期k内的后续状态价值来计算更新当前迭代周期$k+1$内某状态s的价值(听起来像雅可比迭代法):

$$v_{k+1}(s)=\sum_{a\in A}\pi(a|s)(R_s^a+\gamma\sum_{s'\in S}P_{ss'}^a v_k(s'))$$

策略迭代

策略迭代指的是完成一定轮次的策略评估之后，此时我们得到了所有的状态的状态价值。那么我们就可以通过新的状态价值来更新我们的策略，使得我们的策略在某个状态时更倾向于前往状态价值更大的状态。然后，我们使用更新后的策略进行策略评估，更新我们的状态价值函数，重复这个过程。最终，它们都会收敛到最优价值函数和最优策略。

价值迭代

一个策略，如果是最优策略，那么它在某个状态下一定能够产生当前状态下的最优行为。而且通过当前状态下的最优行为到达的后续状态。该策略同样是最优的。
或者换句话说，如果一个策略对于某个状态产生的行为它不是最优行为，那么这个策略就不是一个最佳策略。

> 一个策略能够获得某状态的最优价值当且仅当该策略也同时获得状态所有可能的后续状态的最优价值。
> 对于这句话的理解就是，一个策略获得某状态的最优价值。说明该策略在该状态下产生的行为是一个最优价值行为，最优价值行为就是由当前行为的奖励加上行为导致的后续状态价值的期望组合而成。
> 也就是有:
>
> $$v_*(s) = \max_{a \in A}(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s^{'}))$$

价值迭代与策略迭代的区别

价值迭代算法中迭代的是状态价值函数 $v(s)$，以逼近最优状态价值函数 $v^*(s)$。

> 与策略迭代不同，价值迭代算法**并不需要显式地维护一个策略**。在每次更新状态价值函数后，我们都会自动进行一次最优化策略改进，即选择当前能够使下一个状态价值最大化的行动作为最优策略的行动。因此，价值迭代算法通常会比策略迭代更加高效，因为它将策略评估和策略改进两个过程集成到了一个算法中，无需反复地进行迭代。
> 另外，与策略迭代相比，价值迭代算法通常更容易实现和理解，并且可以处理具有连续状态和行动空间的问题。但它也有一些缺点，例如可能会收敛得比较慢，特别是在状态空间很大的情况下。此外，如果状态或者行动空间非常大，那么存储和计算状态价值函数的成本也可能会很高。
>
> 书上描述的二者的差距：策略迭代每次迭代仅计算相关状态的价值，一次计算即得到最优状态价值。后者在每次迭代时要更新所有状态的价值。
> 这里的"一次计算即得到最优状态价值"指的是基于当前的所有策略中选择了最优的策略计算出的最优状态价值，但是策略迭代同时也需要改善策略。所以根据最优的状态价值还会更新策略继续下一次的策略迭代。

> 而价值迭代是基于各个状态各自能够采取的行动，采取最优行为来更新自己的状态价值函数。也就是说，每次迭代完毕后，我们根据各个状态状态价值的大小选择出来的路径就是最优策略。

> chatGPT给出的描述:
> 在策略迭代中，首先进行一定次数的策略评估得到当前策略下状态的价值函数，然后通过策略改进更新策略，再重新进行策略评估得到新的状态价值函数。这个过程不断重复，直到算法收敛为止。策略迭代的优点是它能够针对特定的策略进行优化，逐步提高策略的质量。
> 而在价值迭代中，每轮迭代会更新所有状态的价值函数，不需要显式地保留策略。价值迭代通常需要多次迭代才能达到最优解，但是一旦完成，从任意状态出发直接选择状态价值最大的状态就可以得到最佳策略的轨迹。因此，价值迭代较为适用于解决具有有限动作空间、可离散化状态空间的强化学习问题。
> 总之，策略迭代和价值迭代都是常用的增量式策略优化方法，在具体应用时需要根据问题的特性选择合适的算法。

价值迭代过程中，价值函数更新的公式为:

$$v_{k+1}(s) = \max_{a \in A}(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s'))$$

这个公式和策略评估的公式非常相似，区别在于策略评估需要根据策略考虑当前状态下所有行为的概率，而价值迭代中仅考虑行为价值最大的行为。
以上就是同步动态规划的内容，迭代法策略评估属于预测问题，使用贝尔曼期望方程进行求解。而策略迭代和价值迭代属于控制问题。