Lecture #17

Proof: Refer to Ch. 8 of the Book I sent via e-mail.

Why bother about contraction maps?

Proposition: Let F be the class of all functions $z : \{1, \ldots, n\} \mapsto \mathbb{R}$. Define $\|z\|_\infty := \max_{1 \le i \le n} |z(i)|$ and the nonlinear map $T : F \mapsto F$ as

$$(Tz)(i) := \inf_{u \in \mathcal{U}} \left\{ c(i,u) + \beta \sum_{j=1}^{n} z(j) P_{ij}(u) \right\}.$$

Then $T(\cdot)$ is a contraction.

Proof. P. 148 in Ch. 8 of the book I sent.

Recap: Bellman's eqn / DP eqn for Infinite horizon discounted cost

OCP: $\min_{\underline{\gamma} \in \Gamma} \mathbb{E}_w \left[\sum_{k=0}^{+\infty} \beta^k c(\underline{x}_k, \underline{u}_k, \underline{w}_k) \right]$

assume $\underline{w}_k$ iid.

s.t. $\underline{x}_{k+1} = f_k(\underline{x}_k, \underline{u}_k, \underline{w}_k)$.

DP eqn: (nonlinear algebraic eqn)

$$W_\infty(\underline{x}) = \min_{u \in \mathcal{U}} \mathbb{E}\left[c(\underline{x}, \underline{u}, \underline{w}) + \beta W_\infty(f(\underline{x}, \underline{u}, \underline{w})) \right]$$

Special case of Bellman's eqn

Instead of $\underline{x}_{k+1} = f_k(\underline{x}_k, \underline{u}_k, \underline{w}_k)$,

we have $\underline{x}_k \sim \text{Markov}(P(u))$,

$P(u) = [P_{ij}(u)]$

$i, j = 1, \dots, n$

Then Bellman's eqn / DP eqn:

$$W_\infty(i) = \min_{u \in \mathcal{U}} \left[c(i,u) + \beta \sum_{j=1}^{n} p_{ij}(u) W_\infty(j) \right]$$

$$i = 1, \dots, n$$

The proposition says the RHS, thought of as a nonlinear operator $T \underline{W}_\infty$, is a contraction.

Theorem:

(1) There exists unique sol$^{\underline{n}}$ $(W_\infty(1), W_\infty(2), \dots, W_\infty(n))^T$ to the set of nonlinear eq$^{\underline{n}}$s:

$$W_\infty(i) = \inf_{u \in \mathcal{U}} \left\{ c(i,u) + \beta \sum_{j=1}^{n} P_{ij}(u) W_\infty(j) \right\},$$

$1 \le i \le n$.

(2) The answer/unique sol$^{\underline{n}}$ from (1) satisfies:

$$W_\infty(i) = \underbrace{\inf_{\gamma \in \Gamma} \mathbb{E}^{\gamma} \left\{ \sum_{k=0}^{+\infty} \beta^k c(x_k, u_k) \,\middle|\, x_0 = i \right\}}_{\text{cost-to-go}}$$

(3) The map T, defined as

$$Tz(i) = \inf_{u \in U} \left\{ c(i,u) + \beta \sum_{j=1}^{n} P_{ij}(u)\, z(j) \right\}, \quad i = 1, \dots, n$$

is a contraction w.r.t. norm $\|z\| = \max_{1 \le i \le n} |z_i|$

(4) $\lim_{n \to \infty} (T^n z)(i) = W_\infty(i) \quad \forall\, i = 1, \dots, n$

for any z.

(5) If $z(i) = \underbrace{0}_{\equiv} \ \forall i$, then $(T^n z)(i) = W_n(i)$,

$\underline{z} = \text{zeros}(n,1)$

where $W_n(i) := \inf_{\gamma \in \Gamma} \mathbb{E}^{\gamma} \left\{ \sum_{k=0}^{n-1} \beta^k c(x_k, u_k) \,\middle|\, x_0 = i \right\}$.

- Everything we said so far holds even if $\mathcal{X}$ (state space) is countable & $\mathcal{U}$ is compact.

Next Main result:

We say a Markov policy is stationary/time-invariant if

$$\underbrace{\{\underline{g}_0, \underline{g}_1, \ldots\}}_{\text{Infinite sequence of functions}} \equiv \{\underline{g}, \underline{g}, \underline{g}, \ldots\}$$

Result: Stationary policy is optimal for Infinite horizon discounted.

