7 Lecture 7. Dynamic programming II

7.1 Policy iteration

In previous lecture, we studied dynamic programming for discrete time systems based on Bellman's principle of optimality. We studied both finite horizon cost

$$J = \varphi(x_N) + \sum_{k=1}^{N-1} L_k(x_k, u_k), \quad u_k \in U_k$$

and infinite horizon cost

$$J = \sum_{k=1}^{\infty} L(x_k, u_k), \quad u_k \in U(x_k).$$

The key ingredients we obtained were the Bellman equations. For finite horizon,

$$J_k^*(x) = \min_{u \in U_k} \{L_k(x, u) + J_{k+1}^*(f(x, u))\}$$

with boundary condition

$$J_N^*(x) = \varphi(x)$$

and for infinite horizon,

$$J^*(x) = \min_{u \in U} \{L(x, u) + J^*(f(x, u))\}.$$

The finite horizon Bellman equation can be solved backward. While for infinite horizon one, normally explicit solution cannot be obtained. In practice, one uses iteration approaches to approximate the solution. There are mainly two types of iteration approaches, namely, value iteration and policy iteration. In previous lecture, we studied value iteration, which starts with an initial function J_0 and updates according to

$$J_{k+1}(x) = \min_{u \in U(x)} \{L(x, u) + J_k(f(x, u))\}.$$

$$u_k \in \arg \min_{u \in U(x)} \{L(x, u) + J_k(f(x, u))\}.$$

The following proposition can be used to guarantee convergence of the algorithm. For proof, see [1].

Proposition 1 (Convergence of value iteration I)**.** *If U is a metric space and the sets*

$$U_k(x, \lambda) = \{u \in U(x) : L(x, u) + J_k(f(x, u)) \leq \lambda\}$$

is compact for all $x \in X$, $\lambda \in \mathbb{R}$ and k, then the value iteration $J_k \uparrow J^*$ pointwisely for any $J_0 \geq 0$ satisfying $J_0(x) \leq \min_{u \in U(x)} L(x, u) + J_0(f(x, u))$ for all $x \in X$, e.g., $J_0 = 0$.

Now let's study policy iteration. PI starts from a policy $u_1(\cdot)$ – stablizing – then solve

$$J_k(x) = L(x, u_k(x)) + J_k(f(x, u_k(x))) \tag{1}$$

for $J_k(\cdot)$. Next, iterate $u_k(\cdot)$ according to

$$u_{k+1}(x) \in \arg \min_{u \in U(x)} \{L(x, u) + J_k(f(x, u))\}. \tag{2}$$

The main result of policy iteration that we are going to prove is the following.

Proposition 2. *Let Assumption 1 (see Lecture note 6) hold. A sequence $\{J_k\}$ generated by the policy iteration algorithm (1), (2) satisfies $J_k(x) \downarrow J^*(x)$ for every $x \in X$.*

(*Proof sketch*) We study the relation between J_1 and J_2: since

$$J_1(x) = L(x, u_1(x)) + J_1(f(x, u_1(x)),$$

by definition of $u_2(\cdot)$, we have

$$J_1(x) \geq L(x, u_2(x)) + J_1(f(x, u_2(x))$$

Let $x_1 = x$ and $x_2, \cdots$ be the sequence under policy u_2, then

$$\begin{aligned} J_1(x) &\geq L(x_1, u_2(x_1)) + J_1(x_2) \\ &\geq L(x_1, u_2(x_1)) + L(x_2, u_2(x_2)) + J_1(f(x_2, u_2(x_2)) \\ &\vdots \\ &\geq \sum_{i=0}^{N-1} L(x_i, u_2(x_i)) + J_1(f(x_N, u_2(x_N)) \\ &\vdots \\ &\geq \sum_{i=1}^{\infty} L(x_i, u_2(x_i)) \\ &= J_2(x) \end{aligned}$$

(by assuming $J_1 \geq 0$) By the way, we also notice that $J_k \geq J_*$ for all k. Thus we conclude that $J_1 \geq J_2$. Repeating the procedure, we get

$$J_1 \geq J_2 \geq \cdots \geq J_k \geq \cdots \geq J_*.$$

So the PI algorithm generates a decreasing sequence. To show that there is no gap, one needs assumption 1.

