APRIL 2021, 伍元凯

# 线性系统、LQR 和强化学习

## 1 线性系统的稳定性、可控性和可观性

### 1.1 稳定性

对于一个线性系统，有

$$x' = Ax(t) + Bu(t) \tag{1}$$

$u(t)$ 是控制输入，$x(t)$ 是状态变量，如果控制输入始终为 0，我们有

$$\frac{dx(t)}{dt} = Ax(t) \tag{2}$$

进而得到

$$\frac{dx(t)}{x(t)} = Adt \tag{3}$$

对等式两边求积分，我们有

$$x(t) = e^{At+C} = e^{A(t-t_0)}x(t_0) \tag{4}$$

对于这个系统，可以利用 $A$ 的特征值进行稳定性分析

$$AT = TD, \quad x' = Tz' = Ax = ATz, \quad z' = T^{-1}ATz = Dz \tag{5}$$

$D$ 是一个对角矩阵，其对角线上的值是 $A$ 的奇异值，$T$ 是 $A$ 特征向量，$x$ 通过特征向量被映射到 $z$。所以我们有

$$z(t) = e^{D(t-t_0)}z(t_0), \quad x(t) = Te^{D(t-t_0)}T^{-1}x(t_0) \tag{6}$$

假设 $D$ 对角线上有元素的实部大于 0，这个系统的状态 $z(t)$ 在 $t$ 趋向于无穷大的时候会趋向无穷大，这个系统会变得不稳定。

### 1.2 可控性

考虑图1中的控制系统，我们有 $y = Cx$, $y$ 为控制器的反馈信号，$u = -Kx$ 为控制信号，对于这个系统，我们有 $x' = (A - BK)x$。显然我们可以通过调节 $K$ 使得这个系统变的稳定（$A - BK$ 的特征值全小于等于 0）。那么什么是

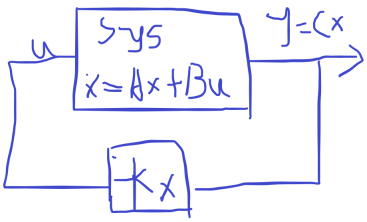


Figure 1: 控制系统

可控制性呢？就是我们可以通过 $K$ 可以把 $x$ 控制到任何状态。这并非是一直可以做到的，如以下这个系统

$$\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = \begin{pmatrix} 1,0 \\ 0,2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \tag{7}$$

因为 $u$ 无法控制 $x_1$，所以这个系统不可控，但如果系统变为

$$\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = \begin{pmatrix} 1,1 \\ 0,2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \tag{8}$$

系统将变得可控，因为我们可以间接的用 $u$ 控制 $x_2$ 来影响 $x_1$。那么如何来分析系统是否可控呢?

$A$ 可以理解为系统状态变量 $x$ 之间的互相影响关系，而 $B$ 可以理解为控制量 $u$ 直接作用于 $x$ 上的影响，那么 $AB$ 就是 $u$ 通过 $A$ 间接给 $x$ 的二次影响, 假设 $x \in \mathrm{R}^n$，我们有可控制性矩阵

$$Co = [B, AB, A^2B, \cdots, A^{n-1}B] \tag{9}$$

这个矩阵表示 $B$ 通过 $A$ 对状态 $x$ 的间接作用，如果这些间接作用可以作用到 $x$ 中的所有变量，我们就认为系统是可控的。因此我们希望 $Co$ 的秩为 $n$，如果秩小于 $n$，就会存在 $Co$ 的某些行为 0 的情况，这就说明通过 $B$ 怎么都没有办法作用到某些 $x$，系统就会变得不可控。所以通过考察 $rank(Co)$ 就可以知道该系统是否可控。如对于公式(7)和公式(8)它们的 $Co$ 分别等于

$$Co_7 = \begin{pmatrix} 0,0 \\ 1,2 \end{pmatrix}, \quad Co_8 = \begin{pmatrix} 0,1 \\ 1,2 \end{pmatrix}, \tag{10}$$

