Ivan Papusha 05/15/2015 12:04

HJB from DP

1 Dynamic Programming

The basic control problem with horizon length N is

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\left\{\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) + g_N(x_N)\right\} \\ \text{subject to} & x_{k+1} = f_k(x_k, u_k, w_k), \quad k = 0, \dots, N-1 \\ & u_k \in U_k(x_k), \quad k = 0, \dots, N-1, \end{array}$$

where the decision variables are the states $x_0, \dots, x_N \in \mathbf{R}^n$ and the control inputs $u_0, \dots, u_{N-1} \in \mathbf{R}^m$, and the expectation is over (random) disturbances $w_0, \dots, w_{N-1} \in \mathbf{R}^q$. Here $f_k : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^q \to \mathbf{R}^n$ are the state transition functions, $g_k : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^q \to \mathbf{R}$ are the stage costs for each $k = 0, \dots, N-1$, and $g_N : \mathbf{R}^n \to \mathbf{R}$ is a terminal cost. The sets $U_k(x_k) \subseteq \mathbf{R}^m$ denote state-dependent control constraints.

For every initial state x_0, the optimal cost $J^\star(x_0)$ of the basic problem is given by $J_0(x_0)$ in the last step of the following algorithm [Ber05, §1.3], which proceeds backward from period $N-1$ to period 0:

$$\begin{aligned} J_N(x_N) &= g_N(x_N), \\ J_k(x_k) &= \min_{u_k \in U_k(x_k)} \mathbf{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + J_{k+1}\big(f_k(x_k, u_k, w_k)\big) \right\}, \\ & \qquad \text{for } k = 0, \dots, N-1. \end{aligned}$$

The optimal policy consists of choosing a minimizing control action $u_k^\star$,

$$\begin{aligned} u_k^\star \in \operatorname*{argmin}_{u_k \in U_k(x_k)} \mathbf{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + J_{k+1}\big(f_k(x_k, u_k, w_k)\big) \right\}, \\ \text{for } k = 0, \dots, N-1. \end{aligned}$$

2 Deterministic Hamilton–Jacobi–Bellman

The basic continuous-time control problem with horizon length T is

$$\begin{array}{ll} \text{minimize} & \int_0^T g(x(t), u(t))\, dt + h(x(T)) \\ \text{subject to} & \dot{x}(t) = f(x(t), u(t)), \quad 0 \le t \le T, \\ & x(0) = x_0 \end{array}.$$

If $V(t,x)$ a continuously differentiable (in t and x) solution to the HJB equation

$$
\begin{aligned}
-\frac{\partial}{\partial t}V(t,x) &= \min_{u\in U}\left\{g(x,u)+\nabla_x V(t,x)^T f(x,u)\right\}, \quad \text{for all } t,x,\\
V(T,x) &= h(x), \quad \text{for all } x,
\end{aligned}
$$

then it is the optimal cost-to-go and a control policy obtained using the minimization is optimal. The function $V : [0,T]\times \mathbf{R}^n \to \mathbf{R}$ is called the value function.

2.1 Derivation Using Dynamic Programming

The following derivation is due to [Ber05, §3.2]. Divide the time horizon $[0,T]$ into N pieces using the discretization interval $\delta = \frac{T}{N}$, and define

$$
x_k \triangleq x(k\delta), \quad u_k \triangleq u(k\delta), \quad k=0,\dots,N.
$$

The first order approximations to the continuous system and its cost function are

$$
\begin{aligned}
x_{k+1} &= x_k + f(x_k,u_k)\cdot\delta\\
J &= \sum_{k=0}^{N-1} g(x_k,u_k)\cdot\delta + h(x_N).
\end{aligned}
$$

Let $J^\star(t,x)$ be the optimal cost-to-go at time t and state x for the continuous-time problem, and $J_d^\star(t,x)$ be the optimal cost-to-go for the discrete-time approximation. The DP equations are

$$
\begin{aligned}
J_d^\star(N\delta,x) &= h(x),\\
J_d^\star(k\delta,x) &= \min_{u\in U}\left\{g(x,u)\cdot\delta + J_d^\star\big((k+1)\cdot\delta, x+f(x,u)\cdot\delta\big)\right\},\\
&\qquad \text{for } k=0,\dots,N-1.
\end{aligned}
$$

