Lecture #16

Stochastic (Discrete time) OCP:

$$\underset{\gamma(\cdot)\in\Gamma}{\text{minimize}} \; \mathbb{E}\left[c_T(\underline{x}(T)) + \sum_{k=0}^{T-1} c_k(\underline{x}_k, \underline{u}_k) \right]$$

$$= (\underline{x}_0, \underline{u}_0, \underline{x}_1, \underline{u}_1, \ldots, \underline{x}_{k-1}, \underline{u}_{k-1}, \underline{x}_k)$$

s.t.

(Discrete-time noisy case, spl. case of MDP)

$$\underline{x}_{k+1} = \underline{f}_k(\underline{x}_k, \underline{u}_k, \underline{w}_k), \qquad \underline{x}_k \in \mathcal{X}, \; \underline{u}_k \in \mathcal{U}$$

Controlled Markov chain

(or, the following MDP):

$$\underline{x}(k) \sim \text{Markov}([P_{ij}]), \quad i,j = 1,\ldots,m \quad \Big\} \; P(\underline{u}) := [P_{ij}(\underline{u})]$$

$$\Updownarrow$$

$$\mathbb{P}(\underline{x}(k+1) = s_j \mid \underline{x}(k) = s_i) = p_{ij} \geq 0.$$

(or, deterministic dynamics)

$$\underline{x}_{k+1} = f_k(\underline{x}_k, \underline{u}_k) \qquad \underline{w}_k \equiv 0 \;\; \forall\, k = 0, \ldots, T-1$$

Given the OCP problem in any of the forms given in the prev. page

Define:

$$V_k(\underline{x}) := \inf_{(\underline{\gamma}_k, \underline{\gamma}_{k+1}, \dots, \underline{\gamma}_{T-1})} \mathbb{E}\left[\left\{c_T(\underline{x}(T)) + \sum_{s=k}^{T-1} c_s(\underline{x}_s, \underline{u}_s)\right\} \Big| \underline{x}_k = \underline{x}\right]$$

We call the above f^n as "value f^n"

Interpretation: "Cost-to-go" from state $\underline{x}$ @ time k

(i.e.) Optimal value of your action that results from applying optimal policy from $\underline{x}_k = \underline{x}$.

2 Main results in DP:

$\boxed{\text{Result \#1}}$ Optimal policy γ^* is a deterministic Markov policy, among the class of all history-dependent randomized policies "Γ".

$\boxed{\text{Result \#2}}$ We derive a recursion on $V_k(\underline{x})$ [called DP eq^n]

DP equations : (derived in the Chapter I emailed)

For Discrete-time noisy case :

$$V_T(\underline{x}) = c_T(\underline{x})$$

$$V_k(\underline{x}) = \inf_{u(\cdot)\in\mathcal{U}} \left\{ c_k(\underline{x},\underline{u}) + \mathbb{E}_{\underline{w}_k}\left[V_{k+1}(f_k(\underline{x},\underline{u},\underline{w}_k)) \right] \right.$$

If c_k also depends on w_k, then we need to put $\mathbb{E}[\cdot]$ outside the curly braces in RHS

$k = T-1, T-2, \ldots, 0$

Markov case :

$$V_T(\underline{x}) = c_T(\underline{x})$$

$$V_k(\underline{x}) = \inf_{u(\cdot)\in\mathcal{U}} \left\{ c_k(\underline{x},\underline{u}) + \sum_{j\in\mathcal{X}} p_{x,j}(u)\, V_{k+1}(j,\underline{x}) \right\},$$

$k = T-1, T-2, \ldots, 0$

where $[p_{ij}(u)]$ is the controlled Markov chain.

Deterministic case:

$\underline{w}_k \sim$ sample path dependency vanishes

so the $E_{w_k}[\cdot]$ can be dropped.

$$V_T(\underline{x}) = C_T(\underline{x})$$

$\therefore$ DP eq$^{\underline{n}}$ becomes:

$$V_k(\underline{x}) = \inf_{u \in \mathcal{U}} \left\{ C_k(\underline{x}, \underline{u}) + V_{k+1}\left(\underline{f_k(\underline{x}, \underline{u})}\right)\right\}$$

We call $V_k(\underline{x})$ as "value function"

Example (Deterministic DP)

An investor receives annual salary/income $\$ x_k$ in the year k. He consumes $\$ u_k$, and adds the rest $(x_k - u_k)$ to his capital, $0 \le u_k \le x_k$. The capital is invested @ interest rate $\theta \times 100\%$.