显然公式(8)中的系统可控，公式(7)中的系统不可控。

间接影响的思考方式可以让我们对 $C$ 的秩为 $n$ 和系统可控的关系提供一定的 intuition。但是数学上如何证明系统可控和 $C$ 的秩的关系呢？在公式(1)两侧乘以 $e^{-At}$，我们有

$$e^{-At}x'(t) - e^{-At}Ax(t) = e^{-At}Bu(t) \tag{11}$$

根据求导的性质有

$$e^{-At}x'(t) - e^{-At}Ax(t) = \frac{d(e^{-At}x(t))}{dt} = e^{-At}Bu(t) \tag{12}$$

等式两侧对 $t$ 求积分在乘以 $e^{At}$

$$x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau \tag{13}$$

利用 $e^{A(t-\tau)} = \sum_{k=0}^{\infty} \frac{A^k(t-\tau)^k}{k!}$，可得

$$x(t) = \sum_{k=0}^{\infty} A^k B\hat{u}_k, \quad \hat{u}_k = \int_{t_0}^{t} \frac{(t-\tau)^k}{k!}u(\tau)d\tau \tag{14}$$

可控意味着改变 $\hat{u}$ 在时间 $[t_0, t]$ 会最终作用到 $x(t) \in \mathrm{R}^n$ 中的所有变量, 而 $x(t)$ 可看做各个矩阵 $B, AB, A^2B, \cdots, A^kB$ 的线性组合，各个矩阵的所有列向量如果能够张开成 $\mathrm{R}^n$ 的空间，那么 $\hat{u}_k$ 最终可以作用到 $x(t)$ 中所有的状态之上。

在这个系统中 $A \in \mathrm{R}^{n\times n}$,根据凯莱–哈密顿定理可知更高次幂的 $A^nB, A^{n+1}B, \cdots$ 都可以由低次幂 $A^0, A^1, \cdots, A^{n-1}$ 的线性组合得到。所以我们只需要考察 $Co = [B, AB, A^2B, \cdots, A^{n-1}B]$ 的秩即可。

同理，如果将可观测性视为通过 $y = Cx$ 这一测量方式能够在 $[t_0, t]$ 中观测到状态 $x$ 的所有变化，我们就认为系统是可观测的，那么可观测性也可以通过 $Ob = [C, AC, A^2C, \cdots, A^{n-1}C]$ 的秩是否为 $n$ 进行判断。

## 1.3 卡尔曼的贡献

可控性和可观性可以说是最优控制理论最基本的理论，这两大性质以及对线性动态系统的分析方式可以说都源于卡尔曼的研究。二战之后，美国的很多公司都设立科学研究中心开展与军事科技相关的研究，卡尔曼加盟了格伦 ·

L · 马丁公司（洛克希德马丁的前身）进行数学研究，在此期间，他发表了卡尔曼滤波以及可控性和可观性 [1] 等重要的理论成果。可以说，可控性和可观性给动态系统提供了基本的分析框架，通过这两种工具可以很快的知道设计的控制系统是否是可行的，解决了缺乏理论工程上只能依靠盲目的试验的问题。到了今天，无数的学者仍然在多智能体以及复杂网络等新问题上拓展卡尔曼的研究。

我时常感叹于如卡尔曼这些杰出科学家的灵感，我们无从得知他是如何灵光咋现出这样美妙的特性，并使用数学的方法对其性质进行精确的分析。想必也是站在巨人的肩膀上、惬意的学术环境以及绝顶的聪明才智共同作用的结果。此外卡尔曼也必然有着常人所没有的问题意识，思考着常人想不到的“大”问题。他曾在一篇论文中提到 [2]"虽然新的问题不断浮现，但我们对控制理论的基础的理解仍然非常肤浅，目前唯一在基础理论进展只有香农的信息论"，“我们需要什么样的信息以及多少信息才能实现有效控制？怎样才能刻画控制系统的性质？”这些可以说都是在前人肩膀上的大问题。

卡尔曼和华人科学家何毓琦先生也颇有渊源，何是哈佛教授，美籍华人数学家、控制理论家，说起来，著名的卡尔曼滤波“Kalman Filter”也是何毓琦先生首先建议命名的呢，那时他正好在格伦 · L · 马丁公司访问卡尔曼。何毓琦教授为民国何竞武将军之子，民国时期的权贵子女成才之人是非常之多的。

