Optimal Control, Lecture 6: VI and PI for RL

Anders Hansson

Division of Automatic Control
Linkoping University

Contents

Optimal Control Problem

$$\begin{array}{ll} \text{minimize} & \sum_{k=0}^{\infty} \gamma^k f(x_k, u_k) \\ \text{subject to} & x_{k+1} = F(x_k, u_k), \quad k \in \mathbf{Z}_+ \end{array}$$

with variables $(u_0, x_1, \ldots)$, where x_0 is given.

Bellman Equation

Assume $f(0,0) = 0$, $F(0,0) = 0$ and that f is strictly positive definite. If there exists a strictly positive definite and quadratically bounded V such that the Bellman equation

$$V(x) = \min_{u \in U(x)} \{f(x,u) + \gamma V(F(x,u))\}$$

holds, then

- (a) $J^*(x) = V(x)$
- (b) The minimizing argument in the Bellman equation is an optimal feedback that results in a globally convergent closed loop system if γ is sufficiently close to one.

The Q-Function

Let $Q(x,u) = f(x,u) + \gamma V(F(x,u))$. Then Bellman equation reads

$$V(x) = \min_u Q(x,u),$$

and

$$\gamma V(F(x,\bar{u})) = \min_u \gamma Q(F(x,\bar{u}),u).$$

By adding $f(x,\bar{u})$ to both sides we get

$$Q(x,\bar{u}) = f(x,\bar{u}) + \min_u \gamma Q\left(F(x,\bar{u}),u\right). \tag{1}$$

The Bellman Q-Operator and VI

Let the Bellman Q-operator be

$$T_Q(Q)(x, \bar{u}) = f(x, \bar{u}) + \min_u \gamma Q\left(F(x, \bar{u}), u\right). \tag{2}$$

Define the VI

$$Q_{k+1} = T_Q(Q_k) \tag{3}$$

with boundary condition $Q_0(x, u) = f(x, u)$.

You will show in Exercise 11.6 that $Q_k(x, u)$ converges to $Q(x, u)$ satisfying (1) as $k \to \infty$.

Generalize VI for Q-Function

Let

$$e(Q) = Q - T_Q(Q),$$

Then (1) is equivalent to as $e(Q) = 0$.

Apply the root finding algorithm

$$Q_{k+1} = Q_k - t_k e(Q_k), \quad k \in \mathbf{Z}_+ \tag{4}$$

- You can initialize with $Q_0 = f$, but there are better ways.
- The step lengths t_k should satisfy $t_k \in (0, 1]$.
- Recover VI for $t_k = 1$.

Proof of convergence on white board.

Q-Learning

It is possible to instead of in each iteration k consider all values of (x, u) only consider one sample (x_k, u_k) at a time.

Theses samples could be generated in a cyclic order or in a randomized cyclic order such that each sample is visited infinitely many times.

We assume that (x, u) belongs to a finite set. Then it holds that

$$\begin{aligned} Q_{k+1}(x_k, u_k) &= Q_k(x_k, u_k) \\ &\quad - t_k \left[Q(x_k, u_k) - f(x_k, u_k) - \min_u \gamma Q(F(x_k, u_k), u) \right] \end{aligned}$$

converges to a solution of $e(Q) = 0$ as k goes to infinity when $t_k \in (0, 1]$ and $\gamma \in (0, 1)$.

Policy Iteration

- Reinforcement learning based on PI is called *self-learning*.
- The policy evaluation step is referred to as a *critic*
- The policy improvement is referred to as an *actor*.
- These type of methods are called *actor-critic* methods.
- In case parametric approximations using ANNs are involved the actor and critic are called *actor networks* and *critic networks*, respectively.

Recap of PI using Value Function

Bellman policy operator:

$$T_\mu(V)(x) = f(x, \mu(x)) + \gamma V\left(F(x, \mu\left(x\right))\right) \tag{5}$$

for a given function μ.

