Computational Learning Theory **Fall Semester, 1999/00**

Lecture 5: November 18

Lecturer: Yishay Mansour *Scribe: Elhanan Borenstein, Keren Saggie, Yossi Mossel*

5.1 Introduction: Discounted Infinite Horizon

5.1.1 Notation

- $\vec{v} : S \Rightarrow R$
- $\|\vec{v}\|_\infty = \max_{s \in S}\{|\vec{v}(s)|\}$
- $H : S \times S \Rightarrow R$
- $\|H\| = \max_{s \in S} \sum_{j \in S} \|H_(s, j)\|$ (If H is a probabilities matrix, $\|H\| = 1$)

- **Probabilities Transition Matrix**

  Usually the topic of discussion will be a series of t steps, and the object of inquiry will be the probability of making a transition from state s to state j after t steps. The transition matrix after t steps, starting from state s, using policy π is:

  $P^t_\pi(j|s) = [P_{d_t} \ldots P_{d_2} \cdot P_{d_1}](j|s) = Prob_\pi(X_{t+1} = j|x_1 = s)$

- **Expectation of Reward**

  The expected value function after t steps, starting from state s, using policy π is:

  $E^\pi_s[\vec{v}(X_t)] = [P^{t-1}_\pi \cdot \vec{v}](s)$

- **The discounted value of policy π**

  $\vec{v}^\pi_\lambda = \sum_{t=1}^{\infty} \lambda^{t-1} P^{t-1}_\pi r_{d_t}$

  (where for deterministic policies, $r_{d_t}(s) = r(s, d_t(s))$ is the immediate reward for transition from s to $d_t(s)$)

Theorem 5.1 *Let Q be a matrix such that $\|Q\| < 1$, then*

1. *There exists $(I - Q)^{-1}$*
2. *$(I - Q)^{-1} = \lim_{N\to\infty} \sum_{i=0}^{N} Q^i$*

(The proof can be found in Puterman's book)

5.1.2 Assumptions

In this section we make the following simplifying assumptions.

1. The immediate reward and the transition probability are stationary. Hence the functions $r(s, a)$ and $p(j|s, a)$ are identical for any time stop. One benefit is that the algorithm can have a finite input.
2. The immediate reward is bounded: $|r(s, a)| < M$.
3. The discounted parameter is $0 \le \lambda < 1$
4. The number of states and actions is finite.

5.2 Calculating the Return Value of a Given Policy

According to Theorem 4.3 from the previous lecture, for each stochastic history dependent policy $\pi = (d_1, d_2, ...) \in \Pi^{HR}$ there exists a Markovian stochastic policy $\pi^{'} = (d_1^{'}, d_2^{'}, ...) \in \Pi^{MR}$ that has the same return, i.e., $v_\lambda^\pi = v_\lambda^{\pi^{'}}$.

Let $\pi \in \pi^{MR}$, then

$$v_\lambda^\pi(s) \quad = \quad E_s^\pi[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)] = \sum_{t=1}^{\infty} \lambda^{t-1} P_\pi^{t-1} r_{d_t}$$

$$\vec{v}_\lambda^\pi \quad = \quad \vec{r}_{d_1} + \lambda P_{d_1} [\underbrace{\vec{r}_{d_2} + \lambda P_{d_2} \vec{r}_{d_3} + \ldots}_{v_\lambda^{\pi^{'}}}]$$

(where $\pi^{'}$ is similar to policy π starting from the second step)

$$\vec{v}_\lambda^\pi = \vec{r}_{d_1} + \lambda P_{d_1} \vec{v}_\lambda^{\pi^{'}}$$

If π is stationary then $\pi^{'} = \pi$ and

$$\vec{v}^{\pi}_{\lambda} = \vec{r}_{d_1} + \lambda P_{d_1} \vec{v}^{\pi}_{\lambda}$$

All the parameters aside from $\vec{v}^{\pi}_{\lambda}$ are known, thus we have a set of linear equations of the form $\vec{x} = r_{d1} + \lambda P_{d_1} \vec{x}$. We will show that these equations have a single solution which is $\vec{v}^{\pi}_{\lambda}$.

5.2.1 Existence of a unique solution

We define a linear transformation L_d: $L_d \vec{v} = \vec{r}_d + \lambda P_d \vec{v}$.
Since $\vec{v}^{\pi}_{\lambda} = L_d \vec{v}^{\pi}_{\lambda}$, $\vec{v}^{\pi}_{\lambda}$ is a fixed point of L_d.

