第十讲
强化学习（II）
Lecture 10
Reinforcement Learning (II)

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

  The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

背景
Background

- 人们关于环境和人类自身的主要知识来源，很大一部分是通过与环境的交互进行学习来获得的。这可能是我们在考察什么是学习时，看到的一个相当广泛的形式。在此过程中，虽然不存在明确的老师，但学习主体通过感觉运动与环境发生了直接的联系。这种联系会产生大量关于因果关系、行动后果以及为实现目标应该做什么的信息。例如，无论我们是学习驾驶汽车还是进行对话，我们都敏锐地意识到我们的环境如何对我们的行为做出的反应，并且我们试图通过我们的行为来影响所发生的事情。实际上，从交互中学习是几乎所有学习和智力理论的基本思想。

  A major source of knowledge about the environment and one's ego oneself throughout a person's life, is acquired by learning via interacting with the environment. This is probably the first idea occurring to us when we think about the nature of learning. During this process, although there is no explicit teacher, it does have a direct sensorimotor connection to its environment. Exercising this connection produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals. For example, whether we are learning to drive a car or to hold a conversation, we are acutely aware of how our environment responds to what we do, and we seek to influence what happens through our behavior. Learning from interaction is a foundational idea underlying nearly all theories of learning and intelligence.

背景
Background

- 另一个深层次的问题是，学习的本质是确实学到了新的知识，达到这一目的的保证是与环境的交互须维持一定的频率或持续一定的时间，而要做到这一点，必须有内在的因素驱动个体这么做。这个内在因素，很多时候便反映为个体看到的采取特定动作与环境交互时所获得的奖励，无论这种奖励为近期还是远期。反之，如果个体看到的是惩罚，则此种行为一般不会持久，很难达到学习的目的。

  When we carry the topic further, it is natural to reflect on the essence of learning, which is to really learn something new. To achieve this, there must be some intrinsic force or motivations that drive the subjects to take some actions to interact with the environment in sufficient frequent or persisting manner. And this intrinsic factor, in most cases are in the form of rewards perceived or anticipated by the subjects when adopt specific actions to interact with the environments. It doesn't matter the rewards are short-term ones or long-terms ones. However, if the subjects would perceive potential punishment, they would stop the interactions, and to discuss learning in this circumstance usually makes no sense.

背景
Background

- 类比生物体及其学习的过程及内在机制，特别是从计算的角度，探索从交互中学习的方法，即强化学习。强化学习的重点不是直接对人或动物如何学习进行理论化，而是主要探索学习形式的实现情况，并评估各种学习方法的有效性。也即，我们要藉由神经科学得到的启发，应用到人工智能的工程研究或实践中来。我们探索能够有效解决科学研究或经济活动中的学习问题的机器学习方法设计，通过数学分析或计算实验的方式评估设计。一言以敝之，我们根据强化学习的侧重，设计与评估从交互中进行目标导向学习的方法。

  In analogous to the learning process and the underlying mechanisms of animals, especially from the computational perspective, we explore to learn from interactions in a computational approach, termed as reinforcement learning (RL). Rather than directly theorizing about how people or animals learn, we primarily explore implementation of the learning and evaluate the effectiveness of various learning methods. That is, we adopt the perspective and inspiration of neuroscience findings to the artificial intelligence research and practice. We explore designs for machines that are effective in solving learning problems of scientific or economic interest, evaluating the designs through mathematical analysis or computational experiments. To sum it up, we design and evaluate methods which merits the characteristics of reinforcement learning, that include but are not confined to the focus on goal-directed learning from interaction.

背景
Background

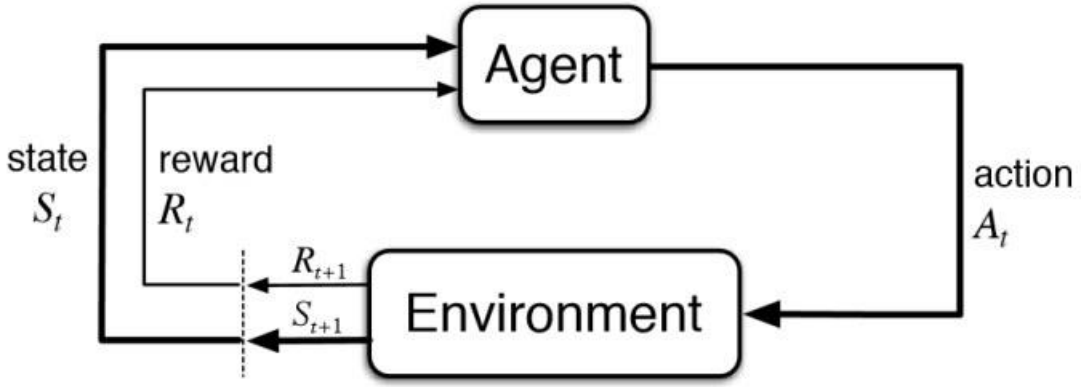
- 在英文中，强化学习和机器学习等以“ing”结尾的术语一样，是一个问题的同时也包含解决该问题的方法，也即涵盖研究某个问题及其解决方法的领域。具体的说，强化学习是研究如何将环境映射到动作以最大化奖励信号这样一个问题及其答案的领域。一般地说，对于强化学习这种范式，学习者不会被告知要采取何种行动，而是必须通过尝试来发现哪些行动会产生最大的回报。在某些情况下，动作不仅可能会影响直接奖励，还会影响下一个状态，并由此影响所有后续奖励。因此，试错搜索和延迟奖励，是强化学习的两个最显著的特征 。

  Reinforcement learning, like topics such as machine learning, whose names end with “ing”, is simultaneously a problem and a class of solution methods that work well on the problem. Concretely speaking, the terminology indicates a field that both studies this problem and its solution methods. Therefore, reinforcement learning is a field to learn how to map the environment to actions so as to maximize the reward signal, both the problem itself and solutions. In such a paragidm, the learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In some challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics, namely, trial-and-error search and delayed reward, are the two most important distinguishing features of reinforcement learning.

