CS 295: Optimal Control and Reinforcement Learning

Winter 2020

Lecture 3: Optimal Control

Roy Fox

Department of Computer Science

Bren School of Information and Computer Sciences

University of California, Irvine

1 Continuous time

Consider a continuous-time MDP with time in a finite interval $t \in [0, T]$, state vector $x_t \in \mathbb{R}^n$, and control vector $u_t \in \mathbb{R}^m$. Let's assume deterministic dynamics, such that the velocity $\partial_t x_t = \dot{x}_t \in \mathbb{R}^n$ is given by $\dot{x}_t = f_t(x_t, u_t)$. Let the instantaneous reward be $r_t(x_t, u_t) \in \mathbb{R}$, and we'd like to maximize the total cost

$$\mathcal{J}(u) = \int_0^T r_t(x_t, u_t) dt,$$

subject to the trajectory x following the dynamics f for the control u, and also having a given fixed start and end positions x_0, x_T.

Following the work of Pontryagin, we can consider the Lagrangian of this optimization problem, with a Lagrange multiplier $\nu_t \in \mathbb{R}^n$

$$\mathcal{J}(u) = \int_0^T (r_t + \nu_t(f_t - \dot{x}_t)) dt.$$

Integrating in parts, we get

$$-\int_0^T \nu_t \dot{x}_t dt = -\nu_T x_T + \nu_0 x_0 + \int_0^T x_t \dot{\nu}_t dt,$$

so that

$$\mathcal{J}(u) = \int_0^T (r_t + \nu_t f_t + x_t \dot{\nu}_t) dt - \nu_T x_T + \nu_0 x_0.$$

Now consider an optimal control u^*, and a perturbed control $u(\epsilon) = u^* + \epsilon h$, with h a fixed perturbation and ϵ its variable magnitude. We have

$$\mathcal{J}(\epsilon) = \int_0^T (r_t(x_t(\epsilon), u_t(\epsilon)) + \nu_t f_t(x_t(\epsilon), u_t(\epsilon)) + x_t(\epsilon) \dot{\nu}_t) dt - \nu_T x_T + \nu_0 x_0.$$

Since $\mathcal{J}(\epsilon)$ is maximized at $\epsilon = 0$, we have

$$0 = \partial_\epsilon \mathcal{J}(\epsilon) = \int_0^T ((\partial_{x_t} r_t + \nu_t \, \partial_{x_t} f_t + \dot{\nu}_t) \, \partial_\epsilon x_t + (\partial_{u_t} r_t + \nu_t \, \partial_{u_t} f_t) h) dt.$$

For this to hold for any perturbation, we need

$$\dot{\nu}_t = -(\partial_{x_t} r_t + \nu_t \, \partial_{x_t} f_t) \tag{1}$$

$$\partial_{u_t} r_t + \nu_t \, \partial_{u_t} f_t = 0. \tag{2}$$

If we consider the Hamiltonia

$$\mathcal{H}_t(x_t, u_t, \nu_t) = r_t(x_t, u_t) + \nu_t f_t(x_t, u_t),$$

we get that

$$\begin{aligned} \dot{x}_t &= \partial_{\nu_t} \mathcal{H}_t && \text{(by } \dot{x}_t = f_t(x_t, u_t)) \\ \dot{\nu}_t &= -\,\partial_{x_t} \mathcal{H}_t && \text{(by (1))} \\ \partial_{u_t} \mathcal{H}_t &= 0 && \text{(by (2))}. \end{aligned}$$

We return to the Hamiltonian in the context of discrete time in Section 3

2 Linear–Quadratic Regulation (LQR)

We now turn to discrete time, but still consider continuous state and action spaces. Such spaces are hard to deal with in practice. It's not always clear how to even represent, let alone optimize, the relevant functions over these spaces.

In this lecture and the next we'll see a family of models over continuous spaces with some particularly nice properties. It arises naturally in theoretical physics and in engineering applications. It gives insight into the behavior of dynamic systems. And it will be computationally easy to represent and optimize.