Iterate starting with initial μ_0:

1. Solve (policy evaluation step)

$$V_k(x) = T_{\mu_k}(V_k)(x), \tag{6}$$

2. Solve (policy improvement step)

$$\mu_{k+1}(x) = \underset{u \in U(x)}{\operatorname{argmin}} \left\{ f(x, u) + \gamma V_k\left(F(x, u)\right)\right\}. \tag{7}$$

Policy Iteration using Q-Function

Let $Q_k(x,u) = f(x,u) + \gamma V_k(F(x,u))$. Then

$$V_k(x) = Q_k\left(x, \mu_k(x)\right)$$

from (6), and therefore

$$V_k(F(x,u)) = Q_k\left(F(x,u), \mu_k(F(x,u))\right).$$

Multiply with γ and add $f(x,u)$ to obtain that Q_k is the solution of

$$Q_k\left(x,u\right) = f\left(x,u\right) + \gamma Q_k\left(F\left(x,u\right), \mu_k(F(x,u))\right). \tag{8}$$

This is the policy evaluation step in terms of the Q-function.

We then obtain a new feedback policy by solving

$$\mu_{k+1}(x) = \underset{u}{\operatorname{argmin}}\, Q_k\left(x,u\right), \tag{9}$$

which is the policy improvement step in terms of the Q-function.

These iterations results in the same solution as (6–7).

LQ Control

$$\begin{array}{ll} \text{minimize} & \sum_{k=0}^{\infty} \gamma^k \left(x_k^T S x_k + u_k^T R u_k \right) \\ \text{subject to} & x_{k+1} = A x_k + B u_k \\ & x_0 \text{ given} \end{array} \quad (10)$$

We guess that

$$Q_k(x, u) = \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} U_k & W_k \\ W_k^T & V_k \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}$$

for some

$$\begin{bmatrix} U_k & W_k \\ W_k^T & V_k \end{bmatrix} \in \mathbf{S}_+^{m+n},$$

where $V_k \in \mathbf{S}_{++}^m$. It then follows from (9) that

$$\mu_k(x) = -L_{k+1} x,$$

where $L_{k+1} = V_k^{-1} W_k^T$.

LQ Control ctd.

The recursion for Q_k in (8) is seen to be satisfied if

$$\begin{bmatrix} U_k & W_k \\ W_k^T & V_k \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} + \gamma \begin{bmatrix} A & B \end{bmatrix}^T \begin{bmatrix} I \\ -L_k \end{bmatrix}^T \begin{bmatrix} U_k & W_k \\ W_k^T & V_k \end{bmatrix} \begin{bmatrix} I \\ -L_k \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix}$$

for a given L_k. This is an algebraic Lyapunov equation which has a positive semidefinite solution since

$$\begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix}$$

is positive semidefinite. This assumes that

$$\sqrt{\gamma} \begin{bmatrix} I \\ -L_k \end{bmatrix} \begin{bmatrix} A & B \end{bmatrix}$$

has all its eigenvalues strictly inside the unit disc. This is true if $\sqrt{\gamma}(A - BL_k)$ has all its eigenvalues strictly inside the unit disc by Exercise 11.1.

Critic Network

It holds that (8) implies

$$\begin{aligned} Q_k(x_0,u_0) &= f\left(x_0,u_0\right) + \gamma Q_k\left(F\left(x_0,u_0\right), \mu_k\left(F\left(x_0,u_0\right)\right)\right) \\ &= f\left(x_0,u_0\right) + \gamma Q_k\left(x_1, \mu_k\left(x_1\right)\right) \\ &= f\left(x_0,u_0\right) + \gamma f\left(x_1, \mu_k(x_1)\right) + \gamma^2 Q_k\left(x_2, \mu_k\left(x_2\right)\right) \\ &\vdots \\ &= f\left(x_0,u_0\right) + \sum_{i=1}^{N-1} \gamma^i f\left(x_i, \mu_k(x_i)\right) + \gamma^N Q_k\left(x_N, \mu_k\left(x_N\right)\right), \end{aligned}$$

where $x_{i+1} = F(x_i, \mu_k(x_i))$ for $1 \leq i \leq N-1$, and $x_1 = F(x_0, u_0)$.

In case N is large and μ_k is stabilizing we have that x_N is close to zero and that also $Q_k(x_N)$ is close to zero.

Critic Network ctd.