Expanding $J_d^\star$ as a Taylor series around $(k\delta,x)$ we obtain

$$
\begin{aligned}
J_d^\star\big((k+1)\cdot\delta, x+f(x,u)\cdot\delta\big) &= J_d^\star(k\delta,x)\\
&\quad + \nabla_t J_d^\star(k\delta,x)\cdot\delta + \nabla_x J_d^\star(k\delta,x)^T f(x,u)\cdot\delta + o(\delta),
\end{aligned}
$$

where $\lim_{\delta\to 0} o(\delta)/\delta = 0$. After substituting back into the DP equations,

$$
\begin{aligned}
J_d^\star(k\delta,x) = \min_{u\in U}\big\{&g(x,u)\cdot\delta + J_d^\star(k\delta,x)\\
&+ \nabla_t J_d^\star(k\delta,x)\cdot\delta + \nabla_x J_d^\star(k\delta,x)^T f(x,u)\cdot\delta + o(\delta)\big\}.
\end{aligned}
$$

We then cancel $J_d^\star(k\delta, x)$ from both sides, divide by δ, and take the limit as $\delta \to 0$. Assuming the discrete-time cost-to-go function yields in the limit its continuous-time counterpart, *i.e.*,

$$\lim_{k\to\infty,\,\delta\to 0,\,k\delta=t} J_d^\star(k\delta, x) = J^\star(t, x), \quad \text{for all } t, x,$$

we arrive at the Hamilton–Jacobi–Bellman equation for the optimal cost-to-go $J^\star(t, x)$,

$$\begin{aligned} 0 &= \min_{u\in U} \left\{ g(x,u) + \nabla_t J^\star(t,x) + \nabla_x J^\star(t,x)^T f(x,u) \right\}, \quad \text{for all } t, x, \\ h(x) &= J^\star(T, x), \quad \text{for all } x. \end{aligned}$$

From here, let $V(t, x) \triangleq J^\star(t, x)$, and subtract the terms that do not depend on u out of the minimum to obtain the HJB equation.

3 Stochastic Hamilton–Jacobi–Bellman

The basic stochastic control problem with horizon length T is

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\left\{ \displaystyle\int_0^T g(x_t, u_t)\, dt + h(x_T) \right\} \\ \text{subject to} & dx_t = f(x_t, u_t)\, dt + \sigma(x_t)\, dW_t, \quad 0 \le t \le T, \\ & x|_{t=0} = x_0. \end{array}$$

The dynamics of the state x_t are governed by an Itō drift-diffusion process in $\mathbf{R}^n$, where $\{W_t \mid t \ge 0\}$ is a standard Wiener process in $\mathbf{R}^q$ and $\sigma : \mathbf{R}^n \to \mathbf{R}^{n\times q}$ is a noise feedthrough function. The cases $q = n$ and $q = 1$ are common. The stochastic HJB equation is

$$\begin{aligned} -\frac{\partial}{\partial t} V(t,x) &= \min_{u\in U} \Big\{ g(x,u) + \nabla_x V(t,x)^T f(x,u) \\ &\qquad\qquad + \frac{1}{2}\,\mathbf{Tr}\left(\nabla_x^2 V(t,x) \cdot \sigma(x)\sigma(x)^T \right) \Big\}, \quad \text{for all } t, x, \\ V(T, x) &= h(x), \quad \text{for all } x. \end{aligned}$$

Note the additional Hessian of the value function, which does not appear in the non-stochastic setting.