As in the previous section, we start with a fully observable, deterministic dynamics model. This time we consider special (time-invariant) transitions that are linear, that is $x_{t+1} = f(x_t, u_t) = Ax_t + Bu_t$, with some matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. We also consider special costs that are quadratic, that is $c(x_t, u_t) = \frac{1}{2} x_t^\intercal Q x_t + \frac{1}{2} u_t^\intercal R u_t$, with some positive semidefinite $0 \preceq Q \in \mathbb{R}^{n \times n}$ and some positive definite $0 \prec R \in \mathbb{R}^{m \times m}$. The linear dynamics and quadratic costs is where this model gets its name.

Sometimes this model can be a good approximation of more general dynamics and costs, and we can simply take the first-order approximation of the dynamics and the second-order approximation of the costs. In this case, the transitions and the cost could become time-variant, that is, we'd have A_t, B_t, Q_t and R_t vary with time. We would also add lower-order terms: a constant drift in f and linear and constant cost terms in c. These extensions complicate the introduction but don't add much insight, so we'll ignore them here.

We say that a symmetric matrix Q is *positive semidefinite*, and denote it $Q \succeq 0$, if for any $x \in \mathbb{R}^n$ we have $x^\intercal Q x \geqslant 0$. We say that a symmetric matrix R is *positive definite*, and denote it $R > 0$, if for any $u \in \mathbb{R}^m$ other than 0 we have $u^\intercal R u > 0$. If either of these don't hold the model becomes degenerate. If Q or R are not positive semidefinite, then there's a direction of x or u where the cost diverges to $-\infty$. If R is positive semidefinite but not positive definite, then there's a subspace of the control space with no cost, where we can optimize for the long term without considering the immediate cost. This is uninteresting, and it complicates the math a bit, so we'll require R to be positive definite.

The total cost in state x_0 and with control sequence $u_0, u_1, \dots$ is

$$\begin{aligned}\mathcal{J}(x_0, u) &= \sum_t c(x_t, u_t) = \sum_t (\tfrac{1}{2} x_t^\intercal Q x_t + \tfrac{1}{2} u_t^\intercal R u_t) \\ \text{s.t. } x_{t+1} &= f(x_t, u_t) = A x_t + B u_t \qquad \forall t \geqslant 0.\end{aligned}$$

Often we use a finite horizon here, sometimes an infinite horizon. It's uncommon to see a discount in this context. In the specific case here that the dynamics are deterministic, we can talk about an episodic horizon, because the state $x = 0$ is absorbing, in a sense: once it's reached no further control is needed, and the cost-to-go (another term for future return, i.e. total future cost) is 0. If there exists a control policy that reaches $x = 0$ from any initial state, we call the dynamics (A, B) *controllable.*

We can unfold the state recursion as

$$x_t = A^t x_0 + A^{t-1} B u_0 + \cdots + A B u_{t-2} + B u_{t-1},$$

or

$$x_t - A^t x_0 = \begin{bmatrix} B & AB & \cdots & A^{t-1}B \end{bmatrix} \begin{bmatrix} u_{t-1} \\ u_{t-2} \\ \vdots \\ u_0 \end{bmatrix}.$$

Suppose A has rank n, so that the process doesn't degenerate into a subspace. Then the uncontrolled dynamics can lead to any state $A^t x_0$. For there to exist a sequence $u_0, \dots, u_{t-1}$ that leads to $x_t = 0$, the entire space needs to be spanned by the columns of the *controllability matrix*

$$\mathcal{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}.$$

There's no point in taking more than n terms in this matrix, because the matrix A satisfies the *characteristic polynomial* of degree n, by the Cayley-Hamilton theorem.

We can use Dynamic Programming to find an optimal control policy, one that minimizes the total cost $\mathcal{J}$. To do that, we need to be able to represent the optimal cost-to-go $\mathcal{J}_t^*$ in time t in a finite way. Since we were able to express x_t as a linear function of x_0 and the control sequence, so that each $c(x_t, u_t)$ is quadratic in those variables, so is the total cost $\mathcal{J}(u)$. But we need to represent the *optimal* cost-to-go $\mathcal{J}_t^*(x_t)$, and we may make an educated guess that this function is quadratic as well. We'll see in a moment that this is true, and part of what makes this setting solvable.

