Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

November 29, 2021

Discrete-Time OCP

$$
\begin{aligned}
&\text{minimize} \quad c_T(x_T) + \sum_{t=1}^{T-1} c_t(x_t, u_t) \\
&\text{subject to} \quad x_{t+1} = f_t(x_t, u_t), \quad \text{for} \quad t = 1, \ldots, T-1 \\
&\text{given} \quad x_1 \quad .
\end{aligned}
$$

The **value function** of a parametric mathematical program is the mapping from program parameters to optimal value. Here the "parameters" (inputs) are x_1 and the value is $c_T(x_T^\star) + \sum_{t=1}^{T-1} c_t(x_t^\star, u_t^\star)$ where $\{x_t^\star\}_{t=2}^{T}$ and $\{u_t^\star\}_{t=1}^{T-1}$.

Bellman Optimality Equations

In the optimal control context, the value function is called the cost-to-go function and satisfies the Bellman optimality equations:

$$
\begin{aligned}
J_T(x_T) &\triangleq c_T(x_T) \\
J_t(x_t) &\triangleq \min_{u_t \in \mathcal{U}(x_t)} \left\{ r_t(x_t, u_t) + J_{t+1}(f_t(x_t, u_t)) \right\} \quad t = 1, \ldots, T-1.
\end{aligned}
$$

The above are DP equations in a discrete-time finite horizon MDP.

Special case: Linear Quadratic Regulation

$$\begin{aligned}
&\text{minimize} \;\; x_T Q_T x_T + \sum_{t=1}^{T} \left(x_t^\top Q x_t + u_t^\top R u_t \right) \\
&\text{subject to} \;\; x_{t+1} = A x_t + B u_t, \;\; t = 1, \ldots T-1 \\
&\text{given} \;\; x_1 \quad .
\end{aligned}$$

Backward Induction for LQR: Ricatti Equation

The local minimization problem in the Bellman optimality equations can be solved in closed-form under the LQR setting.

The cost-to-go function is quadratic and of the form $J_t(x_t) \triangleq x_t^\top P_t x_t$ for all $t = 1, \ldots, T$. and the matrices $\{P_1, \ldots, P_T\}$ can be found by backward induction:

- Set $P_T = Q_T$
- From $t = T - 1, \ldots, 1$:
  - Set $P_t = Q + A^\top P_{t+1} A - A^\top P_{t+1} B (R + B^\top P_{t+1} B)^{-1} B^\top P_{t+1} A$
  - Set $K_t = -(R + B^\top P_{t+1} B)^{-1} B^\top P_{t+1} A$

You can then compute the optimal control at time t in state x_t with $u_t^\star = K_t x_t$ (the optimal controls are linear in the states).

Continuous-Time

Is there also a Dynamic Programming approach to continuous-time OCPs?

$$
\begin{aligned}
&\text{minimize} \;\; c(x(t_f)) + \int_{t_0}^{t_f} c(x(t), u(t))dt \\
&\text{subject to} \;\; \dot{x}(t) = f(x(t), u(t)) \\
&\text{given} \;\; x(t_0) = x_0 \quad .
\end{aligned}
$$

The Hamilton-Jacobi Equations

We can show that the cost-to-go for the above continuous-time problem function satisfies for all x, t:

$$J(x,T) = c(x) \quad \text{and} \quad 0 = \min_{u \in \mathcal{U}} \left\{ c(x,u) + \underbrace{D_2 J(x,t)}_{\text{time derivative}} + \overbrace{D_1 J(x,t)}^{\text{space derivative}} f(x,u) \right\}$$

Note that while the OCP only involves an ODE, the HJB are partial differential equations (PDEs): a specification of how the partial derivatives of a function ought to behave together.

Informal Derivation

Idea: discretize using Euler and apply DP. Let's pick a uniform grid wih n intervals, so that $h = T/n$.

- Discretized dynamics:

$$x_{t+1} = x_t + hf(x_t, u_t) \quad .$$

- Discretized integral cost:

$$c(x_n) + \sum_{k=0}^{n-1} c(x_k, u_k)h \quad .$$

Why?

