代入H-J-B方程，求解关于二元函数$V(x,t)$的偏微分方程：

$$\begin{cases} \dot{V}(x,t) + min_u[\nabla_x V(x,t) f(x, W(V(x,t),x)) + C(x, W(V(x,t),x))] = 0 \\ V(x,T) = D(T) \end{cases}$$

4.4.H-J-B方程*

heaven

智能制造创业 强化学习算法 魔幻巨著创作中

关注他

123 人赞同了该文章

在前面，我们所讨论的都是时间离散的MDP。但是，在传统的最优控制问题中，人们更习惯于讨论时间连续的MDP；本书中对“最优控制”的定义是求解环境已知的MDP。不过，一般语境下的最优控制更侧重于研究动作、状态与时间都连续的MDP。我们将用两章来讨论时间为连续变量的最优控制问题，分别讲解基于价值的方法与基于策略的方法。

需要提前声明的是，接下来的两章H-J-B方程与变分原理，由于考虑的是连续时间的问题，分别需要用到偏微分方程与变分法（常微分方程），需要一定的数学基础。如果没有学过有关的课程，会学起来非常吃力。事实上，这两章的内容能为强化学习提供一些思路，但是和后面强化学习的具体算法关系并不大（因为强化学习研究时间是离散的情况）。因此，这两章我们打上星号，作为补充内容。如果读者看着吃力可以直接跳过，这不会影响后面关于强化学习的学习。

不过另一方面，虽然本章的H-J-B方程并一定需要读者掌握，但是本章一定程度上相当于最优控制中基于价值的算法的一个总结。本章第四节总结了所有和价值有关的方程与公式，读者可以直接前往观看。

下面，让我们先来明确我们要讨论的问题是什么：

1、时间连续的最优控制问题

首先，我们的状态与控制函数不再是序列 $\boldsymbol{x_t}$ 与 $\boldsymbol{u_t}$ 的形式，而是连续函数$\boldsymbol{x_t}$与$\boldsymbol{u_t}$的形式。其次，$\boldsymbol{t}$ 时刻的损失也不再是 $\boldsymbol{C(x_t,u_t,t)}$ 的形式，而应该是 $\boldsymbol{C(x(t),u(t),t)}$ 的形式，也可以将其简记为 $\boldsymbol{C(x,u,t)}$ 。为了简单起见，我们考虑损失函数时齐的情况，即损失为 $\boldsymbol{C(x(t),u(t))}$ 或 $\boldsymbol{C(x,u)}$ 。由于$\boldsymbol{t}$时刻只是一个瞬间，所以瞬间内的损失应该是一个微元。相应地，总损失 $\boldsymbol{J}$ 不再是各个$\boldsymbol{t}$时刻损失的加和，而是各个$\boldsymbol{t}$时刻损失微元的积分。

另外，我们仍然会假定MDP有起始与终止时间 $\boldsymbol{0}$ 与 $\boldsymbol{T}$ ，这意味着我们面对的问题始终是非时齐的。在现实问题中，如果终止时间 $\boldsymbol{T}$ 是有意义的，意味着最后时刻到达的状态往往会比较重要（任务的完成度）。所以，我们不能认为 $\boldsymbol{T}$ 时刻的损失只有一个微元，而应该专门设定一个衡量终止时间损失的函数，突出其重要性。

因此，我们要优化的目标 J 有如下公式：

$$J = \int_0^T C(x(t), u(t))dt + D(x(T))$$

还有一个最重要的问题是，在时间为连续变量的情况下，状态转移关系应该是什么？根据MDP的马氏性，我们必须假定状态在从 t 时刻到 $t + dt$ 时刻内的变化 $\dot{x}(t)dt$ 只和 $x(t), u(t)$ 与t相关，具有 $f(x(t), u(t), t)$ 的形式。为简单起见，我们进一步假定环境也是时齐的，则我们的状态转移关系应该具有如下形式：

$$\dot{x}(t) = f(x(t), u(t))$$

上面给出的是一个确定性的环境。如果时间是连续的，并且环境具有随机性，就要用到随机微分方程相关的内容。本书面向的是具有一定数学基础的同学，微分方程属于理科专业的必修课程，而随机微分方程却不是必修课程，可能只有部分概率或金融相关专业的同学对其有所了解。所以在本章中，我们只讨论环境具有随机性的情况。

综合以上的几点，（时齐的）连续时间的最优控制问题具有如下形式：

给定初始状态： $x(0) = x_0$

环境： $\dot{x}(t) = f(x(t), u(t))$

目标：极小化 $J = \int_0^T C(x(t), u(t))dt + D(x(T))$

上述的问题就是我们在本章中要集中讨论的问题。可以看出，它具有确定、非时齐的环境（因为有终止时间 T ），所以其解出的最优控制具有 $u^* = policy(t)$ 的形式。这和LQR问题的结果类似，但不同在于 t 是连续的，所以我们要求解的不是最优控制序列，而是最优控制函数 $u^*(t)$ 。