Algorithms:

Value Iteration

Good: • Guaranteed convergence

• Rate-of-convergence is geometric

Bad(?): Only asymptotic convergence

(Simply implement the contractive iterates)

e.g. choose arbitrary $W_\infty(i) = \text{rand}(n,1)$

then iterate $W_\infty^{k+1} = TW_\infty^k$
$k = 0, 1, \ldots$

Policy Iteration

Finite step convergence.

<u>Idea</u>: $T(\cdot)$ is not only contractive, but also a monotone (nonlinear) operator

(i.e.) If $z \leq \tilde{z}$,

then $Tz \leq T\tilde{z}$

This suggests we can generate approx. optimal policy δ from a finite set.

Policy Iteration Algorithm:

(Instead of iterating value f's, let us iterate policy)

→ Start with any policy $\{\underline{\gamma_0}, \underline{\gamma_0}, \underline{\gamma_0}, \dots\}$ with γ_0 arbitrary

→ Calculate the cost W_{γ_0}, i.e..

$$W_{\underline{\gamma_0}} = \left(I - \beta P_{\gamma_0}\right)^{-1} c_{\underline{\gamma_0}}$$

→ Is $W_{\underline{\gamma_0}} = \underbrace{T W_{\gamma_0}}$?

$$:= \inf_{u \in U} \left[c(i,u) + \beta \sum_{j=1}^{n} p_{ij}(u) W_{\gamma_0}(j) \right]$$

If yes, then stop.

If not, let

$$\gamma_1(i) = \underset{u(\cdot)}{\operatorname{argmin}} \text{ (above RHS)}$$

Again calculate

$$W_{\gamma_1} = (I - \beta P_{\gamma_1})^{-1} c_{\gamma_1} \quad \ldots \text{ continue.}$$

Another alternative : (Linear Programming solution)

$$\max \sum_{i=1}^{n} z(i)$$

subject to

$$z(i) \le c(i,u) + \sum_{j=1}^{n} \beta P_{ij}(u) z(j) \quad \forall i, \ \forall u$$

decision variable

# of constraints = (# of states) × (# of controls)

The LP constraint $\Longleftrightarrow$

$$z \leq Tz \leq T^2 z \leq T^3 z \cdots \rightarrow W_\infty.$$

(i.e.) Sol$^{\underline{n}}$ of LP is W_∞.

So far, assumed that $\underline{x}$ is available for feedback, i.e., (completely observed) Markov decision process (MDP).

More realistic case, $\underline{x}$ is NOT directly observable.

What if we can only observe $y \neq x$.

POMDP
(Partially Observed Markov Decision Process)

System: $p_{ij}(u)$, $i \in \mathcal{X}$, $u \in \mathcal{U}$

Noisy observation of state:

Y = set of observations

$$p(\underline{y} \mid x) = \mathbb{P}\big(y(t) = y \mid x(t) = x\big)$$

History dependent policies:

$$u(t) = \gamma_t\big(y(0), y(1), \ldots, y(t), u(0), u(1), \ldots u(t-1)\big)$$

$$\underline{\gamma} = (\underline{\gamma_0}, \gamma_1, \ldots, \gamma_{T-1}) \leftarrow \text{policy}$$

The main result (in this Markov chain setting):
we will show that we can solve this POMDP as a 2-step procedure by separating <u>estimation & control</u>.

(whenever these 2 problems can be decoupled, we say "Separation Principle" holds).

<u>State estimation for a Markov Chain:</u>

p_{ij} : Markov chain (no control)

$p(y|x)$: observation probabilities.

<u>Question:</u> What is $x(t)$?

We want
$\mathbb{P}(x(t)=i \mid y(0), y(1), \ldots, y(t))$ for each $i=1,\ldots,n$

Define $y^t := \underbrace{(y(0), \dots, y(t))}_{\text{History upto time } t}$

Denote:

$$p_{t|t}(i \mid y^t) := \mathbb{P}(x(t) = i \mid y^t)$$

Define vector:

$$p_{t|t}(y^t) = [\, p_{t|t}(1|y^t),\ p_{t|t}(2|y^t), \dots, p_{t|t}(N|y^t) \,]$$

= conditional probability distribution of $x(t)$ given y^t

We will show : (Recursive sol^{n})

$$p_{t|t} \xrightarrow[\text{or prediction}]{\text{Time update}} p_{t+1|t} \xrightarrow{\text{Measurement update}} p_{t+1|t+1}$$