## 2 一个线性系统的例子

倒立摆 (图2) 是一个典型的线性系统，在这个系统中，我们定义系统的状态 $y$ 为

$$y = \begin{pmatrix} x \\ x' \\ \theta \\ \theta' \end{pmatrix} \tag{15}$$

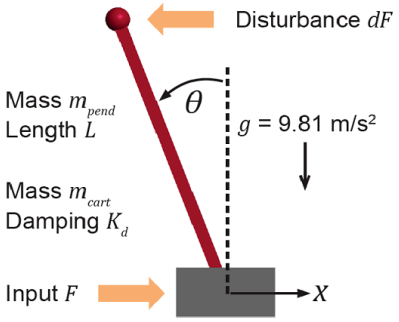


Figure 2: 倒立摆

利用 $y' = Ay + Bu$，有

$$A = \begin{bmatrix} 0, & 1, & 0, & 0 \\ 0, & -d/M, & -mg/M, & 0 \\ 0, & 0, & 0, & 1 \\ 0, & -sd/(ML), & -s(m+M)g/(ML), & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1/M \\ 0 \\ s/(ML) \end{bmatrix} \tag{16}$$

其中 $m$ 是小球的质量，$M$ 是车的质量，$L$ 是摆的长度，$g$ 是加速度，$d$ 是震荡系数，$s$ 是倒立摆位置的参数，$s = 1$ 在上方，$s = -1$ 位于下方。具体的实例见 [3]。对该系统可以进行稳定性分析。

```
import numpy as np
m = 1
```

```
M = 5
L = 2
g = -10
d = 1
s = 1
A = np.zeros([4, 4])
A[0, 1] = 1
A[1, 1] = -d/M
A[1, 2] = -m*g/M
A[2, 3] = 1
A[3, 1] = -s*d/(M*L)
A[3, 2] = -s*(m+M)*g/(M*L)
B = np.array([[0, 1/M, 0, s*1/(M*L)]])
h = np.linalg.eig(A)
h[0]
Out[7]: array([ 0.        , -2.46742895, -0.16651192,  2.43394087])
```

系统存在大于 0 的特征值，并不稳定。系统是否可控:

```
Co = np.concatenate([B.T, np.matmul(A, B.T), np.matmul(A, np.matmul(A, B.T)),
    np.matmul(A, np.matmul(A, np.matmul(A, B.T)))], axis =1)
rank = np.linalg.matrix_rank(Co)
rank
Out[19]: 4
```

$Co$ 的秩为 4，这是一个可控的系统。进而利用 $y' = (A - BK)y$，通过调整 $K$，使得 $A - BK$ 的奇异值小于 0，我们就可以获得一个稳定的系统。

# 3 LQR 控制方法

通过分析可以知道倒立摆是一个可控的系统，但是控制不是免费的，我们对控制可能有各种要求，比如希望对倒立摆系统的控制力不用太大，需要省点电，不希望倒立摆移动太远。对于这样一个问题，我们可以利用 LQR（线性二次调节器）进行求解，对于倒立摆系统，定义控制成本为

$$J = \int_0^\infty \left(y^T Q y + u^T R u\right) dt \tag{17}$$

我们的目标就是找到最小化 $J$ 的 $u = -Kx$。假如我们只希望 $\theta$,$\theta'$ 保持不变，以及尽量使用更小的力，可以定义 $Q$ 和 $R$ 为

$$Q = \begin{pmatrix} 1, & 0, & 0, & 0 \\ 0, & 1, & 0, & 0 \\ 0, & 0, & 100, & 0 \\ 0, & 0, & 0, & 100 \end{pmatrix}, R = 100 \tag{18}$$

$Q$ 和 $R$ 的设计实际就是给不同控制目标一些权重。由于采用二次函数 ($Q = Q^T \geq 0, R = R^T > 0$)，这个成本函数的最小化过程是一个凸优化的过程，我们总是能够找到使得 $J$ 最小的 $K$。

求解 LQR 控制方法，我们需要用到 HJB 方程，即以三位数学家的名字命名的哈密顿-雅可比-贝尔曼方程，贝尔曼的工作可以说是模型预测控制和强化学习的基础，以他名字命名的 Richard E. Bellman Control Heritage Award 可

以说是控制论领域的最高奖项。哈密顿-雅可比-贝尔曼方程处理这一优化问题的方式可以说是非常高明。假设有如下问题：

$$J = \int_0^{\infty} (f(x(t), u(t), t))\, dt, \quad x'(t) = a(x(t), u(t), t). \tag{19}$$