$\therefore$ Dynamics of income :

(i.e.) his income in the $(k+1)^{th}$ yr :

$$x_{k+1} = f_k(x_k, u_k) = x_k + \theta(x_k - u_k)$$

Objective : to maximize total consumption over lifetime T years.

(i.e.) deterministic OCP :

$$\text{maximize} \sum_{k=0}^{T-1} u_k$$

$$u_k = \underline{\gamma}(\underbrace{x_0, u_0, x_1, u_1, \ldots, x_{k-1}, u_{k-1}, x_k}_{\text{History}})$$

History dependent policies.

Question: What is the optimal consumption policy?

Apply DP : $\left.\begin{aligned} c_k(x, u) &= u_k \\ c_T(x) &= 0. \end{aligned}\right\}$ Time invariant OCP $\Updownarrow$ Time homogeneous DP

Define time-to-go : (countdown clock)

$n := T - k$, i.e., $n = 0, 1, \ldots, T$

Here, (x, u) are generic values for (x_k, u_k)

$\therefore$ DP eqn in forward (countdown) time:

$$W_n(x) = \sup_{u(\cdot)\in\mathcal{U}} \Big\{ \overset{u}{\cancel{c_n(x,u)}} + W_{n-1}\Big(\underbrace{x+\theta(x-u)}_{\text{from the RHS of dynamics}}\Big) \Big\}$$

$\boxed{W(\cdot) \text{ is the count-down value fn.}}$

s.t. $W_0(\underline{x}) = 0$ (nothing more to consume), $n = 1, 2, \dots, T$

How to solve: (Let's do first few iterations of DP eqn)

$$W_1(x) = \sup_{0\le u\le x} \Big\{ u + \overset{0}{\cancel{W_0\big(x + \theta(x-u)\big)}} \Big\}$$

$$= \sup_{0\le u\le x} \{u + 0\} = x$$

$$W_2(x) = \sup_{0 \le u \le x} \left\{ u + W_1\big(x + \theta(x-u)\big) \right\}$$

$$= \sup_{0 \le u \le x} \left\{ \underbrace{u + x + \theta(x-u)}_{\text{linear in } u} \right\}$$

($\Rightarrow$ maximum is achieved either @ $u = 0$, or @ $u = x$) This is like Bang-Bang

$$= \max\{(1+\theta)x,\ 2x\}$$

$$= \Big(\underbrace{\max\{1+\theta,\ 2\}}_{\rho_2 \text{ (say)}}\Big)\, x = \rho_2\, x$$

This motivates the guess:

$$W_{n-1}(x) = \rho_{n-1}\, x$$ for some (to-be-determined) constant ρ_{n-1}

Verification of guess:

$$W_n(x) = \max_{0 \le u \le x} \left\{ u + \rho_{n-1}\left(x + \theta(x-u)\right) \right\}$$

$$= \underbrace{\left(\max\left\{ (1+\theta)\rho_{n-1},\ 1+\rho_{n-1} \right\} \right)}_{\rho_n} x$$

$$= \rho_n \quad x$$

$\therefore$ Guess is correct (structurally) & $W_n(x) = \rho_n x$

where ρ_n itself solves a scalar recursion:

$$\rho_n = \rho_{n-1} + \max\{\theta \rho_{n-1},\ 1\}$$

This gives: $\rho_n = \begin{cases} n, & \text{if } n \le n^* \\ (1+\theta)^{n-n^*} n^* & \text{if } n \ge n^* \end{cases}$

where n^* is the least integer

s.t. $(1+\theta)\, n^* \geq 1 + n^*$

$$\Updownarrow$$

$$n^* \geq 1/\theta$$

$$\Updownarrow$$

$$n^* = \lceil 1/\theta \rceil .$$

$\therefore$ Optimal policy is to invest entire income in yrs.

$0, 1, 2, \ldots, T-n^*-1$ (to build up enough capital)

& then consume whole of the income in yrs.

$T-n^*, T-n^*+1, \ldots, T-1$.

