1.强化学习的优雅结构

强化学习（reinforcement Learning）范式与人类的学习方式类似，创建一个智能体通过试错机制与环境互动获得最大化的收益。如图 1 所示，交互过程由马尔可夫决策过程（Markov Decision Process，MDP）定义，在一个时间步 t 中，智能体根据状态S_t，来选择与环境交互的动作A_t，环境接受动作的改变状态发生转换，返回下一时刻的奖励r_{t+1}和状态S_{t+1}。奖励对智能体的策略进行引导和修正，智能体综合短期收益和长期目标制定最优策略。如此循环可以得到一个周期T的样本。（注：本节中描述 MDP 的大写字母为随机变量，对应的小写字母为随机变量的样本）

$$\{S_0, A_0, r_1, S_1, A_1, r_2, \dots\dots S_t, A_t, r_{t+1}, S_{t+1}, \dots\dots S_{T-1}, A_{T-1}, r_T, S_T\} \quad (1)$$

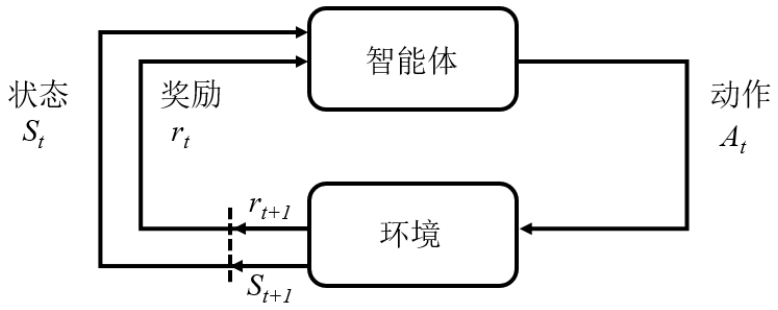

图 1 马尔可夫决策过程

马尔可夫性质是“试错机制”的核心，其保证了智能有效从历史经验中学习到有用的规则，也就是智能体策略表现提升的最根本原因。具体来说，在状态产生转移时$p(S_{t+1}|h_t)$，这个h_t越长对下一时刻的预测就越精准，但是考虑无限长的时间序列是不现实的做法，所以假设这个时间序列满足马尔可夫性质，$p(S_{t+1}|h_t) = p(S_{t+1}|S_t, A_t)$，在决策中这一假设保证了最关键的当前时间步和下一时间步之间的联系，也回答了考虑多长的序列作为历史经验最为合理的问题。

基于马尔可夫性质可以定义出 MDP，这一过程中，基于周期经验序列对奖励的定义显式引导了智能体的决策方向。在(1)的序列中我们可以定义出t时刻的未来回报$R_t = \sum_{t=0}^{\infty} \gamma^t r_t$。其中，$\gamma$是折扣因子，$\gamma < 1$保证未来汇报有界，强化学习算法收敛。贝尔曼方程（Bellman Equation），更愿意称其为动态规划方程，将“决策问题在特定时间点的值$V(s)$”以“来自初始选择的报酬R_t，和初始选择衍生的决策问题的值”的形式表示（2），从而最佳化该决策问题。

$$V(s) = \max_a [R_t + \gamma \sum_s P(s_{t+1}|s_t, a)\, V(s_{t+1})] \quad (2)$$

智能体遵从回报的引导，调整策略 π，优化(S_t, A_t)，以期获得更高的回报。这回报的“期望值”被定义为，动作值函数（Action Value Function）。

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[R_t|s_t, a_t] \quad (3)$$

准确估计$Q^{\pi}(s,a)$的值就可以选择出应对当前状态的最优动作a^*。

$$a^* = arg\max_{A} Q^{\pi}(s,a) \quad (4)$$

状态价值函数也可以被考虑为所有动作的期望，可以定义状态价值函数$V^{\pi}(s_t)$，评价在当前策略的控制下此刻状态的价值。

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}[Q^{\pi}(s_t, a_t)] = \mathbb{E}_{\pi}[R_t|s_t] \quad (5)$$

有监督学习和强化学习，二者一致之处在于优化某个数据分布下的一个分数值的期望，有监督学习是最小化损失函数，强化学习是最大化奖励函数。但二者的优化途径是不同的，有监督学习增强模型对于数据特征的拟合性能，强化学习则是通过改变策略来调整智能体和环境交互数据的分布表现。

2．深度强化学习方法

经典强化学习的理论是优雅的。对于最优策略的核心思想是，在当前状态下采样所有可能的未来轨迹，对采取不同策略得到的未来回报做比较，选择最优的策略执行。值得指出的估计方法有：蒙特卡洛估计法，这是一种对全周期的无偏估计方法，但其估计每种可能性，所以方差极大；时序差分估计法，通过当前 Q 值和下一时间步 Q 值构建梯度，从而估计策略收敛方向，这种方法是学习率敏感的方法，容易发散；在时序差分的基础上演化出 Q-Learning 方法，其是一种最乐观估计方法，即对于下一步 Q 值的选择，总是选择对乐观的估计值，但其用贪心的策略会导致“过估计”问题。

