Dynamic Programming Lecture #8

Outline:

- Controlled Markov chains
- Value iteration proof

Controlled Markov Chains

- State transition probabilities are now a function of control input: $P(u)$
- Example: Machine repair.
  - States $= \begin{cases} 1 & \text{up} \\ 2 & \text{down} \end{cases}$
  - Controls $= \begin{cases} S & \text{stop } \& \text{ repair if needed} \\ C & \text{continue} \end{cases}$
  - Control dependent transitions:

$$P(C) = \begin{pmatrix} Pr(\text{up}|\text{up}) & Pr(\text{down}|\text{up}) \\ Pr(\text{up}|\text{down}) & Pr(\text{down}|\text{down}) \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0 & 1 \end{pmatrix}$$

- For a policy, $u = \mu(x)$,

$$P(\mu) = \begin{pmatrix} \text{—1st row of } P(\mu(1))\text{—} \\ \text{—2nd row of } P(\mu(2))\text{—} \\ \vdots \\ \text{—}k\text{th row of } P(\mu(k))\text{—} \\ \vdots \\ \text{—}n\text{th row of } P(\mu(n))\text{—} \end{pmatrix}$$

Controlled Markov Chains & Random Disturbances

- Original (stage invariant) setup:

$$x_{k+1} = f(x_k, u_k, w_k)$$

- Controlled Markov chains:

$$\mathrm{P}\left[x_{k+1} = j | x_k = i, u_k\right] = p_{ij}(u_k)$$

  What happened to w?

- Connection:

$$x_{k+1} = w_k$$
$$\mathrm{P}\left[w_k = j | x_k = i, u_k\right] = p_{ij}(u_k)$$

- Value iteration:
  - Assume stage cost does not depend on w (no loss of generality)
  - Original setup:

$$J_k(x_k) = \min_{u_k \in \mathcal{U}_k(x_k)} \mathrm{E}\left[g_k(x_k, u_k) + J_{k+1}(f(x_k, u_k, w_k))\right]$$

  - Markov chain setup:

$$J_k(i) = \min_{u_k \in \mathcal{U}_k(i)} g_k(i, u_k) + \sum_{j=1}^{n} J_{k+1}(j) p_{ij}(u_k)$$

- Can also have *stage dependent* probabilities

Cost of Policy

- Policy: $\pi = \{\mu_1, \mu_2, ..., \mu_{N-1}\}$
- Transitions: $P_k(\mu_k)$
- i^{th} coordinate row vector:

$$e_i = \begin{pmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{pmatrix}$$

- Policy dependent cost vector:

$$G_k(\mu) = \begin{pmatrix} g_k(1, \mu(1)) \\ g_k(2, \mu(2)) \\ \vdots \\ g_n(n, \mu(n)) \end{pmatrix} \ \& \ G_N = \begin{pmatrix} g_N(1) \\ g_N(2) \\ \vdots \\ g_N(n) \end{pmatrix}$$

- Cost of trajectory $\{x_0, x_1, ..., x_N\}$ under π:

$$J(x_0, ..., x_N; \pi) = e_{x_0} G_0(\mu_0) + e_{x_1} G_1(\mu_1) + ... + e_{x_N} G_N$$

- *Expected* cost of trajectory given x_0 under π:

$$\sum_{x_0 ... x_N} J(x_0, ..., x_{N-1}; \pi) \mathrm{P}\left[x_0, ..., x_N | x_0, \pi\right]$$

Cost of Policy, cont

- Use $\mathrm{E}\left[\sum_j Z_j\right] = \sum_j \mathrm{E}\left[Z_j\right]$ on

$$J(x_0, ..., x_N; \pi) = e_{x_0} G_0(\mu_0) + e_{x_1} G_1(\mu_1) + ... + e_{x_N} G_N$$

  to rewrite *expected* cost given x_0 under π:

$$\begin{aligned}
\mathrm{E}\left[J(x_0, ..., x_{N-1}; \pi) | x_0\right] &= e_{x_0} G_0(\mu_0) \\
&+ \underbrace{e_{x_0} P_0(\mu_0)}_{q_1 = \mathrm{E}\left[e_{x_1} | x_0, \mu_0\right]} G_1(\mu_1) \\
&+ \underbrace{e_{x_0} P_0(\mu_0) P_1(\mu_1)}_{q_2 = \mathrm{E}\left[e_{x_2} | x_0, \mu_0, \mu_1\right]} G_2(\mu_2) + ... \\
&+ \underbrace{e_{x_0} P_0(\mu_0) P_1(\mu_1) \ldots P_{N-2}(\mu_{N-2})}_{q_{N-1} = \mathrm{E}\left[e_{x_{N-1}} | x_0, \mu_0, \mu_1, ..., \mu_{N-2}\right]} G_{N-1}(\mu_{N-1}) \\
&+ \underbrace{e_{x_0} P_0(\mu_0) P_1(\mu_1) \ldots P_{N-2}(\mu_{N-2}) P_{N-1}(\mu_{N-1})}_{q_N = \mathrm{E}\left[e_{x_N} | x_0, \mu_0, \mu_1, ..., \mu_{N-1}\right]} G_N
\end{aligned}$$

- Notice triangular structure in μ_k dependence!