7.2 Continuous time dynamic programming

In this section, we begin to study dynamic programming for continuous time systems:

$$\begin{aligned} &\dot{x} = f(x, u), \\ &x(t_0) = x_0 \end{aligned} \tag{3}$$

where $x(t) \in X \subseteq \mathbb{R}^n$, $u(t) \in U_t \subseteq \mathbb{R}^m$ and $u(\cdot) \in \mathcal{U}$, and $\mathcal{U}$ is called the *space of admissible control input*. To avoid pathological cases, we assume that the solution to this equation exists and is unique for each $u(\cdot) \in \mathcal{U}$. When the initial condition is clear from the context or is not important to us, we simply write $x(t)$ as the solution to system. If however we want to highlight the initial condition, we may write $x(t, x_0)$ (when the initial time instant is unimportant) or $x(t; t_0, x_0)$. If, further more, we want to include the input, we can write $x(t; t_0, x_0, u)$ for $u \in \mathcal{U}$.

We consider cost functions in *Bolza form*

$$J(u(\cdot)) = \varphi(x(T)) + \int_0^T L(x(s), u(s))\mathrm{d}s \tag{4}$$

where φ and L are both non-negative functions. Likewise, we can consider infinite horizon cost

$$J(u(\cdot)) = \int_0^{\infty} L(x(s), u(s))\mathrm{d}s$$

As before, the objective of optimal control is to seek for an admissible control $u^*(\cdot)$ such that

$$u^*(\cdot) \in \arg\min_{u \in \mathcal{U}} J(u(\cdot)). \tag{5}$$

To apply Bellman's principle, as before, we define cost-to-go and value functions. Again, these are done by simply changing the summation to integration in the discrete time setting. In words, the *cost-to-go function* from $t = s$ with $x(s) = y$ to T is

$$J(s, y; u(\cdot)) := \varphi(x(T)) + \int_s^T L(x(t), u(t))dt,$$

and the *value function* is defined as the optimal value of the cost-to-go under admissible control on the interval $[s, T]$:

$$J^*(s, y) := \min_{u(\cdot) \in \mathcal{U}|_{[s,T]}} J(s, y; u(\cdot)). \tag{6}$$

Here, the set $\mathcal{U}|_{[s,T]}$ is the set of admissible controls that can be implemented on the interval $[s,T]$. More rigorously, $\mathcal{U}|_{[s,T]} = \{u1_{[s,T]} : u \in \mathcal{U}\}$, where $1_{[s,T]}$ stands for the characteristic function of the set $[s,T]$.

Recall that the Bellman principle says: an optimal policy has the property that no matter what the previous decision have been, the remaining decisions must constitute an optimal policy with regard to the state resulting from those previous decisions. Thus for any time instant r, there must hold

$$J^*(s,y) = \min_{u(\cdot)\in\mathcal{U}|_{[s,r]}} \left\{ \int_s^r L(x(t;s,y,u),u(t))\mathrm{d}t + J^*(r,x(r;s,y,u)) \right\}, \quad \forall r \in [s,T]. \tag{7}$$

This equation looks too implicit and is hard to use in practice. The main task now is to derive the celebrated *Hamilton-Jacobi-Bellman* equation based (7), a more tractable form than (7). The key is to note that (7) is satisfied for all $r \geq s$ and hence one can take derivatives when J^* are assumed to be smooth.