对于时间连续的情况，我们依然有LQR问题，即 f 是线性函数、 C, D 为正定二次函数的问题。为了简单起见，我们只考虑时齐的情况，并且胜率交叉项、一次项与常数项。则连续时间LQR问题为：

给定初始状态： $x(0) = x_0$

环境： $\dot{x}(t) = Ax(t) + Bu(t)$

目标：极小化 $J = \int_0^T [x(t)'Qx(t) + u(t)'Ru(t)]dt + x(T)'Dx(T)$

下面让我们介绍基于价值的求解方法——H-J-B方程。

2、H-J-B方程

我们在上一章中说过，基于价值的方法本质上就是要将每一步对于最终结果的影响给强行分离出来，使得我们可以单独对每一步决策求解最优。但是，在时间是连续变量的情况下，没有"一步"这种提法，所以我们要将考虑的是对时间的微元 dt 进行分离。

在时间离散的MDP中，价值 $V(x,t)$ 的定义是“在 t 时刻处于 x 状态，后续按照最佳策略走，能获得的最小损失”。在连续时间的MDP中，我们可以按照完全一样的方式定义价值函数 $V(x(t),t)=min_u[\int_t^T C(x(t),u(t))dt+D(x(T))]$。这里的 $V(x(t),t)$ 与时间离散情况下的 $V(x_t,t)$ 的含义是完全一样的，唯一的区别只在于t变成了连续变量。

定义了价值函数后，我们不难发现有恒等式 $V(x(T),T)=D(x(T))$。由于 D 函数表达式已知，所以我们就知道了 $V(x,T)$ 对于所有 x 的取值。细心的读者可能已经回想起来，上一章讲LQR问题时我们也是率先知道 $V(x,T)$ 对于所有x的取值，二者在逻辑上是一样的。

在上一章时间离散的LQR问题中，我们在知道$V(x,T)$取值之后，还必须找出 $V(x,t)$ 与 $Q(x,u,t)$ 之间满足的方程，才能从 $V(x,T)$ 出发依次推出所有 t 对应的$V(x,T)$。而在时间连续的MDP中，情况也是类似的。但有所不同的是，在时间连续的MDP中任何一个瞬间的控制 $u(t)$ 对于全局的作用只是一个微元。所以无论 t 时刻的$u(t)$取值多少，$Q(x(t),u(t),t)$ 与 $V(x(t),t)$ 的差别只是一个微元。因此，我们就不专门考虑 $Q(x,u,t)$，只考虑如何找出 $V(x,t)$ 函数的内在关系。

具体而言，我们也要设法将有 t时刻对应的价值分解为两部分——这一步立即获得的损失，以及下一步及以后能够获得的损失。由 $V(x(t),t)$ 的定义，我们可以推出公式：

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

对 $V(x(t+dt),t+dt)$ 进行泰勒展开，得：

$$V(x(t+dt),t+dt)=V(x(t),t)+\dot{V}(x(t),t)dt+\nabla_x V(x(t),t)\dot{x}(t)dt+o(dt)$$

将上面两个式子相减再除以 dt，并让 dt 趋于0，得到以下方程：

$$0=\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)\dot{x}(t)+C(x(t),u(t))]$$

将环境 $\dot{x}(t)=f(x(t),u(t))$ 代入，可以得到一个方程。这个方程就是大名鼎鼎的H-J-B方程（Hamilton-Jacobi-Bellman equation）：

$$0=\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)f(x(t),u(t))+C(x(t),u(t))]$$

我们将 $\nabla_x V(x(t),t)f(x(t),u(t))+C(x(t),u(t))$ 称作“哈密顿量”（Hamiltonian），记作 $\mathcal{H}(t)$。我们上面的方程中有一个 $min_u\mathcal{H}$ 的部分，所以我们需要求出 $argmin_u\mathcal{H}$，这也就是我们要求的最优控制 $u^*(t)$。为此，我们可以求出哈密顿量 $\mathcal{H}$ 关于 u 的梯度，将其置于0，即：

$$\nabla_u \mathcal{H} = \nabla_x V(x,t) f'_u(x,u) + C'_u(x,u) = 0$$

由于动力系统 f 与惩罚函数 C 的表达式已知，而 $\nabla_x V(x(t),t)$ 虽然未知，但它里面不含 u 。所以，我们能从上述公式中解出一个 u^* 的表达式，它里面还有 x,t 以及暂时未知的 $V(x,t)$ 。假设 $\mathcal{H}(t)$ 是关于 $u(t)$ 的凸函数，则上面公式的唯一解就是 $argmin_u \mathcal{H}(t)$ ，即 $u^*(t)$ ；若其不是凸的，则我们也可以从上面公式求出 $\mathcal{H}$ 的一些局部最优，并对比分析出最优解到底是哪一个。总的来说，我们可以解出 $u^*(t)$ 有如下表达式（其中 W 是一个已知函数）：

$$u^*(t) = argmin_u \mathcal{H} = W(V(x,t),x)$$

如果将 $u^*(t)$ 的表达式代入上述的方程 $0 = \dot{V}(x(t),t) + min_u \mathcal{H}$ ，就可以得到一个关于二元函数 $V(x,t)$ 的偏微分方程：