概念
Concepts

- 下面我们来从计算的角度介绍强化学习。通俗地讲，强化学习希望通过构建直接从交互中学习的方式以达到特定问题的目标。学习者或代理通过过与环境交互，选择动作，响应环境，而后环境将新的情况呈现给学习者；同时，学习者还会收到环境产生的奖励，代理在通过其选择的动作随着时间的推移寻求最大化的奖励的过程中，逐渐完善策略，达到学习的效果。

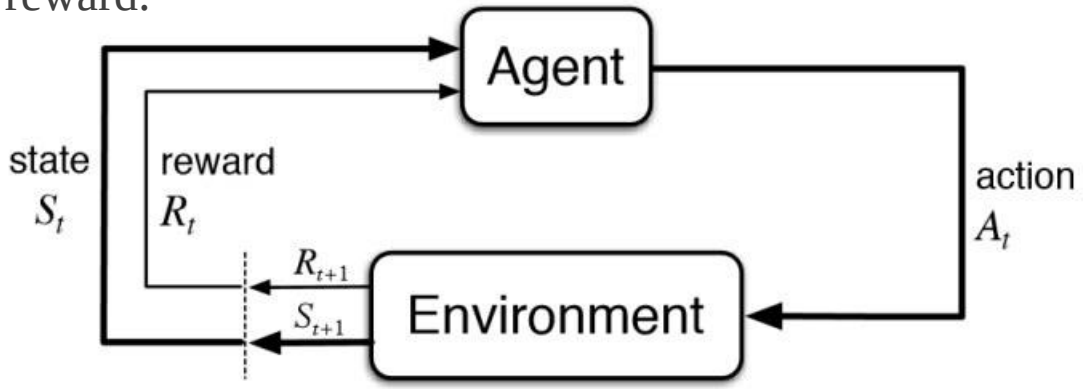
Now we introduce the reinforcement learning (RL) from the computational perspective. In general, RL is by straightforwardly framing the problem of learning from interaction to achieve a goal. The learner or agent interacts with the environment, take actions, and then the environment responding to these actions and presenting new situations to the agent. The environment also gives rise to rewards in numerical values that the agent seeks to maximize over time through its choice of actions. In such a process, the agent polish the its policy and reach the optimal state for learning.



概念
Concepts

- 具体地说，在利用强化学习解决问题的决策框架中，待解决问题的解决步骤，映射为待解决问题上下文或环境中的代理的具体动作，该代理的每个动作都会影响环境，或者说代理可感知的未来状态。选择的动作与环境状态的适配程度，即解决问题过程中每一步的优劣，由奖励信号来衡量；问题解决的理想程度由回报的最大化程度来界定。

  In detail, in the decision-making framework of RL, the procedures to solve a problem maps to the series of actions chosen by the agent in the context of the problem to be solved, or alternatively, the environment. Each action taken by the agent influences the agent's future state. The extent of appropriate action choice, indication of sub-procedure quality, is measured by the scalar reward signal, and the scale of successful problem solving depends on the maximised future reward.

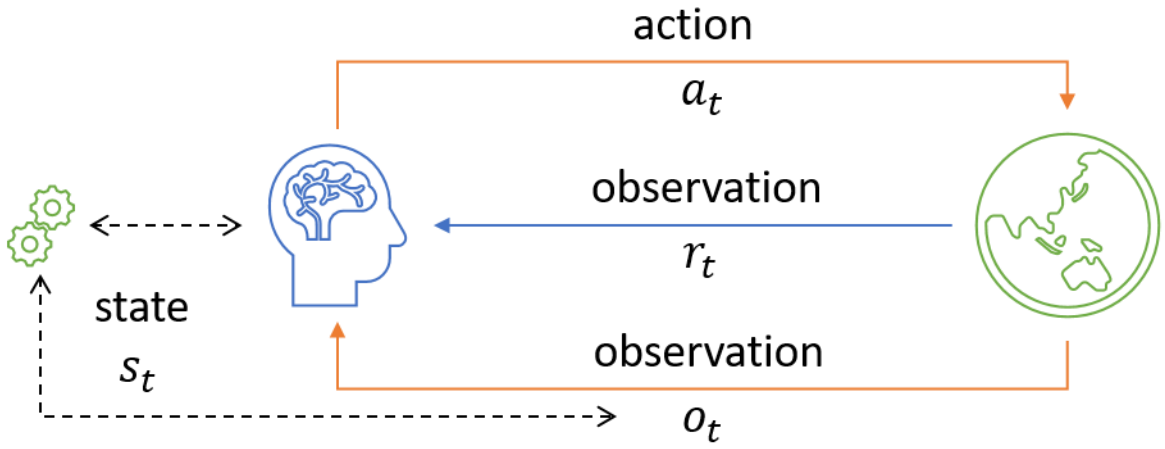


概念
Concepts

- 现在利用强化学习的场景，许多时候都与深度学习相结合。这是因为当环境状态改变时，代理接收环境反馈的状态，更重要地是感知这种状态。当环境状态是视觉的形式时，鉴于深度学习在视觉任务方面的取得，利用深度学习自然是题中之义。同时，当把决策与环境感知融在一起均用神经网络表示时，符合神经网络端到端的特点，近些年来许多成果正是基于此。

  Currently lots of applications that utilize RL actually incorporate DL as well. Actually, to perceive the new state received from the environment is the key for RL to be success, especially the state presents to the agent in some visual way. In this circumstance, to adopt DL which already made the exceptional achievements in vision tasks is a natural and essential choice. It is typically coincided with the “end-to-end” philosophy of DL if the perception of state and decision-making process are represented in one network, and recent grand accomplishments are mainly based on such a architecture.