Theorem 5.2 *For $0 \le \lambda < 1$ and π a Markovian Stationary policy, $\vec{v}^{\pi}_{\lambda}$ is the unique solution for the equation set*

$$\vec{v} = \vec{r}_d + \lambda p_d \vec{v}$$

and is equal to

$$\vec{v}^{\pi}_{\lambda} = (I - \lambda P_d)^{-1} \vec{r}_d$$

Proof: We can write the equation set as

$$\vec{v}(I - \lambda P_d) = \vec{r}_d$$

Since P_d is a probability matrix, $||P_d|| = 1$, and as $\lambda < 1$, $||\lambda P_d|| < 1$.

According to Theorem 5.1, $(I - \lambda P_d)^{-1}$ exists. Thus, a solution $\vec{v} = (I - \lambda P_d)^{-1} \vec{r}_d$ exists.

By the same theorem,

$$\vec{v} = (I - \lambda P_d)^{-1} \vec{r}_d = \sum_{i=0}^{\infty} (\lambda P_d)^i \vec{r}_d = \sum_{i=0}^{\infty} \lambda^i P_d^i \vec{r}_d = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} \vec{r}_d = \vec{v}^{\pi}_{\lambda} \ .$$

We have shown that the solution is the discounted return value of policy π $\square$

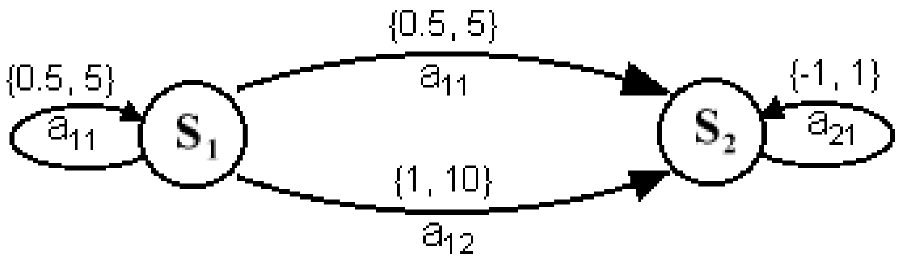

Figure 5.1: Example Diagram

5.2.2 Example:

(Consider the MDP in figure 5.1)

For a policy π, which picks a_{11} in S_1 and a_{21} in S_2 we compute the following values:

$$\begin{aligned} V(S_1) &= 5 + \lambda[\frac{1}{2}V(S_1) + \frac{1}{2}V(S_2)] \\ V(S_2) &= -1 + \lambda[1 \cdot V(S_2)] \end{aligned}$$

Or in matrix notation:

$$\vec{v} = \begin{pmatrix} 5 \\ -1 \end{pmatrix} + \lambda \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix} \vec{v}$$

Solutions are,

$$V(S_2) = -\frac{1}{1-\lambda}$$

$$V(S_1) = \frac{5 - \frac{\frac{1}{2}}{1-\lambda}}{1 - \frac{\lambda}{2}}$$

5.2.3 Properties of the transition matrix:

We show that the matrix $(I - \lambda P_d)^{-1}$ is order conserving.

Lemma 5.3 *The following holds for a probability matrix P and $0 \leq \lambda < 1$:*

1. *If* $\|\vec{u}\| \geq 0$ *then* $\|(I - \lambda P)^{-1}\vec{u}\| \geq \|\vec{u}\| \geq 0$
2. *If* $\|\vec{u}\| \geq \|\vec{v}\|$ *then* $\|(I - \lambda P)^{-1}\vec{u}\| \geq \|(I - \lambda P)^{-1}\vec{v}\|$

3. *If* $\|\vec{u}\| \geq 0$ *then* $\|\vec{u}^T(I-\lambda P)^{-1}\| \geq \|\vec{u}^T\| \geq 0$

Proof: Since $\|P\| = 1$ then $\|\lambda P\| \leq 1$. By theorem 5.1

$$(I-\lambda P_d)^{-1}\vec{u} = \vec{u} + \underbrace{(\lambda P)\vec{u} + (\lambda P)^2\vec{u} + \ldots}_{(sum\ of\ positive\ vectors)} \geq \vec{u} \geq 0$$

□

5.3 Computing the Optimal Policy

Having shown how to evaluate a given policy, we now turn to show how to find an optimal policy.

5.3.1 Optimality Equations

For a finite horizon we gave the following equations:

$$v_n(s) = \max_{a \in A_s}\{r(s,a) + \sum_{j \in S} p(j \mid s,a) v_{n+1}(j)\}$$

For a discounted infinite horizon we have similar equations. We will show that the Optimality equations are:

$$v(s) = \max_{a \in A_s}\{r(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v_(j)\}$$

First we show that maximizing over deterministic and stochastic policies yield the same value.