方程： $\dot{V}(x,t) + \nabla_x V(x,t) f(x, W(V(x,t),x)) + C(x, W(V(x,t),x)) = 0$

边界条件： $V(x,T) = D(x)$

我们可以根据这个方程解出 $V(x,t)$ 。将其代入 $u^*(t) = W(V(x,t),x)$ ，就可以得到 $u^*(t)$ 的表达式 $K(x,t)$ 。这就相当于我们在时间离散的LQR问题中解出的 $u_t^* = K_t x_t$ 。 由于我们有确定的环境 $\dot{x}(t) = f(x(t),u(t))$ ，所以，我们可以求解微分方程
$\dot{x}(t) = f(x(t),u^*(t)) = f(x(t),K(x,t))$ （带有初始条件 $x(0) = x_0$ ）得到 $x(t)$ 。这样一来，我们就可以求出 $u^*(t)$ ，一个只和时间有关，和状态无关的控制。这就相当于我们在时间离散、环境确定的LQR问题中通过 $u_t^* = K_t x_t$ 最终还是解出了序列 u_t^* 。

在上面，我们讲解了用基于价值思想求解时间连续的MDP的方法。具体步骤有些繁琐，可能让读者有些头晕。我们将其写成一张框图：

给定初始状态：$x(0)=x_0$；

环境：$\dot{x}(t)=f(x(t),u(t))$;

目标：极小化$J=\int_0^T C(x(t),u(t))dt+D(x(T))$;

基于价值的方法

1.定义价值：$V(x(t),t)=min_u[\int_t^T C(x(t),u(t))dt+D(x(T))]$;

2.列出H-J-B方程：$\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)f(x(t),u(t))+C(x(t),u(t))]=0$;

3.求解最优控制表达式：$u^*(t)=argmin_u\mathscr{H}=W(V(x,t),x)$;

4.代入H-J-B方程，求解关于二元函数$V(x,t)$的偏微分方程：

$$\begin{cases}\dot{V}(x,t)+min_u[\nabla_x V(x,t)f(x,W(V(x,t),x))+C(x,W(V(x,t),x))]=0\\ V(x,T)=D(T)\end{cases}$$

5.求出$V(x,t)$后代入$W(V(x,t),x)$，得到$u^*(t)=K(x,t)$的表达式；

6.求解关于一元函数$x(t)$的常微分方程：

$$\begin{cases}\dot{x}(t)=f(x(t),u^*(t))=f(x(t),k(x(t),t))\\ x(0)=x_0\end{cases}$$

7.根据$x(t)$与$K(x,t)$得出$u^*(t)$的最终表达式；

在上述的框图中，我们为了方便读者理解，所以将求解的每一步给列了出来。例如对于H-J-B方程$\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)f(x(t),u(t))+C(x(t),u(t))]=0$，由于它含有一个$min_u$的部分，所以我们在解题中要先求出$u^*(t)=W(V(x,t),t)$的表达式，然后代入H-J-B方程，将其变成关于二元函数$V(x,t)$的偏微分方程并求解。但是在概念上，我们可以直接认为$\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)f(x(t),u(t))+C(x(t),u(t))]=0$就是一个关于二元函数$V(x,t)$的非线性偏微分方程，可以直接求解出$V(x,t)$。这样可以使得我们更加简单地理解算法的主要思路。我们可以进一步将算法的主要思路概括为如下的框图：

给定初始状态：$x(0)=x_0$；

环境：$\dot{x}(t)=f(x(t),u(t))$;

目标：极小化$J=\int_0^T C(x(t),u(t))dt+D(x(T))$;

基于价值方法的基本思想

1.定义价值$V(x,t)$;

2.列出H-J-B方程（$V(x,t)$的内在关系）并结合边界条件$V(x,T)=D(x)$解出$V(x,t)$;

3.根据$u^*=argmin_u\mathscr{H}$解出最优控制表达式$u^*=K(x,t)$;

4.根据最优控制表达式$K(x,t)$与环境公式$\dot{x}(t)=f(x,u),x(0)=x_0$解出$x(t)$;

5.根据$x(t)$与$K(x,t)$得出$u^*(t)$的最终表达式;

