ECE 598: Interplay between Control and Machine Learning **Fall 2020**

Lecture 8

Policy Evaluation

Lecturer: Bin Hu, Date:09/29/2020

In this lecture, we discuss how to assess the performance of a given policy. Policy evaluation is an important task. The analysis tools will be tailored into design tools in later lectures.

8.1 Discrete Space Case

Recall that a MDP is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition kernel, R is the reward, and γ is the discount factor. Given a policy π, we want to analyze the associated value function:

$$V^\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \big| a_k \sim \pi(\cdot|s_k), s_0 = s\right].$$

For simplicity, let's first consider the value evaluation of a deterministic policy. If both $\mathcal{S}$ and $\mathcal{A}$ are finite, then the policy π can be represented as a vector

$$\begin{bmatrix} \pi(1) \\ \pi(2) \\ \vdots \\ \pi(n) \end{bmatrix}$$

where $\pi(i) \in \mathcal{A}$ and n is the size of $\mathcal{S}$. Then the value function becomes

$$V^\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, \pi(s_k)) \big| s_0 = s\right].$$

Now we can apply the law of total probability to show:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}R(s_0, \pi(s_0)) + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R(s_k, \pi(s_k)) \big| s_0 = s\right] \\ &= \mathbb{E}R(s, \pi(s)) + \sum_{s' \in \mathcal{S}} \left(\mathbb{E}[\sum_{k=1}^{\infty} \gamma^k R(s_k, \pi(s_k)) \big| s_1 = s'] \right) P(s_1 = s' \big| s_0 = s) \end{aligned}$$

When π is fixed, the state $\{s_k\}$ becomes a Markov chain. We have $P(s_1 = s' \big| s_0 = s) = P(s_1 = s' \big| s_0 = s,\, a_0 = \pi(s)) = P_{ss'}^{\pi(s)}$. Notice $\left[\sum_{k=1}^{\infty} \gamma^k R(s_k, \pi(s_k)) \big| s_1 = s'\right] = \gamma V^\pi(s')$. If we denote $\bar{R}^\pi(s) := \mathbb{E} R(s, \pi(s))$, then the equation on V^π can be rewritten as

$$V^\pi(s) = \bar{R}^\pi(s) + \gamma \sum_{s' \in \mathcal{S}} V^\pi(s') P_{ss'}^{\pi(s)} \tag{8.1}$$

which is the so-called Bellman equation. Recall $\mathcal{S} = \{1, 2, \ldots, n\}$. We can actually rewrite the above Bellman equation in the following matrix form:

$$\begin{bmatrix} V^\pi(1) \\ \vdots \\ V^\pi(n) \end{bmatrix} = \begin{bmatrix} \bar{R}^\pi(1) \\ \vdots \\ \bar{R}^\pi(n) \end{bmatrix} + \gamma \begin{bmatrix} P_{11}^{\pi(1)} & \cdots & P_{1n}^{\pi(1)} \\ \vdots & \ddots & \vdots \\ P_{n1}^{\pi(n)} & \cdots & P_{nn}^{\pi(n)} \end{bmatrix}$$

If we use the following vector notation:

$$V^\pi = \begin{bmatrix} V^\pi(1) \\ V^\pi(2) \\ \vdots \\ V^\pi(n) \end{bmatrix}, \; \bar{R}^\pi = \begin{bmatrix} \bar{R}^\pi(1) \\ \bar{R}^\pi(2) \\ \vdots \\ \bar{R}^\pi(n) \end{bmatrix}, \; P^\pi = \begin{bmatrix} P_{11}^{\pi(1)} & \cdots & P_{1n}^{\pi(1)} \\ \vdots & \ddots & \vdots \\ P_{n1}^{\pi(n)} & \cdots & P_{nn}^{\pi(n)} \end{bmatrix}$$

we can rewrite the Bellman equation as

$$V^\pi = \bar{R}^\pi + \gamma P^\pi V^\pi. \tag{8.2}$$

Therefore, the policy evaluation becomes an equation solving problem. Notice P^π is actually the transition matrix for the Markov chain $\{s_k\}$. Clearly, this matrix is right stochastic and the spectral radius is 1. Therefore, $I - \gamma P^\pi$ is nonsingular for any $0 < \gamma < 1$, and the Bellman equation admits a unique solution

$$V^\pi = (I - \gamma P^\pi)^{-1} \bar{R}^\pi$$

If we want to avoid matrix inversion, we can use an iterative scheme:

$$V_{k+1}^\pi = \bar{R}^\pi + \gamma P^\pi V_k^\pi$$

which is equivalent to a linear time-invariant system:

$$V_{k+1}^\pi - V^\pi = \gamma P^\pi (V_k^\pi - V^\pi)$$

Since the spectral radius of γP^π is γ, we immediately know the above system converges to V^π at a linear rate γ. The above scheme requires knowing P^π. When the model is unknown, we can somehow modify the above scheme and obtain the temporal difference learning method which is model free. We will talk about temporal difference learning next week.