We denote these approximations for different initial values (x^s, u^s) for $1 \leq s \leq r$ as

$$\beta_k^s = f(x^s, u^s) + \sum_{i=1}^{N-1} \gamma^i f\left(x_i, \mu_k(x_i)\right),$$

where $x_{i+1} = F(x_i, \mu_k(x_i))$ for $1 \leq i \leq N-1$, and $x_1 = F(x^s, u^s)$. We then find approximation of Q_k by solving

$$\text{minimize} \quad \tfrac{1}{2} \textstyle\sum_{s=1}^{r} \left(\tilde{Q}(x^s, u^s, a_k) - \beta_k^s\right)^2$$

with variable a_k, where $\tilde{Q}_k$ is an ANN or linear regression.

After this we use the following exact policy improvement step

$$\mu_{k+1}(x) = \underset{u}{\operatorname{argmin}}\, \tilde{Q}\left(x, u, a_k\right). \tag{11}$$

LQ Control

Let $\varphi(x,u) = (x_1^2, x_2^2, u^2, 2x_1x_2, 2x_1u, 2x_2u)$ and

$$\tilde{Q}(x,u,a) = a^T\varphi(x,u),$$

With

$$\begin{bmatrix} \tilde{P} & \tilde{r} \\ \tilde{r}^T & \tilde{q} \end{bmatrix} = \begin{bmatrix} a_1 & a_4 & a_5 \\ a_4 & a_2 & a_6 \\ a_5 & a_6 & a_3 \end{bmatrix}$$

we may write

$$\tilde{Q}_k(x,u,a) = \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} \tilde{P} & \tilde{r} \\ \tilde{r}^T & \tilde{q} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}. \tag{12}$$

Then a_k is the solution to the linear LS problem

$$\text{minimize} \quad \tfrac{1}{2}\textstyle\sum_{s=1}^{r} \left(\varphi^T(x^s,u^s)a - \beta_k^s\right)^2$$

with variable a.

LQ Control ctd.

The solution a_k satisfies the normal equations

$$\Phi_k^T \Phi_k a_k = \Phi_k^T \beta_k,$$

where

$$\Phi_k = \begin{bmatrix} \varphi^T(x^1, u^1) \\ \vdots \\ \varphi^T(x^r, u^r) \end{bmatrix}, \qquad \beta_k = \begin{bmatrix} \beta_k^1 \\ \vdots \\ \beta_k^r \end{bmatrix},$$

whith

$$\beta_k^s = (x^s)^T Sx^s + (u^s)^T Ru^s + \sum_{i=1}^{N-1} \gamma^i \left(x_i^T S x_i + \mu_k(x_i)^T R \mu_k(x_i) \right),$$

where $x_1 = Ax^s + Bu^s$ and $x_{i+1} = Ax_i + B\mu_k(x_i)$ for $1 \leq i \leq N-2$ with initial values x^s, $1 \leq s \leq r$.

It is crucial to choose (x^s, u^s) such that $\Phi_k^T \Phi_k$ is invertible. We realize that we need $r \geq 6$ for this hold.

LQ Control ctd.

The solution to (11) is given by

$$\mu_{k+1}(x) = \operatorname*{argmin}_u \tilde{Q}_k(x, u, a_k) = -\tilde{q}_k^{-1}\tilde{r}_k^T x$$

assuming that $\tilde{q}$ is positive. Here $\tilde{q}_k$ and $\tilde{r}_k$ are defined from a_k. We may hence write

$$\mu_{k+1}(x) = -L_{k+1}x,$$

where $L_{k+1} = \tilde{q}_k^{-1}\tilde{r}_k^T$. It is a good idea to start with some L_0 such that μ_0 is stabilizing.

Linear Programming Formulation

A solution to the Bellman equation for the Q-function can be obtained by solving the Linear Program (LP)

$$\begin{array}{ll} \text{maximize} & \sum_{(x,u)} c(x,u)Q(x,u) \\ \text{subject to} & Q(x,u) \leq f(x,u) + \gamma Q(F(x,u),v), \ \forall (x,u,v) \end{array} \tag{13}$$

where $c(x,u) > 0$ is arbitrary.

- The variables (x,u) has to belong to a finite set.
- The optimization variable is $Q(x,u)$ for all values of x and u in this finite set.
- The LP formulation is often not tractable in general, since there might be many variables and constraints.
- It is possible to approximate $Q(x,u)$ with a linear regression.
- Sampling of constraints may also be used.
- We may use the LP to approximately evaluate a fixed policy μ, which may then be used together with PI.