MATHEMATICAL FOUNDATIONS OF REINFORCEMENT LEARNING

LECTURE IV: APPROXIMATE DYNAMIC PROGRAMMING

SUMMARY: - COMPUTING OPTIMAL FEEBACKS ONLINE
- APPROXIMATE VALUE ITERATION
- APPROXIMATE POLICY ITERATION
- BELLMAN EQN. METHODS.

REFS: - CHAPTER 5, BERTSEKAS BOOK ON RL.
- A. ALLA, M. FALCONE, D.K. "AN EFFICIENT POLICY ITERATION ALGORITHM FOR DYNAMIC PROGRAMMING EQUATIONS", SISC, 2015.

PREVIOUS LECTURES:

INFINITE HORIZON COST:

discount

running cost

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

VALUE FUNCTION: $J^{*}(i) = \min_{\mu} J^{\mu}(i)$

$$i \xrightarrow{u} j$$

V.F. SATISFIES BELLMAN EQUATION:

+1

0

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

VALID $\forall i \in X$

WITH BELLMAN OPERATOR

$$TJ := \min_{u \in U(i)} \sum_{j \in X} \bar{p}_{ij}(u) \left(\bar{g}(i, \mu, j) + \alpha J(j) \right) \Rightarrow J^{*} = TJ^{*}$$

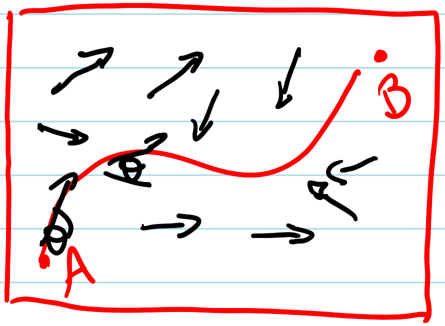
T is a contraction map $\Rightarrow$ fixed point iteration

$$J^*(i) = \lim_{k\to\infty} J^k(i) \quad \text{where } J^{k+1} = T J^k$$

Once $\underline{\underline{J^*}}$ has been computed, μ^* is recovered as

$$\mu^*(i) \in \arg\min_{u \in U(i)} \sum_j p_{ij}(u)\left(\underline{g(i,u,j)} + \alpha \underline{J^*(j)}\right)$$

Example: Zermelo navigation problem



Go from A to B in min time subject to wind

Discrete eqns of motion

$$X_{k+1} = X_k + \Delta t\left(\text{Wind}_x + \bar{V}\cos(\theta_k)\right)$$

$$Y_{k+1} = Y_k + \Delta t\left(\text{Wind}_y + \bar{V}\sin(\theta_k)\right)$$

control

State (X_k, Y_k)

SOME IDEAS FOR ANN'S IN THE CONTROL FIELD

$$\mu^*(i) \in \arg\min \sum_j p_{ij}(u) \left(g(i,u,j) + \alpha \underbrace{J^*(j)}_{\substack{\text{REPLACE BY A LOW} \\ \text{COMPLEXITY ANN} \\ \tilde{J}(j,r)}}\right)$$

EXAMPLE: LQ CONTROL

$x_{k+1} = a x_k + b u_k$, $x_k, u_k \in \mathbb{R}$

$g(x,u,y) = K u^2 + x^2$, K constant

ASSUME A LINEAR ARCHITECTURE $:= \tilde{J}(i,r) = r_0 + r_1 x + r_2 x^2$

$$\mu^*(\underline{x_i}) \in \arg\min_{u} \underbrace{K u^2 + x_i^2 + \bar{r}_0 + \bar{r}_1 (a x_i + b u) + \bar{r}_2 (a x_i + b u)^2}_{g(u) \Rightarrow g'(u) = 0}$$

HOW DO WE INCORPORATE THE APPROX. OF $J^*(i)$ by $\tilde{J}(i,r)$

↳ APPROXIMATE VALUE ITERATION:

$$\hat{J}_{k+1}(i) = \min_{u} \sum_{j} p_{ij}(u)\left(g(i,u,j) + \alpha \tilde{J}(j, r_k)\right)$$

$\forall i \in S_k$, THEN r_{k+1} IS UPDATED ACCORDING TO

$$r_{k+1} \in \arg\min_{r} \sum_{i \in S_k} \| \hat{J}_{k+1}(i) - \tilde{J}(i,r) \|^2$$

↳ DEFINE ARCHITECTURE $\tilde{J}(i,r)$ ✓

↳ INITIAL PARAMETRIC GUESS r_0 ✓

↳ UPDATE $\hat{J}_{k+1}$ USING $\tilde{J}(i,r)$ OVER A SUBSET S_k

↳ USE $\hat{J}_{k+1}(i)$ TO TRAIN r_{k+1}

OBS: - THE ARCHITECTURE $\tilde{J}(i,r)$ REMAINS, ONLY r CHANGES AFTER EVERY ITERATION.