Approximation of the Integral Cost

Consider a problem of the form:

$$\text{find } \ z(t_f) = \int_{t_0}^{t_f} c(x(t))dt \ \text{ such that } \ \dot{x}(t) = f(x(t)) \ \text{ and } \ x(t_0) = x_0 \ .$$

We can solve this problem by forming an augmented IVP:

$$\text{find } \ \tilde{x}(t_f) \ \text{ such that } \ \begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = \tilde{f}(\tilde{x}(t)) = \begin{bmatrix} f(x(t)) \\ c(x(t)) \end{bmatrix} \text{ given } \ \tilde{x}(t_0) = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$$

Approximation of the Integral Cost

Using Euler discretization, we would get:

$$\tilde{x}_{k+1} = \begin{pmatrix} x_k \\ z_k \end{pmatrix} + h \begin{pmatrix} f(x_k) \\ c(x_k) \end{pmatrix}$$

Both components can also written non-recursively. The running "cost so far" z is then:

$$z_n = \sum_{k=0}^{n-1} c(x_k) h$$

Discrete-time Discretized OCP

$$
\begin{aligned}
&\text{minimize} \quad c(x_T) + \sum_{t=0}^{T-1} c(x_t, u_t)h \\
&\text{such that} \quad x_{t+1} = x_t + hf(x_t, u_t) \quad t = 0, \ldots, T-1 \\
&\text{given} \quad x_0
\end{aligned}
$$

Bellman Optimality Conditions on the Discretized-OCP

Substituing the discretized integral cost and dynamics into the discrete Bellman optimality conditions, we get:

$$\tilde{J}(x, nh) = c(x)$$
$$\tilde{J}(x, kh) = \min_{u \in \mathcal{U}} \left\{ c(x, u) + \tilde{J}(x + f(x, u), (k+1)h) \right\}$$

We will now write $\tilde{J}(x + f(x, u), (k+1)h)$ using the Taylor series.

Taylor Approximation

Taking the Taylor series approximation at (x, kh), we get:

$$\tilde{J}(x + hf(x,u), kh + h) = \tilde{J}(x, kh) + hD_2\tilde{J}(x, kh) + hD_1\tilde{J}(x, kh)f(x,u) + o(h)$$

where $\lim_{h\to 0} o(h)/h = 0$.

Why? The first-order Taylor approximation of a multivariate function $f(x, y)$ taken at (a, b) and evaluated at $(x + a, y + b)$ is

$$\begin{aligned} f(x+a, y+b) &\approx f(a,b) + \begin{pmatrix} D_1 f(a,b) & D_2 f(a,b) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \\ &= f(a,b) + D_1 f(a,b)x + D_2 f(a,b)y \end{aligned}$$

Taylor + DP

Plugging the Taylor approximation back in into the DP equations:

$$\tilde{J}(x, kh) = \min_{u \in \mathcal{U}} \left\{ c(x, u) + \tilde{J}(x, kh) + hD_2\tilde{J}(x, kh) + hD_1\tilde{J}(x, kh)f(x, u) + o(h) \right\}$$
$$\Leftrightarrow 0 = \min_{u \in \mathcal{U}} \left\{ c(x, u) + hD_2\tilde{J}(x, kh) + hD_1\tilde{J}(x, kh)f(x, u) \right\} \quad .$$

Because $J(x, kh)$ doesn't depend on u, we can pull it out of the min.

Finally, dividing by h and taking the limit as $h \to 0$:

$$\lim_{k \to \infty, h \to 0, kh = t} \tilde{J}(kh, x, t) = J(t, x) \quad \text{for all } x, t \quad .$$

We then recover the HJB equations:

$$0 = \min_{u \in \mathcal{U}} \{ c(x, u) + D_2J(x, t) + hD_1J(x, t)f(x, u) \} \quad .$$

HJB as a PDE

Theorem (Sufficiency). Let $V(x,t)$ be a solution to the PDE:

$$0 = \min_{u \in \mathcal{U}} \left\{ c(x,u) + D_2 V(x,t) + D_1 V(x,t) f(x,u) \right\} \text{ for all } x, t \quad ,$$

and boundary condition $V(x, t_f) = c(x)$ for all x. Then V is the cost-to-go function J and an optimal policy $\mu(x,t)$ can be obtained by minimizing the expression above given $V(x,t)$.

> ⚠ Potential issue: We assumed that J is differentiable, but this may not be the case, and we may not be able to solve the corresponding HJB equations. But if we happen to do find a solution, analytically, or numerically, then we're in good shape!