Mathematical Foundations of Reinforcement Learning

Lecture II: Dynamic Programming Algorithms

* Dynamic Programming
* Finite Horizon control
* Infinite Horizon control
  * Stationary Policies
  * Bellman operator
  * Value Iteration
  * Policy Iteration

Dynamic Programming

Ingredients:

a) A discrete-time evolution,
$x_1, \ldots, x_m$ states.

b) x_0, a termination state.

c) A control set, node $U(i)$

d) Transition probabilities $P_{ij}(u)$

$$x_i \xrightarrow{u} x_j$$

e) A cost that accumulates over time:

$$\alpha^k \, g(i, u, j)$$

g: running cost; k = iteration

$0 < \alpha \leq 1$: discount factor (depreciation)

THE GOAL: TO FIND AN OPTIMAL POLICY MAP

$$\Pi := \{\mu_0, \mu_1, \dots\}$$

$\mu_k(i)$:= OPTIMAL FEEDBACK AT TIME k AT STATE i

$\mu_k(i) \in U(i)$

FOR A FIXED POLICY, THE SEQUENCE OF STATES BECOMES A MARKOV CHAIN WITH TRANSITION PROBABILITIES

$$P(i_{k+1} = j \mid i_k = i) = P_{ij}(\mu_k(i))$$

↳ Obs: EQUIVALENT TO $\frac{dx}{dt} = f(x(t), u(t))$ WITH $u(t)$ A GIVEN SIGNAL

DIFFERENT DP EQUATIONS FOR DIFFERENT HORIZONS K AND COST g

Finite Horizon Control

Fix a control horizon: $0, \dots, N$

Given a policy $\pi = \{\mu_0, \dots, \mu_{N-1}\}$ and an initial state x_i,

The expected total cost is given by

$$J_N^{\pi}(i) := \mathbb{E}\left[\alpha^N G(i_N) + \sum_{k=0}^{N-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \,\middle|\, i_0 \, i\right]$$

$\alpha^N G(i_N)$: terminal cost

→ w/ respect to the prob. of the MC $\{x_0, x_1, \dots\}$

The optimal cost-to-go (value function) is given by

$$J_N^*(i) = \min_{\pi} J_N^{\pi}(i)$$

INFINITE HORIZON CONTROL

$$J^{\pi}(i) = \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \mid i_0 = i\right]$$

OPTIMAL COST-TO-GO: $J^*(i) = \min_{\pi} J^{\pi}(i)$

FOR ∞-HORIZON PROBLEMS, WE LOOK FOR A STATIONARY POLICY,

$$\pi = \{\mu, \mu, \mu, \ldots, \mu\} \rightarrow J^{\mu}(i)$$

DYNAMIC PROGRAMMING EQUATIONS

FOR FINITE HORIZON CONTROL:

I. THE CASE WITH $N=1$ (INSTANTANEOUS CONTROL)

$$J_1^*(i) = \min_{\mu_0} \sum_{j=1}^{n} p_{ij}(\mu_0(i)) \left(g(i, \mu_0(i), j) + \alpha G(j) \right)$$

$$= \min_{u \in U(i)} \sum_{j=1}^{n} \underbrace{p_{ij}(u)}_{\mathbb{E}[\dots]} \left(\underbrace{g(i, u, j)}_{\text{EXPECTED COST}} + \alpha \underbrace{G(j)}_{\text{FUTURE COST}} \right)$$

PRESENT

$i \xrightarrow[\min!]{u} j$

→ SOLVING 1 OPTIMIZATION FOR EACH i

II. DP FOR AN ARBITRARY HORIZON N

BELLMAN $\Rightarrow$ $$J_N^*(i) = \min_{u \in U(i)} \sum_j p_{ij}\left(g(i,u,j) + \alpha J_{N-1}^*(j)\right) \quad \forall i$$

VERIFICATION: $J_0^*(i) = G(i)$; $J_1^*(i) = \min_{u \in U(i)} \sum_j p_{ij}(u)\left(g(i,u,j) + \alpha G(j)\right)$

FOR AN N-STEP POLICY π_N WITH INITIAL STATE $i_0 = i$, WE SPLIT

$$\pi_N := \{\pi_{N-1}, u\}$$

$$J_N^*(i) \equiv \min_{\pi} J_N^{\pi}(i) = \min_{\{\pi_{N-1}, u\}} \left\{ \sum_j p_{ij}(u) g(i,u,j) + \alpha J_{N-1}^{\pi_{N-1}}(j) \right\}$$

$$= \min_u \sum_j p_{ij}(u) g(i,u,j) + \alpha \min_{\pi_{N-1}} J_{N-1}^{\pi_{N-1}}(j)$$

$$= \min_u \sum_j p_{ij}(u) g(i,u,j) + \alpha J_{N-1}^*(j)$$

II. DP FOR ∞-HORIZON PROBLEMS

$$J^{\pi}(i) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \,\middle|\, i_0 = i\right]$$

TAKING THE LIMIT TO THE FINITE HORIZON $N \longrightarrow \infty$

$$(*) \quad J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \alpha J^*(j)\right)$$

↳ THE MINIMIZER $u^* \to \mu(i)$ OPTIMAL FEEDBACK

NOTATION: $T(J)(i) \equiv (TJ)(i) := \min_u \sum_j p_{ij}(u)\left[g(i,u,j) + \alpha J(j)\right]$

↓

BELLMAN OPERATOR $\overset{(*)}{\Longrightarrow} J^* = TJ^*$

FIXED POINT!