$\blacksquare$

Example of Stochastic DP: (Inventory Control)

Suppose you are the store manager for Walmart

Dynamics of stock:

$$x_{k+1} = x_k + u_k - w_k$$

$k = 0, 1, 2, \ldots, T-1$

x_k = Stock available @ the beginning of the k^{th} period

u_k = Stock ordered (and immediately delivered) at the beginning of the k^{th} period

w_k = Demand during k^{th} period with known probability distribution

Assume that demands $w_0, w_1, \ldots, w_{T-1}$ are indep. r.v.s

Assume $u_k \geq 0$

Negative x (stock) means backlogged demand

cost has 2 components:

(i) $r(x_k)$: penalty for either positive stock (i.e., holding cost for excess inventory) OR negative stock (i.e., shortage cost for unfulfilled demand)

(ii) $c u_k$: purchasing cost (here, c is the cost for per unit ordered)

Want to solve the OCP:

$$\min_{\gamma_k \in \Gamma} \mathbb{E}\left\{ r(x_T) + \sum_{k=0}^{T-1} \left(r(x_k) + c u_k \right) \right\},$$

where $r(.)$ is a piecewise linear cost

piecewise linear cost: $r(x_k) := p\max(0, -x_k) + h\max(0, x_k)$

with slope h when stock > 0

" " p " " < 0.

Assume $p > c$, otherwise $u_k^* = 0 \ \forall\ k = 0, 1, \ldots, T-1$

Recall: OCP is equivalent to

$$\min_{\gamma \in \Gamma} \mathbb{E}\left[\sum_{k=0}^{T-1} \{ r(x_k + u_k - w_k) + c u_k \} \right]$$

Since $\mathbb{E}[r(x_0)]$ cannot be influenced so, we can subtract that.

$\therefore$ DP eqn:

$$V_T(\underline{x}) = 0,$$

$$V_k(\underline{x}) = \inf_{u_k \geq 0} \mathbb{E}_{w_k} \left\{ \underbrace{r(x+u-w_k) + cu}_{\boxed{c_k(x, u, w_k)}} + V_{k+1}\Big(\underbrace{x + u - w_k}_{\boxed{f_k(x, u, w_k)}} \Big) \right\}$$

(assuming w_k as iid)

$$= \inf_{u \geq 0} \mathbb{E}_w \left\{ r(x+u-w) + cu + V_{k+1}(x + u - w) \right\}$$

where $k = T-1, T-2, \ldots, 0$

Can show by induction that $V_k(\underline{x})$ is a non-neg. convex fn. that $\to \infty$ as $|x_k| \to \pm\infty$

Write $\mathbb{E}_w[r(x+u-w)]$ as $H(x+u)$, where $H(y) := \mathbb{E}[r(y-w)]$

Then the DP recursion in the prev. page can be re-written as:

$$V_T(x) = 0$$

$$V_k(x) = \inf_{u \geq 0} \left\{ cu + H(x+u) + \mathbb{E}\left[V_{k+1}(x+u-w)\right] \right\}$$

From the defn, $H(y) := \mathbb{E}[r(y-w)]$, notice that since $r(\cdot)$ is a convex fn, so is $H(\cdot)$. Furthermore, slope of $H(\cdot)$ approaches "h" as $x \to \infty$

and " "$-p$" as $x \to -\infty$

Let $y := x+u$. Then the constraint in "inf" $u \geq 0 \Leftrightarrow y \geq x$, and $cu \equiv cy - cx$, Also, $G(y) := cy + H(y) + \mathbb{E}[V_{k+1}(y-w)]$

Then, $V_k(x) = \inf_{y \geq x} \left(G(y) - cx \right) = \left(\inf_{y \geq x} G(y) \right) - cx$.

Suppose, we can show that "G" is convx and $\lim_{|x| \to \infty} G(x) \to \infty$,

and let $S_k := \arg\min_{y_k} \left(G_k(x_k) \right)$

Then, $V_k(x_k) = G_k(y_k^*) - c x_k$, where

$$y_k^* = \begin{cases} s_k & \text{if } s_k \geq x_k, \\ x_k & \text{otherwise.} \end{cases}$$

(i.e.) Optimal policy is a time-varying threshold policy (i.e., restock to s_k iff $x_k \leq s_k$).

To finish this example, we now prove that $V_k(x_k)$ is cnvx with slope > 0 as $x_k \to \infty$, and with slope < 0 as $x_k \to -\infty$ for $k = 0, 1, \ldots, T-1$. Backward induction proof

Notice that $G_{T-1}(y) = cy + H_{T-1}(y)$ is cnvx since $H_{T-1}(y)$ is cnvx, and its slope approaches $(h+c) > 0$ as $y \to \infty$, and approaches $-p+c < 0$ as $y \to -\infty$

So, $V_{T-1}(x_{T-1}) = c(s_{T-1} - x_{T-1})^+ + H_{T-1}(\max\{s_{T-1}, x_{T-1}\})$

where $z^+ := \max(z, 0)\ \forall z \in \mathbb{R}$.

This f^n. is cnvx with slope > 0 as $x_{T-1} \to \infty$ and slope approaching $-c < 0$ as $x_{T-1} \to -\infty$. Now show this recursively backward.