The Bellman equation for T-step finite-horizon LQR is

$$\mathcal{J}_t^*(x_t) = \min_{u_t} (c(x_t, u_t) + \mathcal{J}_{t+1}^*(f(x_t, u_t))) = \min_{u_t} (\tfrac{1}{2} x_t^\intercal Q x_t + \tfrac{1}{2} u_t^\intercal R u_t + \mathcal{J}_{t+1}^*(A x_t + B u_t)).$$

In the continuous-time limit, this equation is called the Hamilton–Jacobi–Bellman equation. In the discrete-time case, we'll solve it at the same time that we prove by induction that $\mathcal{J}_t^*$ is quadratic. This holds by definition for $\mathcal{J}_T^* = 0$. Now suppose it holds for $\mathcal{J}_{t+1}^*$, that is

$$\mathcal{J}_{t+1}^*(x_{t+1}) = \tfrac{1}{2} x_{t+1}^\intercal S_{t+1} x_{t+1},$$

for some positive semidefinite Hessian $0 \preceq S_{t+1} \in \mathbb{R}^{n \times n}$. So

$$\mathcal{J}_t^*(x_t) = \min_{u_t} \left(\tfrac{1}{2} x_t^\intercal Q x_t + \tfrac{1}{2} u_t^\intercal R u_t + \tfrac{1}{2} (A x_t + B u_t)^\intercal S_{t+1} (A x_t + B u_t) \right).$$

The optimal control is the u_t for which the objective has gradient 0 with respect to $u_t^\intercal$

$$Ru + B^\intercal S_{t+1}(A x_t + B u_t) = 0,$$

which we rearrange to get

$$u_t = -(R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A x_t.$$

We plug this back into the target function to get

$$\begin{aligned}
\mathcal{J}_t^*(x_t) &= \tfrac{1}{2} x_t^\intercal (Q + A^\intercal S_{t+1} A) x_t + u_t^\intercal B^\intercal S_{t+1} A x_t + \tfrac{1}{2} u_t^\intercal (R + B^\intercal S_{t+1} B) u_t \\
&= \tfrac{1}{2} x_t^\intercal (Q + A^\intercal S_{t+1} A) x_t + u_t^\intercal B^\intercal S_{t+1} A x_t - \tfrac{1}{2} u_t^\intercal B^\intercal S_{t+1} A x_t \\
&= \tfrac{1}{2} x_t^\intercal (Q + A^\intercal S_{t+1} A) x_t - \tfrac{1}{2} x_t^\intercal A^\intercal S_{t+1} B (R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A x_t \\
&= \tfrac{1}{2} x_t^\intercal S_t x_t,
\end{aligned}$$

with the symmetric matrix

$$S_t = Q + A^\intercal (S_{t+1} - S_{t+1} B (R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1}) A.$$

The form of this equation is called a discrete-time *Ricatti equation*.

Since we found that $\mathcal{J}_t^*(x_t)$ is quadratic, it shouldn't be surprising that its Hessian S_t is independent of x_t, and that it can be computed without knowing any of the states or the control signals, using the above backward recursion. To show that S_t is also positive semidefinite, we could use the Woodbury matrix identity (not shown here), or just note that, like the immediate cost, the optimal cost-to-go cannot decrease without bound in any direction of the state space.