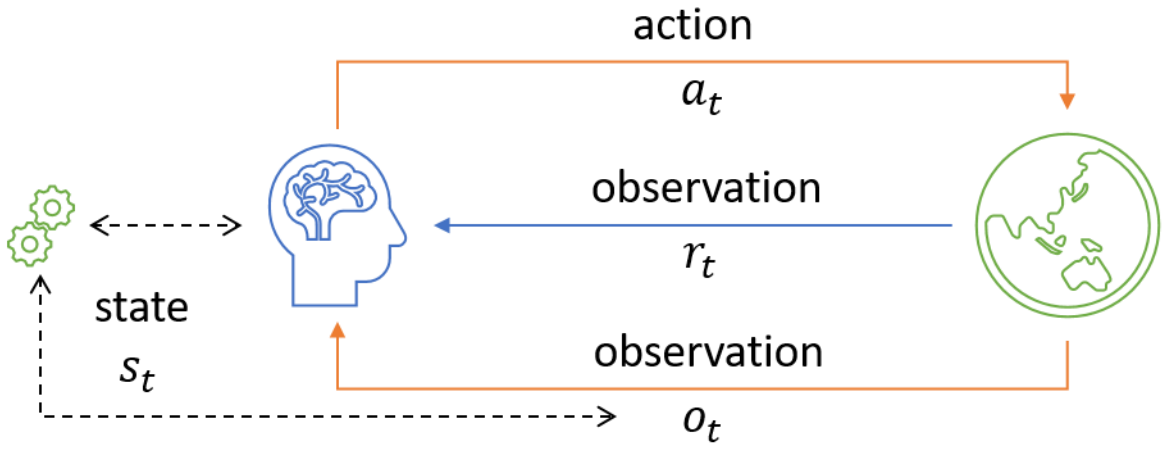
概念
Concepts



- 基于上述原因，目前利用强化学习算法，代理与环境的交互如图所示。虽然代理可以直接观察到当前情形，当仅由目前一次情形观察可能不能决断出当前状态，这由人类的情况亦可知，因此可能需要连续多次观察。由于环境本身并不会缓存代理每次观察时看到的环境的快照，则可能需要代理自己缓存观察快照并根据这些快照决断出当前状态。

  Due to the above reason, the interactions between environment and agent of current RL algorithms is illustrated below. Although agent can directly observe the current situation, however, a sole observation might not be sufficient to generate a current state, analogous to the human's case. Hence, the current state might depend on a consecutively series of observations. However, due to the fact environment will not cache the observations each time the agent perceives, the agent might have to store the observations to deduce the current state.

概念
Concepts



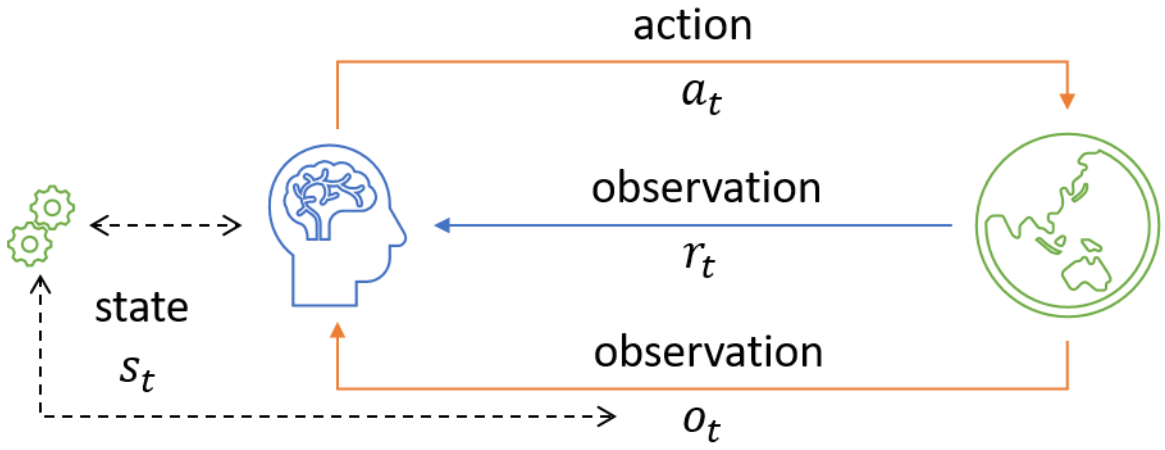
- 在对图示进行深入解读之前，我们先介绍一些概念。

  We introduce some concepts before proceeding to further interpret the diagram.

- 首先是学习者，其可能有多个称呼，可以称为强化学习代理，人工代理，智能体，或干脆就说代理，在具体问题中可能指人或拟人化的事物，或者干脆就是一段代码。

  First is about the learner, which can have quite different names, for example, reinforcement learning agent, artificial agents, intelligent agents, or simply agent. It can be human beings or anthropomorphic objects, or simply, a piece of code.

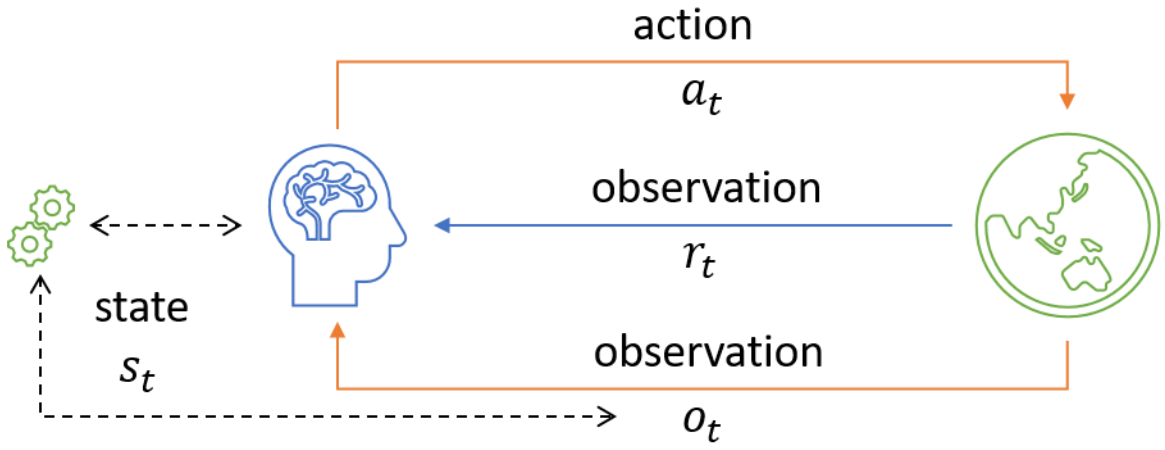
概念
Concepts



- 环境：包含在学习者之外的所有内容，一般均可以归为环境。因此，环境的内涵与外延，须以智能体的选择为参照，可能在一种情况下视为学习者或智能体的对象，在另外一种情况下，先前的智能体的一部分成为新的智能体，而其余部分加上之前的环境，则构成新的环境。

  Environment: roughly we can say that all the things outside the agent constitute the environment. So, the scope of the environment must be by referencing to the definition of the agent. An agent consisting of many parts on one occasion might split itself on another occasion to have only some parts become a new agent, and other parts combined with the previous environment, to form a new environment.

