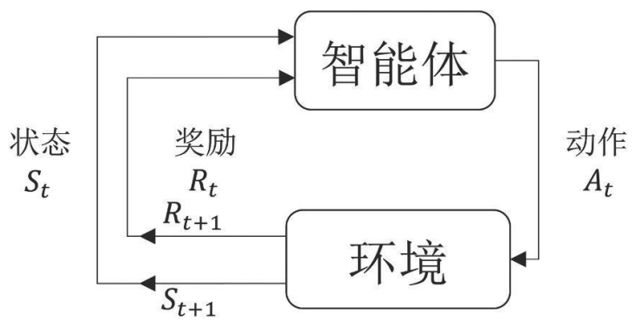
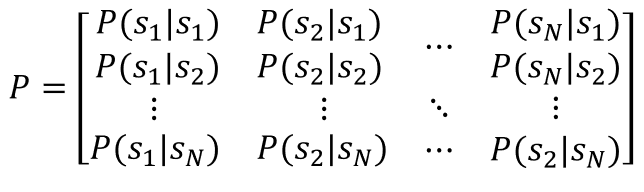
有模型预测与控制：

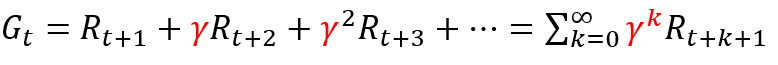
马尔可夫决策过程 (Markov decision process, MDP) 描述了强化学习的一般环境规律。

马尔可夫性质：给定当前状态𝑆\_𝑡，𝑆\_(𝑡+1)取决于𝑆\_𝑡，不用再考虑历史状态。

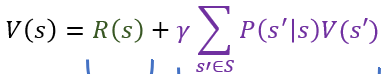
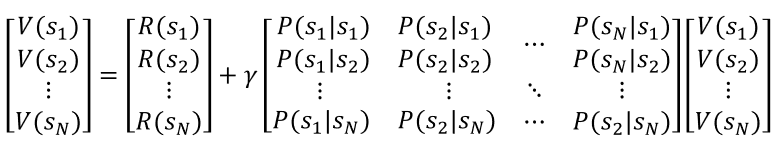
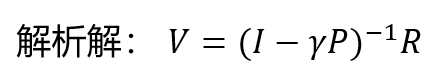
状态转移矩阵 (state transition matrix)

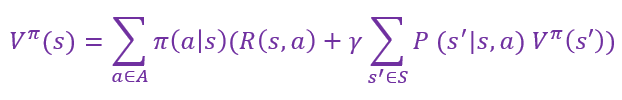
马尔可夫过程 (Markov Process)是满足马尔可夫性质的离散随机过程，又称马尔可夫链 (Markov Chain)。

回报 (return)：从时间𝑡开始，获得的总的折扣奖赏。

折扣因子𝛾：表征未来的奖赏在当前时刻的指数折现

状态值函数 (state value function)：从状态𝑠开始，获得的回报的期望

状态值函数的求解：贝尔曼方程：

 引入策略𝜋（动作a）后：

预测 (prediction)：通过计算价值函数评估给定的策略

控制 (control) ：寻找最优的策略。