We see that in the finite-horizon case the optimal time-variant control policy is also linear

$$u_t = L_t x_t,$$

with the *feedback gain* matrix

$$L_t = -(R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A.$$

3 The co-state and the Hamiltonian

For deterministic dynamics $x_{t+1} = f(x_t, u_t)$ and cost rate $c(x_t, u_t)$, we can ask how sensitive the total cost $\mathcal{J}(u)$ is to small perturbations in the state. More precisely, we can look at the gradient of

$$\mathcal{J}_t(x_t, u) = c(x_t, u_t) + \mathcal{J}_{t+1}(f(x_t, u_t), u)$$

with respect to x_t. This gradient is a row vector called the *co-state* $\nu_t \in \mathbb{R}^n$, and is given by

$$\nu_t = \nabla_{x_t}\mathcal{J}_t = \nabla_{x_t}c_t + \nabla_{x_{t+1}}\mathcal{J}_{t+1} \cdot \nabla_{x_t}f_t = \nabla_{x_t}c_t + \nu_{t+1}\nabla_{x_t}f_t, \tag{3}$$

where $\nabla_{x_t}f_t \in \mathbb{R}^{n\times n}$ is the Jacobian of the dynamics. This is a linear backward recursion, initialized by

$$\nu_T = 0.$$

We can now define the discrete-time Hamiltonian

$$\mathcal{H}_t = c(x_t, u_t) + \nu_{t+1}f(x_t, u_t).$$

The Hamiltonian in control theory is related to the one in physics, but is not the same. Our Hamiltonian is, in a sense, a first-order approximation of the total cost $\mathcal{J}_t$. It has the correct immediate-cost term $c(x_t, u_t)$, and an approximation of the future cost $\mathcal{J}_{t+1}$ to first order in $x_{t+1} = f(x_t, u_t)$. The Hamiltonian defines at the same time the backward dynamics of the co-state

$$\nu_t = \nabla_{x_t}\mathcal{H}_t, \tag{4}$$

by (3), and the forward dynamics of the state

$$x_{t+1} = \nabla_{\nu_{t+1}}\mathcal{H}_t. \tag{5}$$

We can also write our total cost objective as

$$\mathcal{J} = \sum_{t=0}^{T-1}(\mathcal{H}_t - \nu_{t+1}x_{t+1}).$$

In this equation, when the Hamiltonian is plugged in, the co-state ν_{t+1} plays the role of a *Lagrange multiplier* for the constraint $x_{t+1} = f(x_t, u_t)$. It shows that

$$\nabla_{u_t}\mathcal{J} = \nabla_{u_t}\mathcal{H}_t,$$

so that the optimal control is obtained when for all $t \geqslant 0$

$$\nabla_{u_t}\mathcal{H}_t = 0. \tag{6}$$

The system of the nT variables x_t, the mT variables u_t, and the nT variables ν_t may be determined by the nT equations (4), the nT equations (5), and the mT equations (6). In general these equations are nonlinear, and there may be more than one solution, corresponding to local minima.

In the linear–quadratic case, the Hamiltonian is quadratic

$$\mathcal{H}_t = \tfrac{1}{2}x_t^\intercal Q x_t + \tfrac{1}{2}u_t^\intercal R u_t + \nu_{t+1}(Ax_t + Bu_t),$$

and the equations get the linear form

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \\ \nu_t &= \nu_{t+1}A + x_t^\intercal Q \\ Ru_t + B^\intercal \nu_{t+1}^\intercal &= 0, \end{aligned}$$

with the mixed (initial and terminal) boundary conditions of a given x_0 and $\nu_T = 0$. Let's show by induction that the unique solution to these equations is

$$\begin{aligned} \nu_t^\intercal &= S_t x_t \\ u_t &= L_t x_t. \end{aligned}$$

This holds for time T, since $S_T = 0$. Now assume it holds for time $t+1$, and prove for time t. We have

$$0 = Ru_t + B^\intercal \nu_{t+1}^\intercal = Ru_t + B^\intercal S_{t+1} x_{t+1} = (R + B^\intercal S_{t+1} B)u_t + B^\intercal S_{t+1} A x_t,$$

so $u_t = L_t x_t$ and

$$\begin{aligned} \nu_t^\intercal &= A^\intercal \nu_{t+1} + Qx_t = A^\intercal S_{t+1} x_{t+1} + Qx_t \\ &= (Q + A^\intercal S_{t+1} A)x_t + A^\intercal S_{t+1} B u_t \\ &= (Q + A^\intercal S_{t+1} A - A^\intercal S_{t+1} B(R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A)x_t = S_t x_t, \end{aligned}$$

as required.