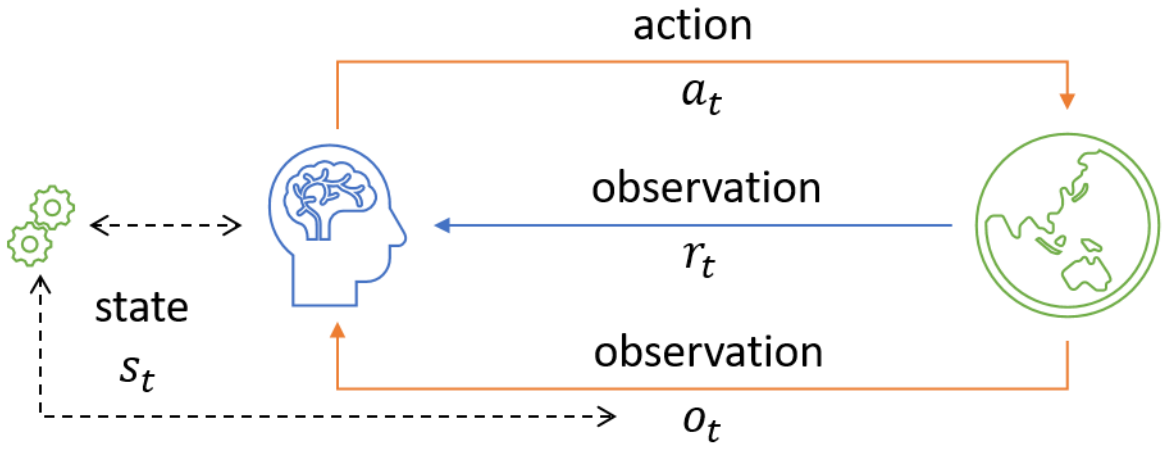
概念
Concepts



- 在上述概念的基础上，我们对图示的解读可表述为：在当前时刻t，代理执行对环境有一观察o_t，同时接收到环境给与的奖励r_t，然后执行动作a_t；同时，环境对代理执行的动作a_t作出反映，演变为代理在下一时刻看到的情形o_{t+1}，同时向代理反馈奖励r_{t+1}。

  Based on what is agent and what is environment, the diagram can be interpreted as: for the current time step t, the agent has a observation of the environment o_t and simultaneously receives the reward r_t from the environment. The agent also execute the action a_t to affect the evolvement of the environment, which cause the environment to present the situation at time step $t+1$ and the next reward r_{t+1}.

概念
Concepts



- 上述描述过程也刻画了强化学习的一个特点，即其是一个顺序决策过程。如果能从形式上刻画此决策过程，则可为决策过程的求解带来便利。而利用马尔科夫决策过程（MDP）的相关知识，则可以做到从理论上达到形式描述的统一，虽然未必严格用概率的知识求解。

  The above interpretation reflects one characteristic of RL, aka, a sequential decision-making process. If such a process can be depicted in a canonical form, it can facilitate the solving process. Based on the Markov theories, a canonical description of RL at least from theoretic prospective can be achieved, though to form the solution might diverse from probability theories.

- 为了对利用强化学习的顺序决策过程（在马尔科夫决策过程理论下）有一个统一的刻画，我们接着介绍一些概念，并考虑用公式的形式表述。

  To depict the sequential decision-making characteristic of RL (under Markov decision process or MDP) in a uniform way, we introduce more concept and consider to express them in formula.

概念
Concepts

- 首先对一个确定问题，过程的每一步t属于自然数的一个子集$\mathcal{T}$，$\mathcal{T}$可能是有限集合，也可能是无限集合。环境可能的状态s的集合记为$\mathcal{S}$，同理，$\mathcal{S}$可能是有限集合，也可能是无限集合。虽然总的动作集合可以记作$\mathcal{A}$，但当代理处于不同的状态时，可以采取的动作a往往与当前状态s相关，即$a \in \mathcal{S}(a)$。代理获得的奖励一般为实数，记为$r \in \mathcal{R}$。由于马尔科夫决策过程依然是基于概率理论，我们用大写字母S_t、A_t表示在t时刻处于特定状态与采取特定动作的随机变量，用小写字母s_t、a_t表示在t时刻实际的状态与采取的动作。
- 一般说来，状态就是代理在当前时刻t观察到的环境的状况s_t。当需要连续n步观察才能得出当前状态时，$s_t = f(o_t, o_{t-1}, \cdots, o_{t-n+1})$。特别对于完全可观察环境，有$s_t = f(o_t)$。在马尔科夫决策过程框架下，环境在代理的作用下的演化并不是固定的，可以是随机的，但却与上一个时刻的状态与代理采取的动作相关，即$p(s') \triangleq \Pr\{S_{t+1} = s' | S_t = s, A_t = a\}$。 显然$p$决定了系统的演化，有时称其为MDP的动态函数。注意，由于对$\forall s \in \mathcal{S}$，$a \in \mathcal{A}$，有$\sum_{s' \in \mathcal{S}} p(s') = 1$，即$p$可视为确定的概率分布函数。