经典强化学习受限于对价值函数的建模，导致对复杂策略的表达能力不足。遂引入神经网络，增强对环境特征的理解，使其在游戏、机器人控制等领域取得了一些成绩。深度 Q 学习（Deep Q Learning）是在强化学习中应用神经网络的里程碑，在此基础上发展出了基于值迭代的一系列算法，并且与策略迭代的算法相结合构建了当前主流的 Actor-Critic 框架。在值迭代算法中，深度 Q 网络（Deep Q Network，DQN）无疑是最核心的设计，其端到端的设计是，输入(S, A)状态动作对，输出值函数的值，也就是拟合（6），神经网络参数 θ，采用（7）梯度优化。

$$Q_t^{tar} = r_{t+1} + \gamma \max_{a} Q(s,a) \quad (6)$$

$$\theta = \theta - \frac{\partial}{\partial\theta}||Q_\theta(s,a) - Q_t^{tar}||^2 \qquad (7)$$

值得讨论的是，Q 学习方法在演化过程中，为了增强训练表现，提出了很多技巧，包括：软更新，设置两个同样参数的 Q 网络，但其中用于计算Q_t^{tar}的网络参数低频更新，以此提升训练的稳定性，同样设置两个 Q 估计器，Double Q-Learning 中维护的两个 Q 网络互相为对方选择动作，也优化了稳定性；经验回放，由马尔可夫性质定义的状态动作序列只考虑相邻状态的变化，因此为了提高效率，可以将采样的经验存储，通过重放的方式提高样本使用效率，一定程度上使样本之间独立，这也是值迭代方法可以进行离策略（off-policy）学习的关键；Dueling 网络，是在网络结构上的改进，引入优势函数$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$，分支出一个专注于优势学习的神经网络头（head），同时分支出另一个神经网络头（head）估计$V^\pi(s)$，从而能获得更丰富的梯度学习值函数（8）。

$$Q(s,a) = V(s) + (A(s,a\_) - \frac{1}{|A|}\sum_{a\_} A(s,a\_)) \quad (8)$$

深度 Q 学习系列的方法不显式定义策略，只通过合理的损失函数直接优化 Q 值，间接提升策略表现。还有一类方法直接讲策略参数化，通过优化神经网络参数优化直接优化策略，被称为策略梯度迭代。经典的 REINFORCE 策略梯度，针对一条完整的轨迹$\tau = (s_0, a_0, \dots, s_T)$奖励之和$R(\tau)$，参数化策略$\pi_\theta$，由$J(\pi_\theta)$代表奖励或策略表现的期望（9），通过(10)梯度下降方法优化参数θ。

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim P(\tau|\theta)}[R(\tau)] \quad (9)$$

$$\theta_{k+1} = \theta_k + \alpha\nabla_\theta J(\pi_\theta) \qquad (10)$$

对于轨迹τ我们可以将其展开为含有策略的形式，

$$P(\tau|\theta) = p(s_0)\prod_{t=0}^{T} P(s_{t+1}|s_t,a_t)\pi_\theta(a_t|s_t) \quad (11)$$

策略梯度$\nabla_\theta J(\pi_\theta) = \int_\tau^{\square} \nabla_\theta P(\tau|\theta)\, R(\tau)$通过变形和带入，我们可以得到显示策略的策略梯度形式（12）

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim P(\tau|\theta)}[\sum_{t=0}^{T}\nabla_\theta \log\pi_\theta(a_t|s_t)\, R(\tau)] \quad (12)$$

当前策略和经验轨迹的绑定也说明其在策略（on-policy）的性质，但该方法由于对每一条轨迹都进行策略更新，导致该方法方差较大。所以，一个直接的改进想法是为奖励引入一个只与状态有关的基线，通过只计算增量的形式减小方差。这种优化思路也与后来的 Actor-Critic 方法，有一些联系，在此说明的是，Actor 是基于策略迭代直接与环境交互的部分，Critic 是基于值迭代的 Q 网络，用来评估当前策略的价值。所以也可以认为 Critic 所