下面，让我们尝试用H-J-B方程的思想来求解一个具体的时间连续的问题。我们还是考虑从最简单的问题——LQR问题开始。

3、用H-J-B方程解LQR问题

假设我们面对的LQR问题如下：

给定初始状态：$\boldsymbol{x(0)=x_0}$；

环境：$\boldsymbol{\dot{x}(t)=x(t)+u(t)}$；

目标：极小化 $\boldsymbol{J=\int_0^T \frac{1}{4}u^2(t)dt+\frac{1}{4}x^2(T)}$；

我们可以粗略想象一下这个MDP的含义—— $\boldsymbol{x(t)}$ 代表某样东西的位置， $\boldsymbol{u(t)}$ 代表我们推它的力度。我们需要在 $\boldsymbol{T}$ 时刻将它推到一个比较接近0的地方，越接近0越好（用 $\boldsymbol{x^2(T)}$ 来衡量最终位置的损失）。另外，在我们推动它的过程中每一个瞬间都要耗费与 $\boldsymbol{u^(t)}$ 成正比的能量。这样，我们就定义出一个时间连续的LQR问题。

首先，我们要定义 $\boldsymbol{V(x,t)}$ 并列出它的H-J-B方程：

$$\boldsymbol{\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)(x(t)+u(t))+\frac{1}{4}u^2(t)]=0}$$

我们必须把上述方程中 $\boldsymbol{min_u}$ 的部分消去，才能解出它。为此，我们必须先把能够极小化哈密顿量 $\boldsymbol{\mathcal{H}}$ 的 $\boldsymbol{u^*}$ 表达式给求出来。我们有：

$$\boldsymbol{\mathcal{H}=\nabla_x V(x,t)(x(t)+u(t))+\frac{1}{4}u^2(t)}$$

我们列出 $\boldsymbol{\mathcal{H}}$ 关于 $\boldsymbol{u}$ 的导数为0的方程：

$$\nabla_u \mathcal{H} = \frac{1}{2}u(t) + \nabla_x V(x,t) = 0$$

上述的方程有唯一解 $u^*(t) = -2\nabla_x V(x(t),t)$ 。为了证明其是最优点，我们还要验证二次充分条件 $\nabla_{uu}\mathcal{H}(u^*(t)) = \frac{1}{2} > 0$ 成立。这便可以推出驻点 $u^*(t) = -2\nabla_x V(x(t),t)$ 确实是 $\mathcal{H}$ 的全局极小点。这就相当于我们求出了 $u^* = W(V(x,t),x)$ 表达式。这也就意味着：

$$min_u \mathcal{H} = \nabla_x V(x(t),t)(x(t) - 2\nabla_x V(x(t),t)) + \frac{1}{4}4\nabla_x V(x(t),t)^2$$

为方便起见，我们将 $\nabla V(x(t),t)$ 简记为 V_x ，把 $V(\dot{x(t)},t)$ 简记为 V_t 。然后，将上述 $min_u \mathcal{H}$ 的表达式代入H-J-B方程，可以得到：

方程： $V_t - V_x^2 + V_x x(t) = 0$ ；

边界条件： $V(x,T) = \frac{1}{4}x^2$ ；

解题进行到这里，我们就得到了一个典型的关于二元函数 $V(x,t)$ 的偏微分方程。要解出这样的方程无疑是有难度的。不过在求解pde时，我们时常会采用一些“连蒙带猜”的技巧。由于这个问题本身也是一个LQR问题，而我们之间看到LQR问题中的 $V(x,t)$ 是关于 x 的正定二次函数，所以我们可以猜测其解具有 $V(x(t),t) = \frac{1}{2}Q(t)x^2(t)$ 的形式，其中的 $Q(t)$ 是一个会随时间变化的正定矩阵。则我们可以推出 $V_x = Q(t)x(t)$ 与 $V_t = \frac{1}{2}\dot{Q}(t)x^2(t)$ 。将其代入上面的H-J-B方程，可以得到：