On the one hand, for any give $u \in \mathcal{U}$, and $r > s$, we have

$$\frac{J^*(s,y) - J^*(r,x(r;s,y,u))}{r-s} - \frac{1}{r-s}\int_s^r L(x(t;s,y,u),u(t))\mathrm{d}t \leq 0, \quad \forall u \in \mathcal{U}$$

Suppose that J^* is continuously differentiable, and that L, u are continuous, then the above implies

$$-\frac{\partial J^*}{\partial s}(s,y) - \frac{\partial J^*}{\partial y}(s,y)f(y,u(s)) - L(y,u(s)) \leq 0, \quad \forall u(s) \in U_s$$

resulting in

$$-\frac{\partial J^*}{\partial s}(s,y) + \sup_{u\in U_s} H\left(y,u,-\frac{\partial J^*(s,y)}{\partial y}\right) \leq 0. \tag{8}$$

where

$$H(x,u,p) = p^\top f(x,u) - L(x,u) \tag{9}$$

On the other hand, for any pair (r,ϵ), with $r > s$, $\epsilon > 0$, there exists a control $u_{\varepsilon,r}$ such that

$$J^*(s,y) \geq \int_s^r L(x(t,s,y,u_{\varepsilon,r}(\cdot)),u_{\varepsilon,r}(t))\mathrm{d}t + J^*(r,x(r;s,y,u_{\varepsilon,r})) - \varepsilon(r-s)$$

or

$$\begin{aligned}
-\varepsilon &\leq \frac{J^*(s,y) - J^*(r,x(r,s,y,u_{\varepsilon,r})}{r-s} - \frac{1}{r-s}\int_s^r L(x(t,s,y,u_{\varepsilon,r}),u_{\varepsilon,r}(t),t)\mathrm{d}t \\
&= -\frac{1}{r-s}\int_s^r \left[\frac{\partial J^*}{\partial s}(t,x(t,s,y,u_{\varepsilon,r})) + \frac{\partial J^*}{\partial y}(t,x(t,s,y,u_{\varepsilon,r})f(x(t,s,y,u_{\varepsilon,r}),u_{\varepsilon,r}(t))\right]\mathrm{d}t \\
&\quad - \frac{1}{r-s}\int_s^r L(x(t,s,y,u_{\varepsilon,r}),u_{\varepsilon,r}(t),t)\mathrm{d}t \\
&= \frac{1}{r-s}\int_s^r \left[-\frac{\partial J^*}{\partial s}(t,x(t,s,y,u_{\varepsilon,r}) + H\left(x(t,s,y,u_{\varepsilon,r}),u_{\varepsilon,r}(t),-\frac{\partial J^*}{\partial y}(t,x(t,s,y,u_{\varepsilon,r})\right)\right]\mathrm{d}t
\end{aligned}$$

Let $r \to s+$ while keeping ϵ fixed, we get

$$\begin{aligned}
-\varepsilon &\leq -\frac{\partial J^*}{\partial s}(s,y) + H\left(y,u_{\varepsilon,r}(s),-\frac{\partial J^*}{\partial y}(s,y)\right) \\
&\leq -\frac{\partial J^*}{\partial s}(s,y) + \sup_{u\in U_s} H\left(y,u,-\frac{\partial J^*}{\partial y}(s,y)\right)
\end{aligned} \tag{10}$$

Since ε is arbitrary, (10) and (8) together imply

$$-\frac{\partial J^*}{\partial s}(s,y) + \sup_{u\in U_s} H\left(y,u,-\frac{\partial J^*}{\partial y}(s,y)\right) = 0, \quad \forall s \in [0,T],\ \forall y \in X$$

or equivalently

$$\frac{\partial J^*}{\partial s}(s,y) + \inf_{u\in U_s}\left\{\frac{\partial J^*}{\partial y}(s,y)f(y,u) + L(y,u)\right\} = 0, \quad \forall s \in [0,T],\ \forall y \in X$$

This is a partial differential equation with dependent variable (s, y). By the definition of value function (see (6)), the PDE is accompanied with boundary condition

$$J^*(T, y) = \varphi(y), \quad \forall y \in X.$$

Summarizing, we have:

Proposition 3. *Suppose that $J^*(s, x)$ defined as (6) is continuously differentiable. Then $J^*(t, x)$ is a solution to the following Hamilton-Jacobi-Bellman (HJB) PDE on $[0, T] \times X$:*

$$-V_t(t, x) + \sup_{u \in U_t} H\left(x, u, -V_x(t, x)\right) = 0, \tag{11}$$

or its equivalent form

$$V_t(t, x) + \inf_{u \in U_t} \{V_x(t, x) f(x, u) + L(x, u)\} = 0, \tag{12}$$

with boundary condition

$$V(t, x) = \varphi(x),$$

where H is defined in (11) and we have adopted the notations $\frac{\partial V}{\partial t} = V_t$ and $\frac{\partial V}{\partial x} = V_x$.

Suppose that $U_t = U$ for all $t \geq 0$. To obtain the optimal control law based on the solution of the HJB equation (11), we can follow Algorithm 1 (called the *verification rule*).

Algorithm 1 The verification rule

1. Solve the optimization problem

$$u^*(x, p) = \arg\sup_{u \in U} H(x, u, p).$$

2. Find a continuously differentiable solution $V(t, x)$ to

$$\begin{aligned} -V_t(t, x) + & H\left(x, u^*(x, -V_x(t, x)), -V_x(t, x)\right) = 0, \\ V(T, x) = & \varphi(x), \end{aligned} \tag{13}$$

   for $(t, x) \in (0, T] \times X$.

3. Then $u^*(t, -V_x(t, x^*(t)))$ is an optimal control and $V(t, x) = J^*(t, x)$.

Example 1 (Double integrator)**.** Consider the double integrator

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u \end{aligned}$$

where $|u| \leq 1$. The objective is to steer the system to the origin in minimum time, i.e.,

$$\min \int_0^{t_f} 1 \mathrm{d}t.$$

The Hamiltonian function is

$$H(x, u, p) = p_1 x_2 + p_2 u - 1.$$

Step 1:

$$u_*(t) = \arg\sup_{u \in [-1, 1]} (p_1 x_2 + p_2 u - 1) = \mathrm{sign} p_2(t).$$

Step 2: The HJB equation reads

$$-\frac{\partial V}{\partial t} + H(x, -\mathrm{sign}(\frac{\partial V}{\partial x_2}), -\frac{\partial V}{\partial x}) = 0$$

or

$$-V_t - V_{x_1}x_2 + |V_{x_2}| - 1 = 0.$$

Since there is no boundary cost, $V(t,0) = 0$.

Step 3: after solving the HJB, the controller is obtained as

$$u = -\text{sign}V_{x_2}$$

which is a feedback controller.

As the discrete time optimal control problem on finite horizon, solving the Bellman equation (11) (when the solution has some regularities) is sufficient to obtain the optimal control. For continuous problems, we have a similar result.

Proposition 4. *If the verification rule Algorithm 1 admits a C^1 solution V, then u^* obtained from the algorithm is an optimal control.*

Proof. Let V be a C^1 solution to the Bellman equation. Let $u(\cdot)$ be any admissible control and $x(\cdot)$ the corresponding trajectory. Then for initial condition x_0,

$$\begin{aligned} V(T,x(T)) - V(t,x(t)) &= \int_t^T \frac{dV(s,x(s))}{ds}\mathrm{d}s \\ &= \int_t^T V_t(s,x(s)) + V_x(t,x(s))f(x(s),u(s))\mathrm{d}s \\ &\geq \int_t^T -L(x(s),u(s))\mathrm{d}s \text{ (use(12))} \end{aligned}$$

from which it follows that

$$V(t,x(t)) \leq \varphi(x(T)) + \int_t^T L(x(s),u(s)\mathrm{d}s.$$

The inequality becomes equality when x is taken to be x_*. Since $u(\cdot)$ is arbitrary, we conclude that $V(t,x)$ is the value function. On the other hand, it is readily checked that u^* is a control that achieves the optimal value. □

Apparently, the most challenging part of the algorithm is the second step, i.e., solving a PDE of the form $F(x,v,v_x) = 0$. But even numerically solving the HJB is quite difficult, which normally causes curse of dimensionality after discretization.