3.1 Derivation Using Dynamic Programming

The extra Hessian term comes from Itō's formula. The definitive sources are [FS06, TBS10] with an informal derivation following the same lines as in §2.1. The first order approximation to the continuous system looks slightly different

$$x_{k+1} = x_k + f(x_k, u_k) \cdot \delta + \sigma(x_k) \cdot \epsilon_k \cdot \delta^{1/2},$$

where $\epsilon_k \sim \mathcal{N}(0, I_q)$ are iid standard normal variables on $\mathbf{R}^q$ inherited from the Wiener process. The DP equations are

$$\begin{aligned} J_d^\star(N\delta, x) &= h(x), \\ J_d^\star(k\delta, x) &= \min_{u \in U} \mathbf{E} \left\{ g(x,u) \cdot \delta + J_d^\star\big((k+1) \cdot \delta, x + f(x,u) \cdot \delta + \sigma(x) \cdot \epsilon \cdot \delta^{1/2}\big) \right\} \\ &= \min_{u \in U} \left\{ g(x,u) \cdot \delta + \mathbf{E}\, J_d^\star\big((k+1) \cdot \delta, x + f(x,u) \cdot \delta + \sigma(x) \cdot \epsilon \cdot \delta^{1/2}\big) \right\}, \end{aligned}$$

Expand $J_d^\star$ as a Taylor series around $(k\delta, x)$ to the second order:

$$\begin{aligned} &J_d^\star\big((k+1) \cdot \delta, x + f(x,u) \cdot \delta + \sigma(x) \cdot \epsilon \cdot \delta^{1/2}\big) \\ &\quad = J_d^\star(k\delta, x) + \nabla_t J_d^\star(k\delta, x)\delta + \nabla_x J_d^\star(k\delta, x)^T \left(f(x,u)\delta + \sigma(x)\epsilon\delta^{1/2} \right) \\ &\qquad + \frac{1}{2} \mathbf{Tr} \left(\nabla_x^2 J_d^\star(k\delta, x) \cdot \sigma(x)\epsilon\epsilon^T \sigma(x)^T \delta \right) + o(\delta^{3/2}) \end{aligned}$$

Using $\mathbf{E}\,\epsilon = 0$ and $\mathbf{E}\,\epsilon\epsilon^T = I_q$, take the expected value of both sides to obtain

$$\begin{aligned} &\mathbf{E}\, J_d^\star\big((k+1) \cdot \delta, x + f(x,u) \cdot \delta + \sigma(x) \cdot \epsilon \cdot \delta^{1/2}\big) \\ &\quad = J_d^\star(k\delta, x) + \nabla_t J_d^\star(k\delta, x)\delta + \nabla_x J_d^\star(k\delta, x)^T f(x,u)\delta \\ &\qquad + \frac{1}{2} \mathbf{Tr} \left(\nabla_x^2 J_d^\star(k\delta, x) \cdot \sigma(x)\sigma(x)^T \delta \right) + o(\delta^{3/2}). \end{aligned}$$

Finally, substitute this expression back into the DP equations, subtract $J_d^\star(k\delta, x)$ from both sides, divide by δ, and take the limit as $\delta \to 0$ to obtain stochastic HJB equations, *cf.* [TBS10, eq. 5].

$$\begin{aligned} -\nabla_t J^\star(t,x) &= \min_{u \in U} \Big\{ g(x,u) + \nabla_x J^\star(t,x)^T f(x,u) \\ &\qquad + \frac{1}{2} \mathbf{Tr} \left(\nabla_x^2 J^\star(t,x) \cdot \sigma(x)\sigma(x)^T \right) \Big\}, \quad \text{for all } t, x, \\ h(x) &= J^\star(T, x), \quad \text{for all } x. \end{aligned}$$

References

[Ber05] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, 3rd edition, 2005.

[FS06] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, 2nd edition, 2006.

[Mol12] Benjamin Moll. Stochastic HJB equations, Kolmogorov forward equations. Economics 521 Lecture 5, available at `http://www.princeton.edu/~moll/ECO521Web/Lecture5_ECO521_web.pdf`, 2012.

[TBS10] Evangelos A. Theodorou, Jonas Buchli, and Stefan Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research*, 11:3137–3181, 2010.