方程： $\frac{1}{2}\dot{Q}(t)x^2(t) - Q^2(t)x^2(t) + Q(t)x^2(t) = 0$ ；

边界条件： $\frac{1}{2}Q(T)x^2(T) = \frac{1}{4}x^2(T)$ ；

我们假定 $Q(t)$ 与 $x(t)$ 都是连续可导函数，而 $x(t)$ 显然不可能处处为0。所以我们可以为方程消去 $x^2(t)$ ，化简可得：

方程： $\frac{1}{2}\dot{Q}(t)?Q^2(t) + Q(t) = 0$

边界条件： $Q(T) = \frac{1}{2}$

要注意我们有如下的积分恒等式：

$$\frac{dQ}{Q(1-Q)} = \frac{dQ}{Q} + \frac{dQ}{1-Q} = d(lnQ) - d(ln(1-Q))$$

因此，我们可以对上述方程进行如下的化简：

$$\frac{1}{2}\dot{Q}(t) - Q^2(t) + Q(t) = 0$$

$$\frac{dQ}{dt} = 2(Q^2 - Q)$$

$$\frac{dQ}{Q^2-Q}=2dt$$

$$d(ln(1-Q))-d(lnQ)=2dt$$

$$ln(\frac{1-Q}{Q})=2t+C$$

$$\frac{1-Q}{Q}=Dexp(2t)$$

$$Q=\frac{1}{1+Dexp(2t)}$$

另一方面，我们根据边界条件可以推出：

$$Q(T)=\frac{1}{1+Dexp(2T)}=\frac{1}{2}$$

因此，我们可以推出 $Dexp(2T)=1$，即 $D=\frac{1}{exp(2T)}$。所以，我们有 $Q(t)=\frac{exp(2T)}{exp(2t)+exp(2T)}$。总的来说，我们通过H-J-B方程解出了 $V(x,t)$ 的表达式：

$$V(x,t)=\frac{1}{2}\frac{exp(2T)}{exp(2t)+exp(2T)}x^2(t)$$

由于我们在前面推出过最优控制关于 V 的表达式 $u^*(t)=-2\nabla_x V(x(t),t)$，所以此时我们可以算出最优控制的表达式：

$$u^*(t)=-2Q(t)x(t)=-2\frac{exp(2T)}{exp(2t)+exp(2T)}x(t)$$

我们最后一个任务就是求出 $x(t)$。我们列出有关$x(t)$的常微分方程：

方程： $\dot{x}(t)=(1-2\frac{exp(2T)}{exp(2t)+exp(2T)})x(t)$；

边界条件： $x(0)=x_0$；

在从这个方程解出 $x(t)$ 的表达式之后，我们就可以推出整个 $u^*(t)$，它是一个只含有时间变量 t 的函数。做到这里，这个问题就算是完全解决了。由于这个方程的解形式比较复杂，我们就不详细求解了。读者可以自己补全剩下的部分。

虽然是求解一个最简单的LQR问题，但我们却经历了如此复杂的运算步骤。下面，我们归纳一下这个过程中我们都经历了哪些步骤：

给定初始状态：$x(0)=x_0$；

环境：$\dot{x}(t)=x(t)+u(t)$;

目标：极小化$J=\int_0^T \frac{1}{4}u^2(t)dt+\frac{1}{4}x^2(T)$;

基于价值的方法

1.定义价值：$V(x(t),t)=min_u[\int_t^T \frac{1}{4}u^2(t)dt+\frac{1}{4}x^2(T)]$;

2.列出H-J-B方程：$\dot{V}(x(t),t)+min_u[\nabla_x V(x(t),t)(x(t)+u(t))+\frac{1}{4}u^2(t)]=0$

3.求解最优控制表达式：$u^*(t)=-2\nabla_x V(x(t),t)$;

4.代入H-J-B方程，求解关于二元函数$V(x,t)$的偏微分方程：

$$\begin{cases}\text{方程：}V_t-V_x^2+V_x x(t)=0;\\ \text{边界条件：}V(x,T)=\frac{1}{4}x^2;\end{cases}$$

5.求出$V(x,t)=\frac{1}{2}\frac{exp(2T)}{exp(2t)+exp(2T)}x^2(t)$;

代入$u^*(t)=-2\nabla_x V(x(t),t)$，得到$u^*(t)=-2\frac{exp(2T)}{exp(2t)+exp(2T)}x(t)$;

6.求解关于一元函数$x(t)$的常微分方程：

$$\begin{cases}\text{方程：}\dot{x}(t)=(1-2\frac{exp(2T)}{exp(2t)+exp(2T)})x(t);\\ \text{边界条件：}x(0)=x_0;\end{cases}$$

7.根据$x(t)$与$u^*(t)=-\frac{exp(2T)}{exp(2t)+exp(2T)}x(t)$得出$u^*(t)$的最终表达式；

上面的这张表高度概括了我们求解这个问题中的各个步骤。并且，它与上一节中我们给出的解题步骤表格是一一对应的，每一步内容是上一节那张表中每一步内容的具体化。读者不妨对照两张表看，或许可以更加深刻地理解上一节中解题步骤表的内容。

至此，我们解出连续时间的LQR问题。我们可以对比一下它的解题步骤与上一章中我们讲的离散时间的LQR问题有什么不一样：

首先一个最显著的不同是，它没有 $Q(x,u,t)$ 。这是因为在连续时间情况下， $Q(x,u,t)$ 与 $V(x,t)$ 只相差一个微元。所以严格地说，如果一定要求出 $Q(x,u,t)$ 的值，则 $Q(x,u,t)$ 在 u 取所有值时都等于 $V(x,t)$ 。但是有细心的读者可能已经发现了，所谓 $Q(x,u,t)$ 与 $V(x,t)$ 只相差一个微元，这个微元应该等于 $[\mathcal{H}(u^*)-\mathcal{H}(u)]dt$ 。所以在某种意义上说，$u^*=argmin_u Q(x,u,t)$ 就等同于 $u^*(t)=argmin_u \mathcal{H}$ ，我们依然可以想象我们是通过极小化 Q 来选择 u 的（但不严格）。总的来说，无论时间是离散还是连续，我们都是根据某个能够在特定状态下衡量控制“好坏”的函数来选择 u^* 的。