- THE EVALUATION OF p_{ij}'s CAN BE COSTLY UNLESS $\# p_{ij} \neq 0$ IS SMALL.

OBS: AVI IS VERY SIMILAR VI, BUT IT IS CRITICAL TO ENSURE THAT

$$\boxed{\left\| \tilde{J}(i, r_{k+1}) - \hat{J}_{k+1}(i) \right\| \leq \delta}$$

WITH δ SUFFICIENTLY SMALL, OTHERWISE THE CONTRACTIVITY OF THE BELLMAN OPERATOR IS LOST

PATHOLOGICAL CASE:

$$\begin{tikzcd} 1 \arrow[r, "\text{Cost}=0"'] & 2 \arrow[loop right, "\text{Cost}=0"] \end{tikzcd}$$

$\Rightarrow$ $p_{11} = 0$, $p_{12} = 1$, $g \equiv 0$, $\alpha > 0$ discount factor

$p_{22} = 1$, $p_{21} = 0$

It's clear that $J^*(1) = J^*(2) = 0$ (because $g \equiv 0$)

USE BELLMAN EQN. $J^*(i) = \sum_j p_{ij} (0 + \alpha J^*(j))$

BELLMAN
$$J^*(1) = \alpha J^*(2)$$
$$J^*(2) = \alpha J^*(2)$$

VALUE ITERATION
$$J_{k+1}(1) = \alpha J_k(2)$$
$$J_{k+1}(2) = \alpha J_k(2)$$

ASSUME A LINEAR APPROXIMATION $\tilde{J}(i, r) = \underline{i \cdot r}$

$$\Rightarrow J_k(1) = r_k, \quad J_k(2) = 2 \cdot r_k.$$

$$\Rightarrow T\tilde{J}_k = (2\alpha r_k, 2\alpha r_k) \Leftrightarrow \hat{J}_{k+1}$$

LINEAR LEAST SQUARES:

$$r_{k+1} \in \underset{r}{\arg\min} \left(\underbrace{r}_{\tilde{J}(1,r)} - 2\alpha r_k \right)^2 + \left(\underbrace{2r}_{\tilde{J}(2,r)} - 2\alpha r_k \right)^2$$

$$\Rightarrow 2(r - 2\alpha r_k) + 2 \cdot (2r - 2\alpha r_k) \cdot 2 = 0$$

$$r_{k+1} = \frac{6}{5} \alpha r_k \quad //$$

$\Rightarrow$ WE NEED $\underline{\alpha} < \frac{5}{6}$ TO CONVERGE $(r^* = 0)$

APPROXIMATE POLICY ITERATION

__PI:__ GIVEN μ_k, SOLVE

$$J^{k+1} = T_{\mu^k} J^{k+1} \quad \text{(POLICY EVAL)}$$

$$\mu^{k+1} = \arg\min_{\mu} \sum p_{ij}(u) \left(g(i,a,j) + J^{k+1}(j) \right)$$

(POLICY UPDATE)

↳ INCORPORATE ANN BY REPLACING J by $\tilde{J}(i, r)$

BELLMAN EQN METHODS

IDEA: SOLVE DIRECTLY BELLMAN EQN BY USING A NN.

$$\min_{r} \sum_{i \in S} \underbrace{\left(\tilde{J}(i,r) - \min_{u} \sum p_{ij}(u) \left(g_{ij}(u) + \tilde{J}(j,r) \right) \right)}_{\text{RESIDUAL OF BELLMAN EQN. } (J^* = TJ^*)}^2$$

RESIDUAL: $D(i,r) \equiv \tilde{J}(i,r) - \min_{u} \sum p_{ij}(u) \left(\underline{\underline{g_{ij}(u)}} + \tilde{J}(j,r) \right)$

$$\Longleftrightarrow \min_{r} \sum_{i \in S} D^2(i,r)$$

APPLYING SGD TO MINIMIZE THE RESIDUAL

i SAMPLE AT K

$$r_{k+1} = r_k - \gamma \underbrace{D(i, r_k) \nabla_r D(i, r_k)}$$

$$= r_k - \gamma \underline{D(i, r_k)} \left(\sum_j p_{ij}(\bar{u}) \nabla_r \tilde{J}(j, r_k) - \nabla_r \tilde{J}(i, r_k) \right)$$

WHERE

$$\boxed{\bar{u} = \arg\min_u \sum_j p_{ij}(u) \left(g(i, u, j) + \tilde{J}(j, r_k) \right)}$$