概念
Concepts

- First for a given problem, every step t of the sequence, we have $t \in \mathcal{T}$, where $\mathcal{T}$ can be an either finite set or infinite set. The collection of all possible states s of the environment is denoted as $\mathcal{S}$, with $\mathcal{S}$ a finite set or infinite set. Although the total actions can be performed by the agent can be denoted as $\mathcal{A}$, but the available actions in a specific state might be restricted, so we usually denote as $a \in \mathcal{S}(a)$. The rewards obtained by the agent are usually real numbers, denoted as $r \in \mathcal{R}$. Due to MDP is based on the probability theory, we use capital letters S_t, A_t denote the random variables such as the state and actions at time step t, lower case letterss_t, a_t to represent the actual state and action taken at time step t.
- Generally, state s_t is what agent perceived or observed from the environment at the current step t. If consecutive observations of the environment are necessary to deduce the current state, we have $s_t = f(o_t, o_{t-1}, \cdots, o_{t-n+1})$. Specifically, for fully-observed environment we have $s_t = f(o_t)$. Under MDP, the evolvement of environment is not deterministic but with stochasticity, which depends on the intermediate state and action at last time, so we have $p(s') \triangleq \Pr\{S_{t+1} = s' | S_t = s, A_t = a\}$. p is called the dynamic of MDP since it decides the progress of the environment. Since for $\forall s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have $\sum_{s' \in \mathcal{S}} p(s') = 1$, which means p is just a deterministic probability distribution function.

概念
Concepts

- 强化学习中，与生物系统类似，驱动代理进行学习提高的，是从环境传递给代理的称为奖励的信号$R_t \in \mathcal{R}$。代理的目标，粗略来讲，是最大化它收到的奖励总量，即长远地看，代理所采取的策略，不是最大化即时奖励，而是最大化累积奖励。因此对强化学习的探讨，我们假设与生物系统类似，都存在一个奖励假设：我们的目标或目的都可以被认为是接收到的标量奖励值的累积，或其期望值的最大化。

  In reinforcement learning, in analogous to the biological system, what drives the agent to learn or improve is the reward signal $R_t \in \mathcal{R}$ passing from the environment to the agent. Roughly speaking, the agent's goal is to maximize the total amount of reward it receives. This means in the long run, the strategy that agent takes is to maximize not immediate reward, but cumulative reward. Therefore, to the context of RL, we can assume the existence of the reward hypothesis just as the biological subjects, that all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum or expectation of the received scalar rewards.

概念
Concepts

- 由于代理的目标与奖励信号r息息相关，我们可将奖励与神经科学中奖励系统的奖励的概念相比照。虽然在生物系统中，奖励可以描述成快乐或痛苦的体验，但从计算的角度，强化学习中奖励信号一般采用数字化的标量。同时，强化学习中奖励一般不说惩罚，因为值为负的奖励信号可视为惩罚。一般地说，奖励信号是智能体调整策略的主要依据。如果策略选择的动作之后奖励变低，那么将来可能会更改策略以在该情况下选择其他动作。通常，奖励是环境状态和所采取动作的函数，可以是固定函数，也可以是随机函数。

  Since the goal of agent is closely related to the reward signal r, we can compare it with the counterpart in the reward system in the neuroscience. We might think of rewards as analogous to the experiences of pleasure or pain in a biological system, however, we have to digitize the reward from the computational perspective. Meantime, in reinforcement learning, we mainly talk about the positive reward solely instead of punishment since the value can be negative to indicate punishment. Usually, the reward signal is the primary basis for altering the policy; if an action selected by the policy is followed by low reward, then the policy may be changed to select some other action in that situation in the future. In general, reward signals are functions of the environment states and the actions taken, they may appear in the form of deterministic or stochastic.

概念
Concepts

- 我们说，代理的目标是最大化它在长期与环境交互的过程中所获得的总奖励，这句话怎么理解呢？或怎样形式地表示呢？假设从t时刻开始，代理所收到的奖励为R_t，R_{t+1}，…，R_T，定义$G_t \triangleq R_t + R_{t+1} + \cdots + R_T$，称为在步骤$t$的回报。此为当$T$有限的情况，当$T$无限时，为了保证$G_t$有定义，通常会引入一个衰变因子$\gamma$，即$G_t \triangleq R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$。或$G_t = R_t + \gamma G_{t+1}$。由于不同的动作选择会有不同的轨迹，即$G_t$可能会有不同的取值，因此我们说，代理的目的是最大化G_t的期望。

  An question remaining unanswered is how to interpret or how to formally express that the agent's sole objective is to maximize the total reward it receives over the long run. Assume begin from step t, the sequence of rewards received by the agent are R_t, R_{t+1}, …, R_T, etc., then define $G_t \triangleq R_t + R_{t+1} + \cdots + R_T$, as return for the agent at time step t. This is for T is finite. When T is infinite, an discounting factor γ will be introduced, that is $G_t \triangleq R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$, or $G_t = R_t + \gamma G_{t+1}$. Since different actions lead to different trajectories of the system, which means G_t may have different values, therefore, the goal of the agent is to maximize the expectation of G_t.

概念
Concepts

- 在明晰目标之后，代理为了达到目标，必然会采用最优的策略π。策略可类比于神经科学中的决策，即从感知的环境状态到在这些状态下要采取的行动的映射。策略是强化学习中智能体要学习的核心，因为它本身就足以智能体确定在当下感知的环境状态下的行为。一般来说，策略可以是固定的，指定特定的动作；也可以是随机的，指定每个动作的概率。因此策略的实现，在一些情况下可能是一个简单的函数或查找表，而在其他情况下，策略可能涉及大量的计算。

  To achieve the goal set, a agent must adopt the optimal strategy or policy π. The concept policy is analogous to the concept of decision-making in neuroscience, aka, roughly corresponding to what in psychology would be called a set of stimulus–response rules or associations. The policy is the core of a reinforcement learning agent in the sense that it alone is sufficient to determine behavior from the perceived state of the environment. In general, policies may be deterministic, specifying a particular action; or they can be stochastic, specifying probabilities for each action. Hence, the implementation of policy in some cases may be a simple function or lookup table, whereas in others it may involve extensive computation.