另外，在时间离散的情况下，我们是利用公式 $Q(x,u,t)=C(x,u)+V(f(x,u),t+1)$ 与 $V(x,t)=min_uQ(x,u,t)$ 从后往前依次交替地算出所有 t 对应的 $Q(x,u,t)$ 与 $V(x,t)$。事实上，这两个公式也可以联立为只关于 $V(x,t)$ 的方程，即 $V(x,t)=min_u(C(x,u)+V(f(x,u),t+1))$。我们完全可以通过这个方程迭代求解出所有 $V(x,t)$，然后再根据它求解 Q 与 u^*，与连续时间的情况一样。不过，由于我们本来就要把 $Q(x,u,t)$ 与 $V(x,t)$ 都解出来，所以采用了迭代求解与 V 的方法；而在时间连续的情况下，由于没有 $Q(x,u,t)$，所以我们就只能利用关于$V(x,t)$的公式解出 $V(x,t)$；

简单地总结一下，在离散时间的问题中，我们迭代地对所有 t 分别求出 $Q(x,u,t),V(x,t)$ 与最优控制 $u^*(t)$。而在连续时间的问题中，我们先直接求出 $V(x,t)$，再通过优化关于 V 的 $\mathcal{H}$ 来求出 $u^*(t)$。我们之前讲过，对于确定的环境，价值是确定的，可以由边界条件与环境给出的转移关系决定。而最佳控制则是由价值算出来的。从这个角度说，后者解法其实是更贴近问题本质的，前者只是在后者基础上结合时间离散的特殊性简化而来的，不利我们理解基于价值问题的本质。

4、关于价值的方程

讲到这里，我们已经对于状态与动作是离散或是连续、时间是离散或是连续的、环境是否具有随机性的MDP分别讲解了基于价值的方法。想必读者已经发现，它们本质上都是求解一个方程。下面，让我们来总结一下我们遇到过的所有方程。

首先，我们这里说的方程指的都是关于某个未知函数的方程，其目标是为了解出某个我们不知道的函数，并可以使得我们利用这个函数来求解最佳策略。可以说，这个函数是方程的“主角”。回顾一下符合上述定义的函数，包括 V_π,V,Q_π,Q 等。要注意的是，它们都是某种关于“好坏”的度量，更具体地说，它们衡量的都是未来能够获得的期望效用。当我们说“基于价值的思想”时，一般是指解出这些函数中的一个或几个，因为它们在广义上都可以被称为是“和价值有关的函数”；但我们说“价值”的时候，一般特指 V 函数。

然后，我们来回顾一下我们遇到过哪些关于 V_π,V,Q_π 或 Q 的方程：

对于状态、动作、时间都为离散变量，环境随机且时齐的MDP（更一般地假定策略也是随机的，用 $\pi(a|s)$ 表示），并且假定效用随着时间指数衰减。我们就可以列出关于 V_π 的方程：

$$V_\pi(s)=\Sigma_a\pi(a|s)[R_s^a+\gamma\Sigma_{s'}P_{s,s'}^aV_\pi(s')]$$

同样的问题中，我们有关于 Q_π 的方程（这里，$Q_\pi(s,a)$ 表示在 s 立即采取 a，后续再按照 π 给出的策略走以获得的期望价值）：

$$Q_\pi(s,a)=R_s^a+\gamma\Sigma_{s'}P_{s,s'}^aV_\pi(s')$$

考虑到 $V_\pi(s)=\Sigma_a\pi(a|s)Q_\pi(s,a)$，则我们可以将上面的方程简化为只是关于 Q_π 本身的方程：

$$Q_\pi(s,a) = R_s^a + \gamma \Sigma_{s'} P_{s,s'}^a (\Sigma_{a'} \pi(a'|s') Q_\pi(s',a'))$$

如果环境是非时齐的，则 π 与 V_π 应为包含 t 的函数，方程如下所示：

$$V_\pi(s,t) = \Sigma_a \pi(a|s,t)[R_s^a + \gamma \Sigma_{s'} P_{s,s',t}^a V_\pi(s',t+1)]$$