假设 $V(x(t), t)$ 代表从指定时间 $t$、指定状态 $x$ 出发，到 $t$ 后续最小惩罚的表达式。有

$$V(x(t), t) = min_u[(\int_t^{t+dt} (f(x(t), u(t), t))\, dt) + V(x(t + dt), t + dt)] \tag{20}$$

这里的 $V(x(t + dt), t + dt)$ 是从 $t + dt$ 出发后续的最优损失, 积分项是 $[t, t + dt]$ 之间的最优损失，我们对在 $t$ 这一瞬间采取的最优策略比较感兴趣，所以 $dt$ 趋向于 0。对 $V(x(t + dt), t + dt)$ 进行泰勒展开

$$V(x(t + dt), t + dt) = V(x(t), t) + \bigtriangledown_t V(x(t), t)dt + \bigtriangledown_x V(x(t), t)x(t)dt + o(dt) \tag{21}$$

将公式(20)和(21)相减，再除以 $dt$，有

$$0 = \bigtriangledown_t V(x(t), t) + \min_u[\bigtriangledown_x V(x(t), t)x(t) + f(x(t), u(t), t)] \tag{22}$$

将 $x(t) = a(x(t), u(t), t)$ 代入得到

$$0 = \bigtriangledown_t V(x(t), t) + \min_u[\bigtriangledown_x V(x(t), t)a(x(t), u(t), t) + f(x(t), u(t), t)], \tag{23}$$

$\bigtriangledown_x V(x(t), t)a(x(t), u(t), t) + f(x(t), u(t), t)$ 就是哈密顿量，我们可以哈密顿量关于 $u(t)$ 的梯度，继而求出能够最小化哈密顿量的 $u(t)$ ，得到它关于 $x(t)$ 的表达式。而解出 $u(t)$ 关于 $x(t)$ 和 $V(x, t)$ 的方程之后再代入公式(23)就可以得到最优的 $u(t)$ 对应于 $x(t)$ 的解。

这个方法也就是动态规划 (Dynamic Programming) 的连续版。关于 Dynamic Programming 的命名也非常有意思，当时贝尔曼任职于兰德公司，兰德公司受雇于美国空军，当时的空军负责人并不太支持纯粹的数学研究，贝尔曼选择了 (dynamic) 这一词来表示其解决的是时变的问题，让项目投资者印象深刻，Programming 主要是为了迎合军事上训练和后勤等等的需要（这些应用当时都使用 programming），这样他就可以在军方支持下进行自己的数学研究了。而最早发现贝尔曼方程与经典物理中的哈密顿-雅可比方程关联的正是卡尔曼 [4]。

回到 LQR 问题，利用 HJB 方程有哈密顿量

$$\min_u[\bigtriangledown_x V(Ax + Bu) + x^T Qx + u^T Ru] \tag{24}$$

在线性系统中，$\bigtriangledown V = 2x^T S$，得到最优的 $u^* = -R^{-1}B^T Sx$。其中的 $S$ 是未知量，代入到 HJB 方程有 $0 = SA + A^T S - SBR^{-1}B^T S + Q$（LQR 的 $V(x, t)$ 对 $t$ 的导数为 0），解出 $S$ 关于 $A, B, R, Q$ 的表达式后，就可以得到每个状态 $s$ 下对应的最优状态 $u$。这个公式也被称为代数黎卡提方程。

**小结:** 稳定性、可控性这些优美的性质所针对的都是线性动力系统，LQR 控制方法也是对线性动力系统求解的方法。即系统需满足 $x' = Ax(t) + Bu(t), u(t) = -Kx(t)$，有人会说世界本质上是非线性的，不存在严格符合线性系统规律的系统。但线性系统的数学性质也是复杂世界涌现出的一种模式，譬如说文中的倒立摆系统。数学家们正是受到自然界的这些模式的启发利用数学定义出了线性系统，然后推导衍生出了这种系统的最优解和各种各样的性质，线性系统上的性质进而也可以衍生到更复杂的非线性系统。可以说简洁的数学将卡尔曼脑中对可控性和可观测性的理解公理化，后人们学习数学阅读这些定义，就可以精确化的了解到卡尔曼的思想，这就是数学的魅力所在，它以最简洁最精确的方式将哲人的思想保留下来，传播开去，它同样可以依靠逻辑和公式推演，使得我们能在巨人的肩膀上研究更为复杂的系统。