Value Iteration Revisited

- Rewrite (for some function α and row vector β):

$$\begin{aligned}&\mathrm{E}\left[J(x_0,...,x_{N-1};\pi)|x_0\right] = \\ &\quad \alpha(x_0,\mu_0,...,\mu_{N-2}) + \beta(x_0,\mu_0,...,\mu_{N-2})\left(\,G_{N-1}(\mu_{N-1}) + P_{N-1}(\mu_{N-1})G_N\,\right)\end{aligned}$$

- Accordingly:

$$\begin{aligned}&\min_{\mu_0,\ldots,\mu_{N-1}} \mathrm{E}\left[J(x_0,...,x_{N-1};\pi)|x_0\right] = \\ &\quad \min_{\mu_0,\ldots,\mu_{N-1}} \alpha(x_0,\mu_0,...,\mu_{N-2}) \\ &\quad +\beta(x_0,\mu_0,...,\mu_{N-2})\begin{pmatrix} g_{N-1}(1,\mu_{N-1}(1)) + \sum_{j=1}^n P_{N-1}^{1j}(\mu_{N-1}(1))G_N(j) \\ g_{N-1}(2,\mu_{N-1}(2)) + \sum_{j=1}^n P_{N-1}^{2j}(\mu_{N-1}(2))G_N(j) \\ \vdots \\ g_{N-1}(n,\mu_{N-1}(n)) + \sum_{j=1}^n P_{N-1}^{nj}(\mu_{N-1}(n))G_N(j)\end{pmatrix}\end{aligned}$$

- Minimizing over μ_{N-1} results in term-by-term minimization:

$$J^*_{N-1} = \begin{pmatrix} \min_{u_{N-1}} g_{N-1}(1,u_{N-1}) + \sum_{j=1}^n P_{N-1}^{1j}(u_{N-1})G_N(j) \\ \min_{u_{N-1}} g_{N-1}(2,u_{N-1}) + \sum_{j=1}^n P_{N-1}^{2j}(u_{N-1})G_N(j) \\ \vdots \\ \min_{u_{N-1}} g_{N-1}(n,u_{N-1}) + \sum_{j=1}^n P_{N-1}^{nj}(u_{N-1})G_N(j)\end{pmatrix}$$

Value Iteration Revisited, cont

- μ_{N-1} eliminated (replaced by J^*_{N-1}):

$$\min_{\mu_0,\ldots,\mu_{N-1}} \mathrm{E}\left[J(x_0,...,x_{N-1};\pi)|x_0\right] = \\ \min_{\mu_0,\ldots,\mu_{N-2}} \alpha(x_0,\mu_0,...,\mu_{N-2}) + \beta(x_0,\mu_0,...,\mu_{N-2})J^*_{N-1}$$

- Repeat analysis, but now J^*_{N-1} plays role of G_N:

$$\min_{\mu_0,\ldots,\mu_{N-1}} \mathrm{E}\left[J(x_0,...,x_{N-1};\pi)|x_0\right] = \\ \min_{\mu_0,\ldots,\mu_{N-2}} \alpha(x_0,\mu_0,...,\mu_{N-3}) + \beta(x_0,\mu_0,...,\mu_{N-3})\left(\, G_{N-2}(\mu_{N-2}) + P_{N-2}(\mu_{N-2})J^*_{N-1}\,\right)$$

- Minimization over μ_{N-2} produces J^*_{N-2} as function of J^*_{N-1}
- Repeated application is same as value iteration:

$$\begin{aligned} J_k(x_k) &= \min_{u_k} \mathrm{E}\left[g_k(x_k,u_k) + J_{k+1}(f_k(x_k,u_k,w_k))\right] \\ &= \min_{u_k} g_k(x_k,u_k) + \sum_{j=1}^{n} J_{k+1}(j)\mathrm{P}\left[x_{k+1} = j | x_k, u_k\right] \end{aligned}$$