对于环境非时齐的 Q_π 的方程我们就不详细列出了，留给读者思考。

如果我们要求解的不是“基于策略的价值的动作”，而是“价值”。由于它强调的是“最佳动作”对应的后果，所以与上面相比，它将 $\Sigma_a \pi(a|s)$ 的部分给修改为了 max_a，而别的地方几乎没什么变化。因此，我们可以将上面的几条方程分别简化为：

$$V(s) = max_a[R_s^a \gamma \Sigma_{s'} P_{s,s'}^a V(s')]$$

$$V(s,t) = max_a[R_s^a \gamma \Sigma_{s'} P_{s,s'}^a V(s',t+1)]$$

对于非时齐的 Q 函数，读者可以尝试自己列出有关方程。

在动作、控制都连续，时间却是离散的LQR问题中，环境是非随机的（ $x' = f(x,u)$ ）、非时齐的（到 T 终止），由于我们考虑的是总损失，则我们列出的关于 V 方程为：

$$V(x,t) = min_u(C(x,u) + V(f(x,u),t+1))$$

同理，我们列出关于 $Q(x,u,t)$ 的方程为：

$$Q(x,u,t) = C(x,u) + V(f(x,u),t+1) = C(x,u) + min_u Q(f(x,u),u,t+1)$$

在对于状态与动作是连续的MDP，考虑 V_π 或 Q_π 是有一定困难的，因为 π 的形式是一个连续函数。读者可以尝试考虑一下这种情况下的Bellman方程，我们这里不详细列出了。

在本章中，我们列出了关于 $V(x,t)$ 的H-J-B方程。想必读者可以感觉到，这与前面的LQR问题中的方程也是很相似的。主要的不同在于环境从 $x' = f(x,u)$ 进一步变成了微分的形式，所以衡量下一个状态的价值的项有所不同。除此之外，可以说和之前几乎是一模一样的：

$$0 = \dot{V}(x(t),t) + min_u[\nabla_x V(x(t),t) f(x(t),u(t)) + C(x(t),u(t))]$$

至此，我们列举出了我们之前见过的所有方程，它们的解法各不相同。不过，当我们将方程中未知函数解出后（由于这个函数与“价值”有关），就能够利用它来求解最佳策略。

有的读者可能会觉得上述方程多而复杂。但是，有敏感的读者想必也能发现，这些方程之间都具有某种相似性，例如每一个方程的右边似乎都有两项，分别代表立即的效果，与下一个状态的后续效果；例如总会出现 max （对于奖励）或 min （对于损失）等等。事实上，我们不用刻意去记住每一种MDP对应的具体方程是什么，因为MDP可以根据动作与状态的连续性、环境的随机性、时

间的连续性、时齐性分为很多类，其对应的方程也有很多不同，很难通过死记硬背的方法记下来。最关键的还是要理解问题、理解价值的含义，这样就可以直接通过问题的条件列出我们想要求的关于价值的函数的方程。

下面，让我们总结一下列出这些方程的步骤：

首先，根据需要选择要研究的“基于价值”的函数是 V, Q 还是 V_π, Q_π ；再根据环境是否时齐判断函数的变量，例如是 $V(s)$ 还是 $V(s,t)$ ；这样，我们就确定了函数具体形式。当然，更重要的是要理解这个函数的意义。

然后，根据函数的含义确定关于它的恒等式。例如， Q 就等于立即采取 a 的后果， V 就等于采取最好的 a 的后果， V_π 就等于按照 π 来选择 a 的后果。一般而言，这个“后果”分为立即奖励或损失以及后续的“后果”。前者一般可以直接列出（例如 R_s^a 或 $C(x,u)$ ，而后者需要结合MDP的状态转移关系来表示（例如 $\Sigma_{s'}P_{s,s'}^a V(s)$ 或 $Q(f(x,u),u,t+1)$ ），等等。

列出方程后，我们要考虑如何求解。例如对于时间连续的情况，可能要用微分方程；对于时间离散、且有边界条件的情况，可能可以用迭代法从边界条件出发依次求出其在整个定义域的取值——这在某种程度上相当于离散化地求解含有边界条件的微分方程；而对于只有方程没有边界条件的情况，我们可以产生一个随机初始解，然后从初始解出发，用雅克比迭代的方法收敛到最优解。不过这种情况下要注意讨论一下迭代法的收敛性，或者讨论一下方程是否只有唯一的解。

以上，就是比较一般的、基于价值方法的求解思想。有的读者看完了这几段的内容难免还是会觉得有些抽象。笔者认为，一个比较好的解决方法是读者可以尝试针对时间离散的MDP的所有的不同具体情况（在状态与动作连续或离散、环境时齐或非时齐、环境随机或确定），分别列出关于 V, Q, V_π, Q_π 的方程。当读者能够将这各种情形下的方程都列出来，并能从直观上感觉它们符合逻辑的时候，便真正掌握了基于价值的思想。

上面我们特别强调了在时间离散的情况下，是因为时间是连续的与时间是离散的会有比较大的不同。例如，如果时间连续且环境随机，就要涉及到随机微分方程的内容。由于一般的理科院系是不学这方面内容，且它与本书的主题——强化学习常用算法——的关系相对比较小，所以我们在本书中不做特别要求。