概念
Concepts

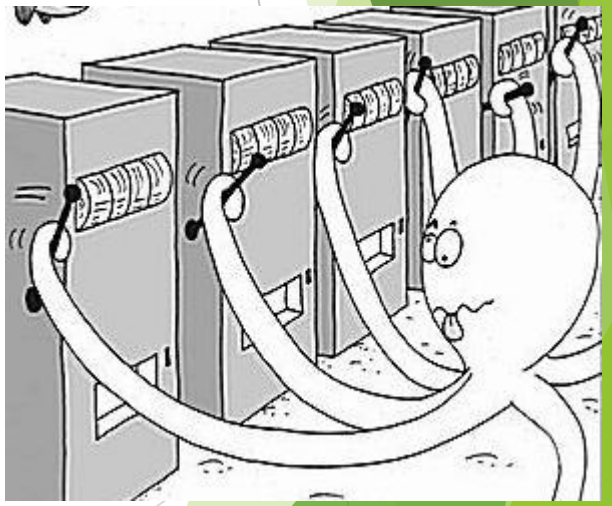
- 一个我们一直没有探讨的问题是，我们为什么需要策略？强化学习与其它类型的机器学习的一个很大的不同是“探索”和“利用”的权衡。为了使获得的奖励最大化，代理可能倾向于“利用”过去尝试过的且能有效产生奖励的动作。但是要发现这样的动作，它必须充分“探索”或“尝试”以前没有选择过的动作。智能体必须利用它已经体验过的动作来获得奖励，但它也必须探索才能在未来做出更好的动作选择。例如，在随机任务中，必须多次尝试每个动作才能获得对其预期奖励的可靠估计。探索与开发利用的权衡问题，实际上本身也是策略，我们正是通过代理的目标，来优化这个策略，使得达到最优策略的同时，也将问题解决。

  A question remains unanswered is why we need the policy? One challenge distinguishing RL from other kinds of machine learning method, is the trade-off between exploration and exploitation. To obtain as much reward, an RL agent must exploit actions tried in the past and found to be effective in producing reward. But to discover such actions, it has to explore actions un-selected before. In summary, the agent not only has to exploit what it has already experienced in order to obtain reward, but also to explore action in order to make better selections in the future. For example, on a stochastic task, each action must be tried many times to gain a reliable estimate of its expected reward. The balance between exploration and exploitation itself is a policy. We optimize such a policy via the goal of agent, to reach the optimal policy as well as solve the problem.

概念
Concepts

- 在上面两个问题明晰的情况下，我们来看一个具体的策略。我们考虑多臂赌博机问题，如图所示。当用户投注之后，可以选择去拉动赌博机的不同拉杆。当选择拉杆不同时，从不同出口吐出的币的数目不一。一个策略是每5次中，我们前4次选择在所有的尝试中，平均吐出的币最多的那个拉杆（利用策略），最后一次随机选择一个拉杆（探索策略）。当我们每个拉杆尝试比如至少10次之后，更改为新的策略，在新策略下，每次只拉平均出币最多的拉杆。

  When we are clear of the above questions, we resort to specific policies. Let's consider the muilti-armed bandit problem. After the customer bets, he can choose to pull arbitrary levers. However, different levers will trigger the machine to output different number of coins. One particular policy, for every five trying, the first four attempts is to choose the lever which outputs the maximal average coins, and the last one just choose randomly. After we repeat this strategy to have each levered pulled at least ten times, we switch to a new strategy, which always chooses the lever with maximal average coin output.



概念
Concepts

- 对于上面的第一种策略π_1，我们假设有N个臂，每个臂用i表示。臂i第j次出币个数为$r_j^{(i)}$，臂i的总共被拉动的次数$c^{(i)}$，$a_t = \pi(t) \in [1, \cdots, N]$表示在第$t$次拉动哪个臂。第二种策略记为$\pi_2$。

  For the first policy above denoted as π_1, we first assume that there are total N arms, each represented as i. For the arm i, the output coins for j-th time is $r_j^{(i)}$, the total number of i-th arms pulled is denoted as $c^{(i)}$, $a_t = \pi(t) \in [0, 1, \cdots, N]$ represents which arm pulled at time step t. The second policy is denoted as π_2.

$$\pi_1(t) = \begin{cases} \underset{i}{\operatorname{argmin}} \left\{ \mathbb{E}\left[r_j^{(i)} \middle| j = 1,2, \cdots, c^{(i)} \right] \right\}, i \in [1, \cdots, N], t \not\equiv 4 (\operatorname{mod} 5) \\ (\operatorname{rand}() \bmod N) + 1, t \equiv 4 (\operatorname{mod} 5) \end{cases}$$

$$\pi_2(t) = \underset{i}{\operatorname{argmin}} \left\{ \mathbb{E}\left[r_j^{(i)} \middle| j = 1,2, \cdots, c^{(i)} \right] \right\}, i \in [1, \cdots, N]$$

概念
Concepts

- 一般情况下，我们需要间接地、形式化地表示策略，为此我们引入价值函数。既然希望价值函数与策略相关，则其不能像奖励信号那样主要去衡量短期意义上的好坏，而能表明的是长期意义上的好坏。因此，一个状态的价值能够是一个智能体可以期望的从此状态开始，累积到未来（如学习结束）的奖励总和。换而言之，智能体通过交互去达到特定的环境状态的考量中，奖励决定了直接可取性，而价值表明了考虑后续状态及对应的奖励之后的长期可取性。例如，一个状态可能总是产生较低的即时回报，但其后续状态经常产生高的回报，因此仍然具有较高的价值。

  Generally, we need to express the policy in an indirect and formal way, for this we introduce the value function. Since we expect the value function correlates to the policy, value function should specify which state is good in the long run instead what is good in an immediate sense just as the reward signal. Hence, the value of a state indicates the total amount of reward an agent can expect to accumulate over the future, starting from that state. Alternately speaking, whereas rewards determine the immediate, intrinsic desirability of specified environmental states reached via interaction, values indicate the long-term desirability of states after taking into account the states that are likely to follow and the rewards available in those states. For example, a state might always yield a low immediate reward but still have a high value because it is regularly followed by other states that yield high rewards.