Method of characteristics

A well-known approach to solving PDE of the form

$$F(x,v,v_x) = 0, \quad x \in \Omega \subset \mathbb{R}^n. \tag{14}$$

with boundary condition

$$v(x) = g(x), \quad x \in \partial\Omega,$$

is via the so-called *method of characteristics*. Here v is a real valued function and F is assumed to be a continuous mapping from $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ to $\mathbb{R}$. In addition, we assume Ω to be compact with smooth boundary.

The idea of the method of characteristics is to turn the first order PDE (14) into a set of ODEs. Given a point $y \in \partial\Omega$ and a curve $x : [0,1] \to \bar{\Omega}$, with $x(0) = y$. We examine the values of $v(x)$ along this curve, see Figure 7.2.

Introduce the notation

$$(p_1, \cdots, p_n)^\top = (v_{x_1}, \cdots, v_{x_n})^\top.$$

For convenience, denote

$$\begin{aligned} v(s) &=: v(x(s)) \\ p(s) &=: p(x(s)) = v_x(x(s)). \end{aligned}$$

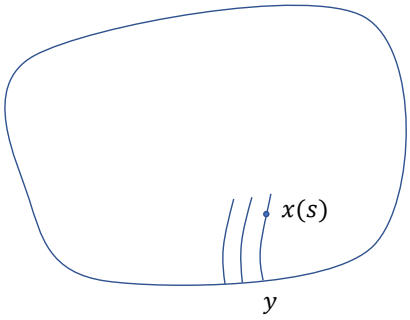

Figure 1: Method of characteristics.

Differentiating v and p w.r.t. s, we find

$$\dot{v} = \sum_{i=1}^{n} v_{x_i}\dot{x}_i = \sum_{i=1}^{n} p_i\dot{x}_i$$
$$\dot{p}_i = \sum_{j=1}^{n} v_{x_i x_j}\dot{x}_i$$

where $\dot{x}_i$ stands for the derivative of x_i w.r.t. s. Further, differentiating (14) w.r.t. x_i, we get

$$\frac{\partial F}{\partial x_i} + \frac{\partial F}{\partial v}v_{x_i} + \sum_{i=1}^{n}\frac{\partial F}{\partial p_i}v_{x_i x_j} = 0.$$

Now if the curve $x(s)$ is chosen such that $\dot{x}_i = \partial F/\partial p_i$ (this time, we call x a characteristic curve), one can easily obtain the following

$$\dot{v} = \sum_{i=1}^{n} p_i\frac{\partial F}{\partial p_i},$$
$$\dot{p}_i = -\frac{\partial F}{\partial x_i} - \frac{\partial F}{\partial v}p_i, \quad i = 1, \cdots, n$$

or in more compact form

$$\begin{cases} \dot{x} = F_p^\top \\ \dot{v} = F_p p \\ \dot{p} = -F_x^\top - F_v p \end{cases} \tag{15}$$

The above equation is a system of ordinary differential equations with boundary condition

$$x(0) = y, \quad v(0) = g(y), \quad p(0) = v_x(y)$$

for $y \in \partial\Omega$. Thus by varying the initial condition y, we can obtain *local solutions* near $\partial\Omega$ of the PDE (14). In general, however, the solution cannot be extend globally to the entire region Ω. For example, when two characteristic curves meet in Ω, singularity occurs.