得到的 Q 值也是一种基线。

Actor-Critic 方法为基础框架的算法，是现今主流的深度强化学习方法。对于 Actor 的优化主要在于，通过引入优势函数定义的策略梯度，减少估计的方差，（估计时方差的产生因素在于轨迹总和奖励的不稳定，故用优势函数代换轨迹总和奖励，只计算增量部分，可以减小方差）得到新的策略梯度$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim P(\tau|\theta)}[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, A_\varphi(s_t, a_t)]$。对于在各类任务中性能优秀的近端策略优化算法（Proximal Policy Optimization，PPO）其核心改进思路也是通过重要性采样技术处理 Actor 的优势函数，从而给出了梯度下降的最佳步长，平衡了梯度更新的速度和质量。对于 Critic 的优化，深度确定性策略梯度算法（Deep Deterministic Policy Gradient，DDPG）延续软更新的思路，维持两个不同频率更新的 Q 网络，保证 Q 学习目标的稳定性，同时，通过确定性策略使 Actor 输出一个确定的值（而不是类似其他策略梯度算法输出一个概率分布）。另一个值得讨论的算法是异步策略梯度（Asynchronous Advantage Actor-Critic，A3C），该方法采用联邦（Federated）思想，实例化多个 Actor-Critic 网络与环境交互，异步更新全局网络参数，从而汇聚出一个优势策略。

3.多智能体层次化强化学习方法

在现实问题中有很多情况是低反馈且长尾的，奖励稀疏且要通过多个步骤达成一个任务，这对常规强化学习方法的能力提出挑战。为解决这类问题，层次化强化学习方法将复杂的、长时间步的任务进行分解，进而取得较好的效果。主流的任务分解思路分为两类，一类是在长尾任务中设置多个子目标，推进子目标的完成，从而达到最终的决策目标，另一类是需要反复组合调用多种“原子”行动，完成特定的任务。

第一类，基于子目标的层次化学习，如图 2 所示，通常被分解为两个步骤，上层步骤输出子目标，下层步骤根据子目标完成相应的动作。通常两个步骤由两个子智能体完成决策，他们各司其职，异步工作，上层智能体由时序敏感型网络构建，观察环境状态，捕获任务转换时机，隔若干步发布子目标，下层智能体接受多个输入头，综合环境状态和来自上层智能体的子任务，负责持续与环境交互，达成任务目标，二者相互透明交付结果，隔离进行梯度提升。这一类层次化强化学习方法，在训练时还有很多技巧，一类思路是为两类智能体设置不同的奖励机制，另一类思路是上层智能体输出的子目标是动态规划的，让上层的子目标不断逼近最终决策目标。

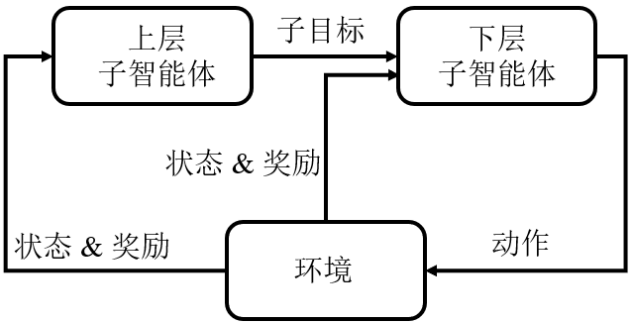

图 2 基于子目标的层次化学习

第二类，基于技能的层次化学习，如图 3 所示，和基于子目标的层次化学习方法类似，有两层智能体构成，其中上层智能体负责学习执行特等任务需要组织何种特定技能，下层智能体学习一些与任务无直接关联的通用动作——技能。针对两类智能体的构建有两类重要的训练方法，一类是针对下层子智能体的，为每种任务设计专门的网络结构是一件费力且低效的工作，因此往往采用随机神经网络来构建下层子智能体，随机神经网络随机化构建输入层到隐藏层的参数，故通过一种结构能广泛学习各类型的技能动作。另一类是针对上层子智能体的元学习方法，对于不同任务下层子智能体策略在训练完成后不会修改，而上层子智能体的策略会根据任务不同做相应调整，也就是在两类智能体更新参数时认为另外的一类子智能体只是环境而已，这种训练的方式，使上层智能体对下层技能的应用和泛化能力更强。

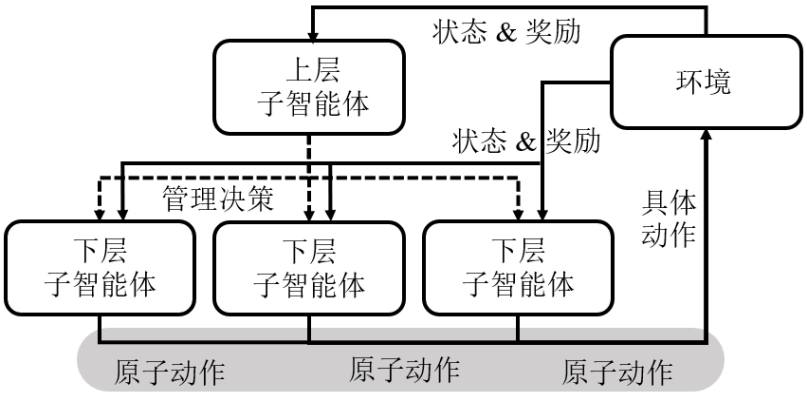

图 3 基于技能的层次化学习