概念
Concepts

- 我们来定义价值函数，首先，要明确价值函数是在固定当前策略的前提下定义的。例如，对于给定的状态s，不同的策略导致的后续动作选择继而的状态转变，所导致的不同轨迹所对应的回报G是不同的，我们正是根据代理的目标，通过学习，或者说策略迭代，从而达到最优策略的。对于给定的策略π，第一种价值函数称为状态价值函数，其定义如下式所示，代表按照策略π，从状态s开始，回报的期望值。

  We now define the value function. First we should be aware that the value function is defined in the premise of fixed policy. For example, for the given state s, different policies lead to different consecutive actions and state transitions, resulting in different trajectories and corresponding returns G. By learning or policy iteration to target the goal of agent, we find the optimal policy and solve the problem. For a given policy π, the first is call state-value function, defined as below, indicating begin from state s and according to π, the expectation of returns.

$$v_\pi(s) \triangleq \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{+\infty}\gamma^k R_{t+k}\,|S_t = s\right] \quad s \in \mathcal{S}$$

概念
Concepts

- 第二种价值函数称为状态动作价值函数，或简称动作价值函数，其定义如下式所示，代表按照策略π，从状态s和动作s开始，回报的期望值。

  The second is call state-action-value function, defined as below, indicating begin from state s and action s, according to π, the expectation of returns.

$$q_\pi(s,a) \triangleq \mathbb{E}_\pi[G_t|S_t=s,A_t=a] = \mathbb{E}_\pi\left[\sum_{k=0}^{+\infty}\gamma^k R_{t+k}\,|S_t=s,A_t=a\right] \quad s\in\mathcal{S}, a\in\mathcal{A}(\pi(s))$$

- 价值函数v_π和q_π的一个特点是可以根据经验估计，这点在我们介绍价值函数的另一个基本属性，即满足递归关系后，更加清晰 。

  The value functions v_π and q_π can be estimated from experience, this is more obvious when a fundamental property of value functions, aka, recursive relationships are established.

概念
Concepts

- 我们首先推导v_π和q_π各自满足的递归关系。

  We first deduce the recursive relationship satisfied by v_π and q_π independently.

$$
\begin{aligned}
v_\pi(s) &\triangleq \mathbb{E}_\pi[G_t|S_t=s] = \mathbb{E}_\pi[R_t+\gamma G_{t+1}|S_t=s] \\
&= \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\left[r+\gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}=s']\right] \\
&= \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r+\gamma v_\pi(s')]
\end{aligned}
$$

$$
\begin{aligned}
q_\pi(s,a) &\triangleq \mathbb{E}_\pi[G_t|S_t=s,A_t=a] = \mathbb{E}_\pi[R_t+\gamma G_{t+1}|S_t=s,A_t=a] \\
&= \sum_{s'} p(s'|s,a)\left[r+\gamma \pi(a'|s')\mathbb{E}_\pi[G_{t+1}|S_{t+1}=s',A_{t+1}=a']\right] \\
&= \sum_{s'} p(s'|s,a)\left[r+\gamma \pi(a'|s')q_\pi(s',a')\right]
\end{aligned}
$$

概念
Concepts

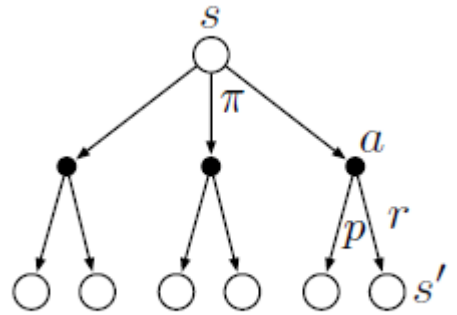
- 我们再推导v_π和q_π相互满足的递归关系。

  We first deduce the recursive relationship satisfied by v_π and q_π intervalently.

$$v_\pi(s) \triangleq \mathbb{E}_\pi[G_t|S_t = s] = \sum_a \pi(a|s)\mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \sum_a \pi(a|s)q_\pi(s,a)$$

$$q_\pi(s,a) \triangleq \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \sum_{s'} p(s'|s,a)\left[r + \gamma\mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right]$$

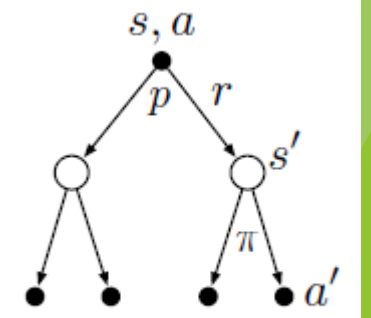
$$= \sum_{s'} p(s'|s,a)[r + \gamma v_\pi(s')]$$



Backup diagram for v_π

- v_π和q_π 各自与相互之间的关系式称为贝尔曼方程，并可表示成如图所示的回溯图。

  The equations related v_π and q_π solely or mutually are called Bellman equations, which can be represented as back up diagrams as right.



q_π backup diagram

概念
Concepts

- 当有了这些价值函数之后，一个问题是如何求最优策略。如前面所讲，解决强化学习任务，是通过寻找一个从长远来看能获得最多回报的策略实现的。首先，如果对于所有的状态，策略π的预期回报大于或等于π'的预期回报，则称策略π等于或优于策略π'。换而言之，$\pi \geq \pi'$当且仅当对所有$s \in \mathcal{S}$，有$v_\pi(s) \geq v_{\pi'}(s)$。这样一来，价值函数定义了策略的一个偏序排序。如果一个策略等于或优于其他所有策略，我们将其记为最优策略π^*，其对应的状态价值函数，称为最优状态价值函数，定义为$v_*(s) \triangleq \max_\pi v_\pi(s)$。同理，最优动作价值函数定义为$q_*(s,a) \triangleq \max_\pi q_\pi(s,a)$。注意最优策略$\pi^*$可能不止一个。