When a stochastic policy is used, the Bellman equation still holds. We only need to slightly modify the definitions of $\bar{R}^\pi$ and P^π. For example, now we have $\bar{R}^\pi(s) = \mathbb{E}\left[R(s,a) \big| a \sim \pi(\big| s)\right]$. I will let you figure out how to modify P^π. In general, when a fixed stochastic policy is used, the state $\{s_k\}$ still becomes a Markov chain and P^π is the associated transition matrix. Then V^π can still be solved via the Bellman equation.

8.2 Continuous Space Case

For simplicity, let's consider the LQR setup:

$$x_{t+1} = Ax_t + Bu_t \tag{8.3}$$

we focus the policy evaluation for a linear policy $u_t = -Kx_t$. Substituting $u_t = -Kx_t$ into (8.3) leads to $x_{t+1} = (A - BK)x_t$. Hence we have

$$x_t = (A - BK)^t x_0 \tag{8.4}$$

We denote the spectral radius as ρ. If K stabilizes the system (8.3), then $\rho(A - BK) < 1$ and $x_t \to 0$ at a geometric rate. The quadratic cost becomes

$$\mathcal{C}(K) = \mathbb{E}_{x_0 \sim \mathcal{D}} \quad x_0^\mathsf{T} \left(\sum_{t=0}^{\infty} ((A - BK)^\mathsf{T})^t (Q + K^\mathsf{T} RK)(A - BK)^t \right) x_0 \tag{8.5}$$

When $\rho(A - BK) \geq 1$, the above cost blows up to infinity. It makes sense to restrict the policy search within the class of stabilizing K. When $\rho(A - BK) < 1$, we know $\sum_{t=0}^{\infty}((A - BK)^\mathsf{T})^t(Q + K^\mathsf{T}RK)(A - BK)^t$ will converge to a fixed constant matrix. We denote this matrix by P_K. Therefore, it is reasonable to parameterize the value function as $x_0^\top P_K x_0$ which is a quadratic function of x_0. When a nonlinear policy is used, we typically need to parameterize the value function as a neural network.

Bellman equation for policy evaluation. From the above discussion, we have already known $P_K = \sum_{t=0}^{\infty}((A - BK)^\mathsf{T})^t(Q + K^\mathsf{T}RK)(A - BK)^t$. The bellman equation can be derived as follows.

$$\begin{aligned}
x_0^\mathsf{T} P_K x_0 &= \sum_{t=0}^{\infty} x_t^\mathsf{T} (Q + K^\mathsf{T} RK) x_t \\
&= x_0^\mathsf{T}(Q + K^\mathsf{T}RK)x_0 + \sum_{t=1}^{\infty} x_t^\mathsf{T}(Q + K^\mathsf{T}RK)x_t \\
&= x_0^\mathsf{T}(Q + K^\mathsf{T}RK)x_0 + x_1^\mathsf{T} P_K x_1 \\
&= x_0^\mathsf{T}(Q + K^\mathsf{T}RK)x_0 + x_0^\mathsf{T}(A - BK)^\mathsf{T} P_K (A - BK)x_0 \\
&= x_0^\mathsf{T}\left(Q + K^\mathsf{T}RK + (A - BK)^\mathsf{T} P_K (A - BK)\right) x_0
\end{aligned}$$

Therefore, the Bellman equation takes the following form:

$$P_K = Q + K^\mathsf{T}RK + (A - BK)^\mathsf{T} P_K (A - BK) \tag{8.6}$$

For any fixed K, the above equation is a linear equation of P_K. Hence the existence and uniqueness of the solution to the above Bellman equation can be established using linear equation theory.

To obtain a closed-form solution for P_K, we need to introduce the Kronecker product and the vectorization operation. The Kronecker product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is denoted by $A \otimes B$ and given by:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

where a_{ij} is the (i, j)-th entry of A. Clearly, we have $A \otimes B \in \mathbb{R}^{pm\times qn}$. Notice $(A \otimes B)^T = A^T \otimes B^T$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ when the matrices have compatible dimensions.