# 4 强化学习和最优控制理论的关系

对于最优控制理论的现实应用，最为麻烦的过程当属如何将手头上的系统通过数学推导和近似构建为一个可分析的系统，然后再利用数学工具对其进行求解。而今天的深度强化学习直接解决的则是超高维状态观测（图像）下的高度非线性系统（围棋）的控制决策问题。对于过于复杂的控制决策问题，要将其完全数学化公理化是非常困难的，利用深度强化学习来解决这一问题实际上利用的是所谓神经网络的万能逼近能力。当然由于深度强化学习并没有严格的稳定性、可控性等等的证明，所以也受到了一些传统控制领域的研究者的质疑 [5]。

强化学习并非毫无理论根基可言，从最基础的角度，首先其解决的马尔科夫决策过程（MDP）的问题。即

$$P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a) \tag{25}$$

$t+1$ 时刻的状态 $s'$ 只由 $t$ 时刻的状态 $s$ 和动作 $a$ 决定。马尔科夫决策过程这一概念是 1960 年代由斯坦福经济工程系的教授 Ronald A. Howard 在其著作《Dynamic programming and Markov processes》[6] 中所系统介绍的。当时的西尔斯 · 罗巴克公司（1980 年之前美国最大的零售商）研究应该如何给用户发邮件吸引用户购买商品，Howard 当时正在 MIT 读博士，为了解决这一问题，构建出了马尔科夫决策过程，在它的书中叫做有奖励的马尔科夫过程，奖励就是公司的收益减去寄邮件的成本，而状态则定义为用户的购物历史 [7]。用户的购买行为本身就非常复杂，马尔科夫决策过程的源头本身就是为了解决这种无法精确描述的系统行为，所以马尔科夫决策过程中看待系统的状态转移通过的是一种充满不确定性的一种统计建模的视角，这就和很多工业系统并不一致，工业系统中工程师为了保证系统的鲁棒性，本身就将系统中的不确定性限定到很低，我们也往往有确定性的公式去描述这些系统的动力学模型。所以从根源上来看，强化学习相较于控制理论可能更适合宏观的决策。

强化学习的第二大基础是基于贝尔曼公式 (也就是贝尔曼最优原理)，值得一提的是 Howard 认为自己的工作是间接受到贝尔曼影响的，虽然他当时并不知道贝尔曼在研究动态规划，当他认为自己当时的导师是知道贝尔曼的工作的，并在指导过程中间接影响了他。贝尔曼最优原理的表述为：一个过程的最优策略具有这样的性质，即无论其初始状态及初始决策如何，其以后诸决策对以第一个决策所形成的状态作为初始状态的过程而言，必须构成最优策略。这个原理的实质即不管过去的过程如何，只从当前的状态和系统的最优化要求出发，作出下一步的最优决策 [8]。贝尔曼将这一思想用数学的方式定义出来就是

$$V(x) = \max_{a \in \Gamma(x)} \{F(x, a) + \beta V(T(x, a))\}. \tag{26}$$

$F(x, a)$ 是给定状态 $x$ 和动作 $a$ 时的奖励，而 $T(x, a)$ 则为状态转移方程，$V(x)$ 是最优的状态价值函数。贝尔曼公式将整体最优的问题分解为若干子问题，使得其变得可解，而其公式的连续版本 HJB 公式在上文中也已经介绍。不管是目前深度强化学习中使用的动作价值函数还是状态价值函数都是基于贝尔曼最优原则得到的。贝尔曼的最优原则会给人带来一定的困惑，如图3所示，如果不考虑之前的策略，路径 [1,2,3,4] 和 [1,2,4] 都是最优的路径，但如果使用贝尔曼的 DP 算法可以轻松解出 [1,2,4] 才是最优的路径。其实这一困惑源于贝尔曼的语言中忽略掉了它寻找最优