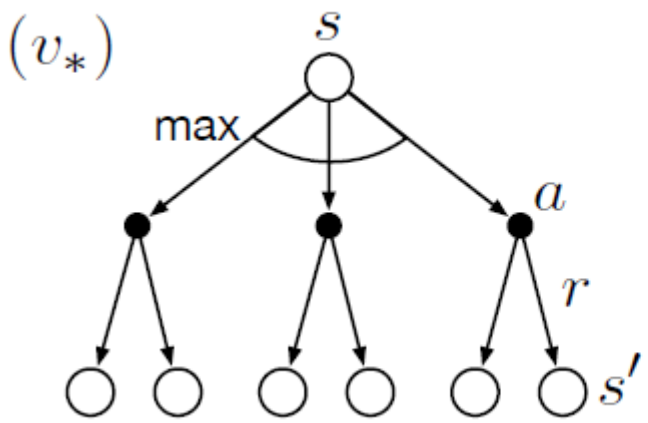
  After we defined all these value functions, a natural question is how to find the optimal policy. As stated before, roughly speaking, to solve a RL problem is by finding a policy to target the maximal expected return over the long run. First, we say a policy π is better than or equal to a policy π' if its expected return is greater than or equal to that of π' for all states. In other words, $\pi \geq \pi'$ if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in \mathcal{S}$. Value functions define a partial ordering over policies. If one policy is equal to or better than all other policies, it is denoted as an optimal policy π^*. The corresponding state-value function is called the optimal state-value function, denoted as $v_*(s) \triangleq \max_\pi v_\pi(s)$. As the same, the action value functions is defined as $q_*(s,a) \triangleq \max_\pi q_\pi(s,a)$. Note that the optimal policy maybe more than one.

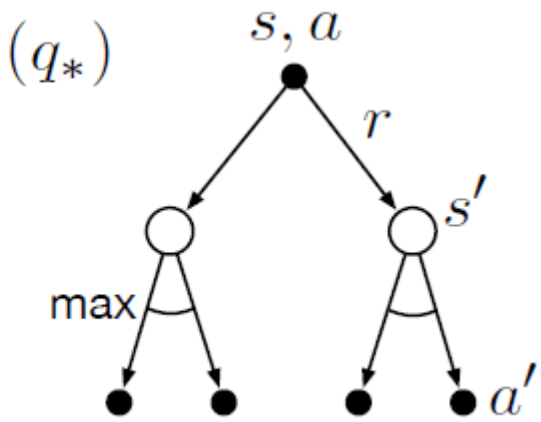
概念 Concepts

- 对于最优状态价值函数与最优动作价值函数，也有相应的Bellman方程，推导如下；方程对应的回溯图如下图所示：

  There are also Bellman equations for the optimal state value function and action value function, as deduced below. The corresponding back-up diagrams are also illustrated below:

$$v_*(s) = \max_{a \in \mathcal{S}(a)} q_{\pi^*}(s,a) = \max_a \mathbb{E}_{\pi^*}[G_t | S_t = s, A_t = a]$$
$$= \max_a \mathbb{E}_{\pi^*}[R_t + \gamma G_{t+1} | S_t = s, A_t = a] = \max_a \mathbb{E}[R_t + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$
$$= \max_a \sum_{s'} p(s'|s,a)[r + \gamma v_*(s')]$$
$$q_*(s,a) = \mathbb{E}[R_t + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] = \sum_{s'} p(s'|s,a)[r + \gamma v_*(s')]$$





概念
Concepts

- 当我们有$v_*(s)$和/或$q_{\pi^*}(s,a)$之后（实际上由Bellman公式，这两者可以互推），则可以确定最优策略π^*，如对给定s，$\pi^*(s) = \operatorname{argmin}_a q_{\pi^*}(s,a)$。注意，$q_*(s,a)$的确定主要是按照定义，由所有可行策略根据偏序关系得来，即$q_*(s,a) = \max_\pi q_\pi(s,a)$。因为在此之前，我们不知道最优策略，否则会出现鸡生蛋、蛋生鸡问题。

  When we have $v_*(s)$ and/or $q_{\pi^*}(s,a)$ (actually, according to Bellman equation, one can obtain one from the other), we can decide on the optimal policy. For example, for a given state s, $\pi^*(s) = \operatorname{argmin}_a q_{\pi^*}(s,a)$. Notably, $q_*(s,a)$ is obtained according to the definition, aka, in accordance with the partial, so $q_*(s,a) = \max_\pi q_\pi(s,a)$. Actually we don't know the optimal policy before hand, otherwise it gives rise to the chicken or the egg problem.

- 实际中，通常经过策略改进来达到最优策略。如对策略π与π'而言，如果对某个s而言，有$q_\pi\big(s,\pi'(s)\big) \geq v_\pi(s)$，则可构造如下$\pi''$，显然$\pi'' \geq \pi$。

  In practice, the optimal policy is attained via policy improvement. Suppose we have two policies, π and π', and for a given state s, we have $q_\pi\big(s,\pi'(s)\big) \geq v_\pi(s)$, then we can construct a third policy π'' in the following, obviously we have $\pi'' \geq \pi$.

$$\pi''(s) = \begin{cases} \pi'(s) & s = s \\ \pi(s) & s \neq s \end{cases}$$

概念
Concepts

- 实际上，强化学习需要代理与环境达到足够多次的交互才能完成目标。如果对于环境能够建模，则可以借助计算机，进行快速交互实验。因此强化学习一般分为基于模型的方法和无模型方法。基于模型方法需要对环境行为或环境演进进行模拟，从而推断环境的演进方式。例如，给定状态和动作，模型能够预测在此前提下的下一个状态及对应的奖励。模型并非必须的，使用规划的方法来解决强化学习问题才需要模型。而在解决强化学习问题时使用试错的方式解决时，一般基于无模型方法。

  In fact, RL requires sufficient interactions with the environment to fulfill the task. However, if the environment can be modelled, the experiment can be accelerated much faster by harnessing the computer. So RL is categorized into model-based methods and model-free methods. Model-based methods require to mimic the behaviour or evolution of the environment, or more generally, that allows inferences to be made about how the environment will behave. For example, given a state and action, the model can predict the resultant next state and the corresponding reward. Models are not always needed, they are mostly used for the planning method. Methods for solving RL problems with explicit trial-and-error learners are called model-free methods.