Next, let vec denote the standard vectorization operation that stacks the columns of a matrix into a vector. For example, we have

$$\mathrm{vec}\left(\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 2 \\ 4 \\ 6 \end{bmatrix}.$$

An important fact is that we always have $\mathrm{vec}(AXB) = (B^\mathsf{T} \otimes A)\,\mathrm{vec}(X)$. Therefore, we have

$$\mathrm{vec}\left((A - BK)^\mathsf{T} P_K (A - BK)\right) = \left((A - BK)^\mathsf{T} \otimes (A - BK)^\mathsf{T}\right) \mathrm{vec}(P_K)$$

Then we can vectorize both sides of the Bellman equation (8.6) to obtain

$$\mathrm{vec}(P_K) = \mathrm{vec}(Q + K^\top RK) + \left((A - BK)^\mathsf{T} \otimes (A - BK)^\mathsf{T}\right) \mathrm{vec}(P_K)$$

which can be easily solved for P_K:

$$\mathrm{vec}(P_K) = \left(I - (A - BK)^\mathsf{T} \otimes (A - BK)^\mathsf{T}\right)^{-1} \mathrm{vec}(Q + K^\mathsf{T} RK)$$

Now we have a closed-form solution for P_K. Using properties of the Kronecker product, one can show $\left(I - (A - BK)^\mathsf{T} \otimes (A - BK)^\mathsf{T}\right)$ is nonsingular under the assumption $\rho(A-BK) < 1$. We skip the details here. The key message here is that for any stabilizing K, we can solve (8.6) to obtain P_K and then the value function for K is $V(x) = x^\mathsf{T} P_K x$.

For general nonlinear policy, the existence and uniqueness conditions for Bellman equation are much more complicated. Consider a nonlinear system $x_{t+1} = f(x_t, u_t)$ with some nonlinear policy $u_t = K(x_t)$. Then the Bellman equation takes the form of $V(x) = C(x, K(x)) + V(f(x, K(x)))$.

Finally, we consider LQR with process noise: $x_{t+1} = Ax_t + Bu_t + w_t$ where w_t is an IID process noise. Given a linear policy K, it is straightforward to use induction to show

$$V(x) = r_K + \quad x^\mathsf{T} \left(\sum_{t=0}^{\infty} \gamma^t ((A - BK)^\mathsf{T})^t (Q + K^\mathsf{T} RK)(A - BK)^t\right) x \tag{8.7}$$

where r_K is some extra term introduced by the noise w_t. Therefore, we can parameterize the value function as $x^\intercal P_K x + r_K$. Therefore, we have

$$V(x) = x^\intercal (Q + K^\intercal RK)x + \gamma \left(\mathbb{E}((A - BK)x + w)^\intercal P_K((A - BK)x + w) + r_K\right) \tag{8.8}$$

Notice w is independent from x and has a zero mean, we have

$$\mathbb{E}((A - BK)x + w)^\intercal P_K((A - BK)x + w) = x^\intercal (A - BK)^\intercal P_K(A - BK)x + \mathbb{E}(w^\intercal P_K w)$$

Notice that the left side of (8.8) is just $x^\intercal P_K x + r_K$. Hence (8.8) can be rewritten as

$$x^\intercal P_K x + r_K = x^\intercal (Q + K^\intercal RK)x + \gamma x^\intercal (A - BK)^\intercal P_K(A - BK)x + \gamma\mathbb{E}(w^\intercal P_K w) + \gamma r_K$$

To ensure that the quadratic functions on the left and right sides of the above equation are the same, the following have to be true:

$$\begin{aligned} x^\intercal P_K x &= x^\intercal (Q + K^\intercal RK)x + \gamma x^\intercal (A - BK)^\intercal P_K(A - BK)x \\ r_K &= \gamma\mathbb{E}(w^\intercal P_K w) + \gamma r_K \end{aligned}$$

Hence, the Bellman equation becomes

$$P_K = Q + K^\intercal RK + \gamma(A - BK)^\intercal P_K(A - BK)$$

and $r_K = \frac{\gamma}{1-\gamma}\mathbb{E}(w^\intercal P_K w) = \frac{\gamma}{1-\gamma}\,\mathrm{trace}(PW)$ where W is the covariance matrix of w_t.