General result: Back to DP Theory

$\underline{\gamma}^* \in \Gamma$, (optimal policy) is a non-randomized (deterministic) Markovian policy

(i.e) $\gamma^* \in \Gamma_M \subset \Gamma$ ~ class of all history dependent randomized policies.

Infinite horizon Problems:

$$\min_{\underline{\gamma} \in \Gamma} \mathbb{E}\left[\sum_{k=0}^{+\infty} c_k(\underline{x}_k, \underline{u}_k, \underline{w}_k) \,\middle|\, \underline{x}(0) = \underline{x} \right]$$

called Total cost

→ may NOT be well defined (series sum may NOT converge or may diverge even when "c" is bdd.)

NOTE :

$$\sum_{k=0}^{+\infty} a_k \text{ converges} \implies \lim_{k\to\infty} a_k = 0$$

(contrapositive)

$$\lim_{k\to\infty} a_k \neq 0 \implies \sum_{k=0}^{+\infty} a_k \text{ diverges.}$$

This motivates introducing discount factor

$$0 < \beta < 1$$

a number

$\boxed{\text{Total discounted cost}}$

$$\min_{\underline{\gamma} \in \Gamma} \mathbb{E}\left[\sum_{k=0}^{+\infty} \beta^k c_k(\underline{x}_k, \underline{u}_k) \,\middle|\, \underline{x}(0) = \underline{x} \right] \quad \text{---} (*)$$

$\beta \approx 0$: Tomorrow/ future is NOT imp.

If $\beta \approx 1$: All days are important (approx.)

$\boxed{\text{Total discounted cost is ALWAYS well defined if } f^{u}_{\pi},\ c_k(\cdot) \text{ is bdd.}}$

"MYOPIA"

Today $1 = $1

Tomorrow $1 = 90¢ today.

(geometric fall of USD)

Day after tomorrow : 81¢

$\beta = 0.9$

Finite horizon Discounted cost :

Let $W(x, n) = \min\limits_{\gamma \in \Gamma} \mathbb{E}\Big[\sum\limits_{k=0}^{T-1} \beta^k c_k(x_k, u_k, w_k) \Big| x(0) = x \Big]$

DP eq$^{\underline{n}}$: (when x is Markov (P))

$$W(x,n) = \min_{u \in \mathcal{U}} \left\{ c(x,u) + \sum_{j \in \mathcal{X}} p_{xj}(u) \cdot \beta W(j, n-1) \right\}$$

$$W(x,0) = 0.$$

Infinite horizon Discounted Cost:

$$W(x,\infty) = \min_{u \in \mathcal{U}} \left\{ c(x,u) + \beta \sum_{j \in \mathcal{X}} p_{xj}(u) W(j,\infty) \right\} \quad \text{---} (**)$$

same function.

Algebraic nonlinear eq$^{\underline{n}}$

nonlinear since $c(x,u) \neq 0$.

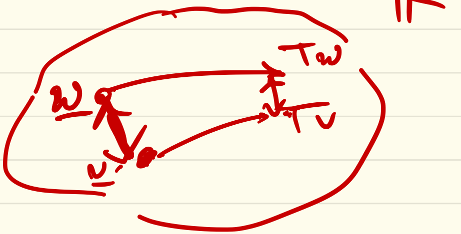
One eq$^{\underline{n}}$ for each x (when x is enumerable)

$\therefore$ x eq$^{\underline{n}}$s in x unknowns.

Question :

Solution → Exists ?
→ unique ?
→ Algorithm to solve it?

Contraction Map.



$F \subseteq \mathbb{R}^N$ closed subset

$0 < \beta < 1$

$$d(Tw, Tv) \le \beta\, d(w, v)$$

The map T is called contractive.

$T : F \mapsto F$

$w, v \qquad Tw, Tv$

Contraction Mapping Thm.

$F \subseteq \mathbb{R}^N$
closed set.

Suppose $\exists$ a scalar β where $0 < \beta < 1$, s.t.

$$\| T\underline{w} - T\underline{v} \| \leq \beta \, \| \underline{w} - \underline{v} \|$$

$$\forall \; \underline{v}, \underline{w} \in F$$

Then

① $\exists \; \underline{z} \in F$ s.t. $T\underline{z} = \underline{z}$, called fixed pt. of map. $T(\cdot)$

② The fixed pt. $\underline{z}$ is unique

③ start with any $\underline{w} \in F$, then

$$\lim_{n \to \infty} T^n \underline{w} = \underline{z} \quad (\text{Fixed pt. iteration})$$