To solve the HJB using method of characteristics, we first need to write the equation (13) into the standard form $F(x, v, v_x) = 0$ for some F. For that, let $x_{n+1} = t$ and $\tilde{x} = (x, x_{n+1})$. Then (13) can be written as $-v_{x_{n+1}} + H(x, u^*(x, v_x), -v_x) = 0$, or $-v_{x_{n+1}} + \tilde{H}(x, v_x) = 0$ for some scalar function $\tilde{H}$. Let $\tilde{p} = (p_1, \cdots, p_n, p_{n+1})$, then F takes the form

$$F(\tilde{x}, v, \tilde{p}) = -p_{n+1} + \tilde{H}(x, p).$$

Hence the first line of (15) reads

$$\begin{aligned} \dot{x} &= F_p^\top = \tilde{H}_p^\top \\ \dot{x}_{n+1} &= \frac{\partial F}{\partial p_{n+1}} = -1 \end{aligned} \tag{16}$$

Notice that $\partial F/\partial v = 0$, the third line of (15) reads

$$\begin{aligned} \dot{p} &= -\tilde{H}_x \\ \dot{p}_{n+1} &= -\frac{\partial \tilde{H}}{\partial x_{n+1}} = 0 \end{aligned} \tag{17}$$

and the second line of $\dot{v} = p^\top \tilde{H}_p - p_{n+1}$. In the above formulas, the only relevant ones are the first lines of (16) and (17), i.e.,

$$\begin{cases} \dot{x} = \tilde{H}_p^\top \\ \dot{p} = -\tilde{H}_x^\top \end{cases} \tag{18}$$

This is the *Hamiltonian equation* we have already met in PMP.

Notice all the previous derivations are based on the assumption that the value function is continuously differentiable. This is almost never met in practice. What's worse, the HJB equation may not have continuously differentiable solutions! This problem turns out be non-negligible and must be handled with care. This is why we cannot derive the maximum principle from the HJB equation: the regularity issue here is an essential difficulty.

7.3 Infinite horizon problems

Consider the time-invariant system

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

with cost

$$J = \int_0^\infty L(x(t), u(t))\mathrm{d}t$$

where $L \geq 0$, $u(t) \in U \subseteq \mathbb{R}^m$ for all $t \geq 0$. It is easy to notice that the value function in this case is time independent and thus can be written as $J^*(x)$. Further more, the HJB equation reads

$$\sup_{u\in U} H(x, u, -V_x) = 0$$

because $J^*(t,y) = J^*(r,y)$ for all t, r. Here $H(x,u,p) = p^\top f(x,u) - L(x,u)$, or equivalently

$$\inf_{u\in U}\{V_x f(x,u) + L(x,u)\} = 0. \tag{19}$$

In the LQR setting, for the system

$$\dot{x} = Ax + Bu \tag{20}$$

and cost

$$J = \int_0^\infty x^\top Qx + u^\top Ru\mathrm{d}t \tag{21}$$

the HJB equation (19) reads

$$\inf_{u\in U}\{V_x(Ax+Bu) + x^\top Qx + u^\top Ru\} = 0$$

As before, choose $V = x^\top Px$, then the above formula turns into $\inf_{u\in U}\{2x^\top P(Ax+Bu) + x^\top Qx + u^\top Ru\} = 0$. The minimum on the left hand side is achieved at

$$u^* = -R^{-1}B^\top Px$$

with minimum zero if

$$A^\top P + PA + Q - PBR^{-1}B^\top P = 0. \tag{22}$$

This equation in P is called *algebraic Riccati equation (ARE).*

Recall that $V(x_0) = \min J$, thus $x_0^\top Px_0$ is the optimal cost.

Proposition 5. *Consider the LTI system (20) and cost function (21) with $Q \geq 0$, $R > 0$. Assume (A,B) is controllable, (A,C) is observable, where C is full row rank satisfying $C^\top C = Q$. Then the ARE has a unique symmetric solution P which is positive definite. Further more, the optimal control is given by a static state feedback $u = -R^{-1}B^\top Px$ and the optimal cost is $x_0^\top Px_0$.*

Proof. The proof of this proposition is essentially the same as the discrete time case and is thus left as an exercise. □

References

[1] Dimitri P Bertsekas. Value and policy iterations in optimal control and adaptive dynamic programming. *IEEE transactions on neural networks and learning systems*, 28(3):500–509, 2015.