MORE NOTATION: $T_\mu J(i) := \sum_j P_{ij}(\mu(i)) \cdots$

REPLACING $\min_{u \in U(i)}$ BY $\mu(i)$

OPTIMALITY: $J^* = TJ^* \leq T_\mu J^*$, FOR ANY POLICY μ

MORE NOTATION II: $P_\mu \in \mathbb{R}^{n \times n}$, $P_\mu^{ij} = P_{ij}(\mu(i))$

$$\Rightarrow \boxed{T_\mu J} = \underset{\in \mathbb{R}^n}{\bar{g}_\mu} + \underset{\in \mathbb{R}^n}{P_\mu J}$$

WHERE $\bar{g}_\mu(i) = \sum_{j=0}^{n} P_{ij}(\mu(i))\, g(i, \mu(i), j)$

MORE NOTATION III: $(T^k J)(i) = T(T^{k-1} J)(i) \cdots, T^0 J(i) = J(i)$

(SAME FOR $(T_{\mu m} T_{\mu m} J)(i)$)

PROPERTIES OF T AND T_μ

MONOTONICITY: J_F $J(i) \leq \bar{J}(i)$, $i=1,\ldots,n$

FOR ANY n-DIMENSIONAL VECTORS J AND $\bar{J}$, FOR ANY STATIONARY POLICY μ WE HAVE

$$(T^k J)(i) \leq (T^k \bar{J})(i)$$

$$(T_\mu^k J)(i) \leq (T_\mu^k \bar{J})(i)$$

PROOF: USE THE DEFINITION OF $T^k J$ AND INDUCTION ....

$$T^0 J = J \leq \bar{J} = T^0 \bar{J}$$

$$T^k J = T(T^{k-1} J) = \min_{u \in U} \sum_j p_{ij}(u)\{g(i,u,j) + \alpha T^{k-1} J\} \leq \min_u \ldots + \alpha T^{k-1} \bar{J}$$

$$\| \quad T(T^{k-1}\bar{J}) = T^k \bar{J}$$

FIXED POINT PROPERTIES

ASSUME THERE EXISTS A TERMINATION STATE x_0 SUCH THAT $p_{00}(u) = 1$, $g(0,u,0)=0$ $\forall u \in U(0)$ AND AT LEAST ONE POLICY μ SUCH THAT THE PROBABILITY OF REACHING x_0 AFTER n STAGES IS >0.

THEN: a) THE OPTIMAL COST-TO-GO $J^* \in R^M$ IS THE UNIQUE SOLUTION TO $J^* = T J^*$

b) $\lim_{K\to\infty} T^{(K)} J = J^* \quad \forall\, J \in R^M$

c) A STATIONARY POLICY IS OPTIMAL IIF

$$T_\mu J^* = T J^*$$

Obs: T IS A CONTRACTION MAPPING

$$\|TJ - T\bar{J}\|_\varepsilon \leq \beta \|J - \bar{J}\|_\varepsilon$$

$$0 \leq \beta < 1, \quad \forall J, \bar{J}.$$

WHERE $\|J\|_\varepsilon = \max_{i=1,\dots,n} \frac{|J(i)|}{\xi(i)}$

BANACH FIXED POINT SUGGESTS A SOLUTION METHOD:

VALUE ITERATION

GIVEN ANY $J^0(i), \forall i \Leftrightarrow J^0 \in \mathbb{R}^n$, ITERATE

$$J^{k+1} = TJ^k, \text{ UNTIL CONVERGENCE}$$

+: Simple, Robust

−: Slow depending on β if $\beta \to 1$

ALTERNATIVE FOR SOLVING $J^* = (TJ^*)$:

- NONLINEAR SYSTEM OF EQNS. (VI $\Leftrightarrow$ FPI)
  → USE NEWTON'S METHOD "$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$"

POLICY ITERATION:

GIVEN μ^0: I) POLICY EVALUATION STEP:

SOLVE FOR J^{k+1} $\quad J^{k+1} = T_{\mu^k} J^{k+1}$

II) POLICY UPDATE STEP:

$$\mu^{k+1}(i) = \underset{u}{\operatorname{argmin}} \sum_j p_{ij}(u)\left[g(i,u,j) + \alpha J^{k+1}(j)\right]$$

Eventually $J^k \longrightarrow J^*$, $\mu^k \rightarrow \mu^*$

+ : NEWTON METHOD , FASTER CONVERGENCE

− : NEWTON METHOD , NEED GOOD INITIAL GUESS μ_0

SUMMARY:

* DYNAMIC PROGRAMMING
* FINITE HORIZON CONTROL
* INFINITE HORIZON CONTROL
* STATIONARY POLICIES
* BELLMAN OPERATOR
* VALUE ITERATION
* POLICY ITERATION