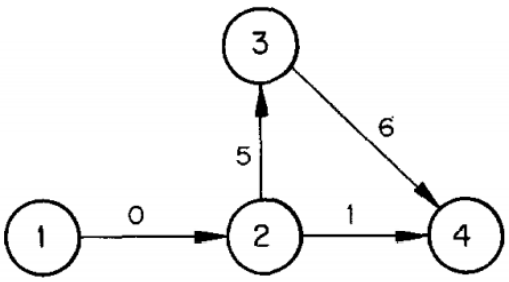


Figure 3: 一个例子，根据贝尔曼最优原理，路径 [1,2,3,4] 和 [1,2,4] 都为最优路径。

的过程中基于的最初策略本身也需要是最优的假设，价值函数 $V(1)$ 和 $V(2)$ 显然等于 $-1$（距离的负数），所以在 2 节点应该选择的是直接到达 4。可见将思想数学化是最精确的表达，也是最易于避免歧义的表达。

在贝尔曼等式中有一个瞬时的收益 $F(x,a)$，也就是强化学习中奖励函数。大多的深度强化学习方法可以说都是利用环境提供的海量奖励进行训练的，这就是基于时序差分方法 (Temporal difference)。TD 算法是现代强化学习之父 Sutton 教授受到 Arthur Samuel 的人工智能国际象棋软件的启发而发明的，Sumuel1949 年从 MIT 毕业之后集中于研究如何让计算机玩游戏，因为他认为这会对使用计算机解决普遍的问题有好处。他设计了一种算法，该算法并不直接搜索哪一种路径能够最终导向胜利，他给每一个落子位置都给予分数，计算机根据这一分数来选择最优的落子位置 [9]。这样一种思路在今天的 AlphaGO 仍然被沿用，只不过给予分数的函数变成了深度极深的神经网络。

Sutton 正是基于 Samuel 的工作，提出了 TD 算法

$$V(s) \leftarrow V(s) + \alpha(\overbrace{r + \gamma V(s')}^{\text{The TD target}} - V(s)) \tag{27}$$

$s$ 和 $s'$ 分别是上一个时刻和当前时刻的状态，$r + (s')$ 被称为 TD 目标函数。在现实的大多算法中我们一般利用动作价值函数而非状态价值函数对网络，有

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{b \in a} Q(s',b) - Q(s,a)], \tag{28}$$

其中 $a$ 是采取的动作，这表明每一步只有 $Q(s,a)$ 是被更新的。下面给出 TD 算法收敛性的一个证明。

将公式(28)写作

$$Q_{t+1}(s_t,a_t) = (1-\alpha)Q_t(s_t,a_t) + \alpha[r + \gamma \max_{b \in a} Q(s_{t+1},b)] \tag{29}$$

将 $Q*(s_t,a_t)$ 定义为最优 Q 函数，定义

$$\triangle_t(s_t,a_t) = Q_t s_t, a_t - Q^*(s_t,a_t) \tag{30}$$

等式两侧都减去 $Q^*(s_t,a_t)$ 得到

$$\triangle_t(s_t,a_t) = (1-\alpha)\triangle_t(s_t,a_t) + \alpha[r + \gamma \max_{b \in a} Q(s_{t+1},b) - Q^*(s_t,a_t)]. \tag{31}$$

用 $F_t s, a$ 表示 $r(s,a,X(s,a)) + \gamma \max_{b \in a} Q(s_{t+1},b) - Q^*(s_t,a_t)$，等式变为

$$\triangle_t(s_t,a_t) = (1-\alpha)\triangle_t(s_t,a_t) + \alpha F_t(s,a), \tag{32}$$

如果上述公式中的 $\triangle_t(s_t,a_t)$ 收敛到 0，显然 $Q_t s_t, a_t$ 最终会收敛到最优的 $Q^*(s_t,a_t)$。

要证明 $\triangle_t(s_t,a_t)$ 收敛到 0 需要套用一个随机过程收敛性定理，这个定理数学上证明可见于 [10]。

**Theorem .1** *对于定义在 $\mathrm{R}^n$ 随机过程序列 $\{\triangle_t\}$,$\triangle_t(x) = (1-\alpha)\triangle_t(x) + \alpha F_t(x)$ 最终会收敛到 0，如果满足以下条件：*

1. $\sum_t \alpha = \infty, \sum_t \alpha^2 < \infty$;
2. $\|E(F_t(x)|\mathcal{F}_t)\|_W \leq \gamma \| \triangle_t \|_W$, 其中 $\gamma < 1$;
3. $var[F_t(x)|\mathcal{F}_t] \leq C(1 + \| \triangle_t(x)\|_W^2)$，$C \geq 0$