由于时间是否连续为MDP带来的区别比较大，我们一般将时间是连续变量时的方程称为H-J-B（Hamilton-Jacobi-Bellman）方程，而将时间是离散变量时的方程称为Bellman方程，以将二者加以区分。Bellman方程在强化学习中是十分重要的，例如我们会根据如下Bellman方程来设计出一个十分经典的算法——DQN：

$$Q(s,a)=R_s^a+\gamma\Sigma_{s'}max_{a'}P_{s,s'}^aQ(s',a')$$

至此，在最优控制问题中，基于价值的思想就彻底讲完了。希望读者好好吸收其中的思想，以便能够在后面的强化学习问题中正确地应用它。

思考题

1、为什么说 $V(x,t)-Q(x,u,t)$ 等于 $[\mathcal{H}(u^*)-\mathcal{H}(u)]dt$？

2、请总结一下连续时间和离散时间的最优控制问题有什么区别?

3、本章中那个LQR问题最终的解 $u^*(t)$ 是多少？补全求解过程中未完成的最后一步。

4、对于时间是离散变量的MDP，分别考虑在状态与动作连续或离散、环境时齐或非时齐、环境随机或确定的情况下，列出有关于 V,Q,V_π,Q_π 的方程（可以做一些额外假定，例如状态动作连续的情况下， π 是确定性策略而非随机策略）。

编辑于 2020-08-04 11:24

最优控制　强化学习 (Reinforcement Learning)　偏微分方程

欢迎参与讨论

16 条评论　默认　最新

Yingru Li
请问HJB方程在未知转移方程未知cost的情况下有人研究过吗
2020-08-27　回复　1

heaven 作者 ▸ **Yingru Li**
时间连续在现实中不太好实现
2020-08-27　回复　喜欢

Yingru Li ▸ **heaven**
当然是说强化学习，就是想问有没有人系统的研究过，证明一些东西。像是Q Learning。
2020-08-27　回复　喜欢

展开其他 1 条回复 ›

GUESS

本篇文章由‘机器学习与控制论’转载表于‘强化学习轻松入门’专栏，该专栏专研于RL的基础理论推导与介绍，欢迎大家关注与讨论。

2020-08-21 回复 1

彬迅

写得好清晰，谢谢博主啦

01-31 回复 喜欢

dongqi

博主 有pi2算法相关的讲解吗，新手小白 看了一些论文里的pi2算法，感觉很懵逼。。

2023-05-28 回复 喜欢

陈道之

书什么时候出版?

2022-08-28 回复 喜欢

heaven 作者

关注我等消息，预计明年

2022-08-28 回复 喜欢

新时代优秀青年

二次这个方程会有两个解的情况吗

2022-05-08 回复 喜欢

逍遥药师

你好，这个现在求解一般是通过神经网络来逼近的吗，有类似的程序吗？谢谢了

2021-05-19 回复 喜欢

drageon

为何最大化汉密尔顿量的U就是最优控制啊！！！

2021-02-21 回复 喜欢

heaven 作者

你从离散的观点看，再变成无穷小量极限，道理差不多的嘛

2021-02-22 回复 喜欢

Story

请问在直接求解HJB方程非常困难的情况下，我可不可以先找一个不错的可行解，然后逐渐逼近HJB方程解（就像最优化里面的梯度下降法一样）？这是一种解HJB方程的方法吗？请问在哪里可以找到系统介绍解HJB方程的方法？

2020-09-12 回复 喜欢

heaven 作者

这是一种pde，我举的例子比较特殊，是线性二次的。一般的话就看看pde怎么解吧

2020-09-12 回复 喜欢

兔兔

写得很好，关注了

2020-07-27 回复 喜欢

欢迎参与讨论

强化学习轻松入门
一起来学习深度强化学习吧!

强化学习知识大讲堂
专注强化学习算法分享，欢迎投稿到该专栏

机器学习与控制论
当维纳遇上图灵

DQN

推荐阅读

深度学习求解偏微分方程（13）BURGERS’ ...

在深度学习求解偏微分方程中，有一些普遍用来测试的测试函数，比如BURGERS’ EQUATION，DARCY FLOW，NAVIER-STOKES EQUATION。 今天介绍BURGERS’ EQUATION的含义...

Kelle... 发表于深度学习求...

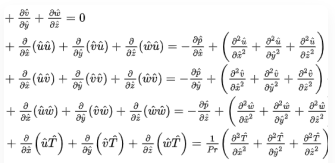
常见的可求解析解的微分方程

数学天才琪... 发表于微分方程的...

偏微分方程8th

1. 证明下面 Neumann 边值条件热方程至多只有一个光滑解. \ [\begin{cases} u_t-\Delta u=f \quad in \ U_T \\ \qquad\ \frac{\partial u}{\partial u}=0 \quad on \ \partial...

三川啦啦啦 发表于三川的作业...



有限体积法（12）基本方程推导——无量纲化

珐式小蛋糕