显然我们只需要判断我们针对 Q 学习定义的 $\triangle_t(x)$ 是否符合以上的条件就可以判断 Q 学习是否收敛了。首先 $0 < \alpha < 1$，条件 1 符合。

下面考察条件 2，

$$E(F_t(x)|\mathcal{F}_t) = \sum_{y \in X} P_a(x,y)[r(a,x,y) + \gamma \max_{b \in a} Q(y,b) - Q^*(x,a)]] = (\mathcal{H}Q_t)(x,a) - (\mathcal{H}Q^*)(x,a) \tag{33}$$

我们将 $(\mathcal{H}Q_t)(x,a)$ 定义为 $\sum_{y \in X} P_a(x,y)[r(a,x,y) + \gamma \max_{b \in a} Q(y,b)$, $Q^*(x,a)$ 是收敛后的最优动作价值函数，所以显然有 $Q^*(x,a) = \mathcal{H}Q^*)(x,a)$。考察 $(\mathcal{H}Q_t)(x,a) - (\mathcal{H}Q^*)(x,a)$ 有 $(\|\mathcal{H}Q_t)(x,a) - (\mathcal{H}Q^*)(x,a)\|_\infty \leq \gamma \|Q_t - Q^*\|_\infty = \gamma \| \triangle_t \|_\infty$，显然条件 2 也满足。

再看条件 3，利用 $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$ 以及 $E[Q^*(x,a)|\mathcal{F}_t] = Q^*(x,a)$，有

$$var[F_t(x)|\mathcal{F}_t] = E[(r(x,a,X(x,a)) + \gamma \max_{b\in a} Q(y,b) - (\mathcal{H}Q_t)(x,a))^2] = var[r(x,a,X(x,a)) + \gamma \max_{b\in a} Q(y,b)|\mathcal{F}_t] \quad (34)$$

唯一贡献方差的是 $r(x,a,X(x,a))$，而且 $r$ 有界，显然 $var[F_t(x)|\mathcal{F}_t] \leq C(1 + \| \triangle_t(x)\|_W^2)$。

上述证明只是证明了收敛性 ($t \to \infty$)，我们并不知道基于 TD 的算法到底多快可以收敛，现实之中，深度强化学习往往需要大量的采样才可以收敛，这也是限制深度强化学习现实工业应用的一大问题，因为需要一个非常逼真的仿真器来生成大量的样本，然而这些仿真样本和现实样本之间会存在差异性，对深度强化学习的实际工业应用有所限制。

和控制理论相比，强化学习并非缺乏理论，它针对的是更为普适的强人工智能问题，它将环境建模为一种未知且充满不确定性的状态转移方程，这和控制论中将环境试图还原为一种可数学分析的动态系统的方式有所不同，两者在二战后的五六十年代共享相同的基础，但逐渐渐行渐远。对于实际问题，复杂的高层级决策，深度强化学习是不错的选择，而底层的控制，目前还需要依赖可靠的，可精确分析的控制论方法。

一些值得思考的问题：在深度强化学习对状态变量建模时是否可以引入控制论中的可观测性原理进行分析？是否可以对构建的马尔科夫过程进行可控性分析？

# 5 参考文献

[1] Kalman, R. E. (1963). Mathematical description of linear dynamical systems. Journal of the Society for Industrial and Applied Mathematics, Series A: Control, 1(2), 152-192.

[2] Kalman, R. E. (1960, August). On the general theory of control systems. In Proceedings First International Conference on Automatic Control, Moscow, USSR (pp. 481-492).

[3] Steve Brunton `https://www.youtube.com/channel/UCm5mt-A4w61lknZ9lCsZtBw`

[4] Kalman, R. E. (2021). 16. The Theory of Optimal Control and the Calculus of Variations (pp. 309-332). University of California Press.

[5] `https://www.zhihu.com/question/360460405?utm_source=wechat_timeline`

[6] Howard, R. A. (1960). Dynamic programming and markov processes.

[7] Howard, R. A. (2002). Comments on the origin and application of Markov decision processes. Operations Research, 50(1), 100-102.

[8] Bellman, R. (1966). Dynamic programming. Science, 153(3731), 34-37.

[9] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

[10] Jaakkola, T., Jordan, M. I., Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. Neural computation, 6(6), 1185-1201.