Theorem 5.4 *For all* $\vec{v}$ *and* $1 > \lambda \geq 0$

$$\max_{d \in \Pi^{MD}}\{r_d + \lambda P_d \vec{v}\} = \max_{d \in \Pi^{MR}}\{r_d + \lambda P_d \vec{v}\}$$

Proof:
Since $\Pi^{MD} \subseteq \Pi^{MR}$, the right side of the equality is at least as large as the left.
Now we show that the left side is at least as large as the right.
For $\pi \in \Pi^{MR}$ and v we define

$$\forall s \in S \qquad w_s(a) = r(s,a) + \lambda \sum_{j \in S} p(j \mid s,a) v(j)$$

Fix a state $s \in S$. The value of π is a weighted average of $w_s(a)$, and we have

$$\max_{a \in A_s}\{w_s(a)\} \geq \sum_{a \in A_s} q_\pi(a) w_s(a) \ .$$

Hence

$$\max_{d \in \Pi_{MD}} \{r_d + \lambda P_d \vec{v}\} \geq r_\pi + \lambda P_\pi \vec{v} \ ,$$

For any $\pi \in \Pi_{MR}$.
This shows that the left hand side in the theorem is at least as large as the right hand side, which completes the proof. □

Let us define the non-linear operator L:

$$L\vec{v} = \max_{d \in \Pi^{MD}} \{\vec{r}_d + \lambda P_d \vec{v}\}$$

Therefore we can state the optimality equation by

$$v = \max_{d \in \Pi^{MD}} \{\vec{r}_d + \lambda P_d \vec{v}\} = L\vec{v}$$

It still remains to be shown that these indeed are optimality equations, and that the fixed point of the operator is the optimal return value.

Theorem 5.5 *Let v^*_λ be the optimal return value with parameter $1 > \lambda \geq 0$*

1. *If $v \geq Lv$ then $v \geq v^*_\lambda$*
2. *If $v \leq Lv$ then $v \leq v^*_\lambda$*
3. *If $Lv = v$ then $v = v^*_\lambda$*

Proof: We start by proving (1)

$$\begin{aligned} v &\geq \max_{d \in \Pi^{MD}} \{r_d + \lambda P_d v\} \quad (\textit{given}) \\ &= \max_{d \in \Pi^{MR}} \{r_d + \lambda P_d v\} \quad (\textit{by theorem } 5.4) \end{aligned}$$

this implies that for any policy d,

$$\begin{aligned} v &\geq r_d + \lambda P_d v \\ &\geq r_d + \lambda P_d r_d + (\lambda P_d)^2 v \\ &\geq \sum_{i=0}^{n-1} (\lambda P_d)^i r_d + (\lambda P_d)^n v \quad (\textit{where } [P^0 = I]) \end{aligned}$$

For a given policy d

$$v_\lambda^d = \sum_{i=0}^{\infty} (\lambda P_d)^i r_d$$

By subtraction

$$v - v_\lambda^d \geq (\lambda P_d)^n v - \sum_{i=n}^{\infty} (\lambda P_d)^i r_d$$

Since $\lambda^n \|v\| \geq \|\lambda^n P_d^n v\|$ and $\lambda < 1$, we have that for any $\epsilon > 0$ there exists an $N > 0$, such that for all $n > N$ we have $\|\lambda^n P_d^n v\| \leq \frac{\epsilon}{2}$

Since $|\vec{r}_d| < M$ we can write:

$$-\frac{\lambda^n M}{1-\lambda} \cdot \vec{1} \leq \sum_{k=n}^{\infty} \lambda^k P_d^k r_d \leq \frac{\lambda^n M}{1-\lambda} \cdot \vec{1}$$

For a large enough n we have

$$\| \sum_{k=n}^{\infty} \lambda^k P_d^k r_d \| \leq \frac{\epsilon}{2} \ .$$

By this we derive that

$$\forall s \in S \quad v(s) \geq v_\lambda^d(s) - \epsilon$$

and for all d

$$v \geq v_\lambda^d - \epsilon$$

Thus

$$v \geq \max_{d \in \Pi^{MR}} \{v_\lambda^d\} - \epsilon = \max_{d \in \Pi^{MD}} \{v_\lambda^d\} - \epsilon = v_\lambda^* - \epsilon$$

As this is true $\forall \epsilon > 0$

$$v \geq v_\lambda^*$$

(If we assume that there is a state s such that $v(s) < v_\lambda^*(s)$ we pick $\epsilon = \frac{v_\lambda^* - v(s)}{2}$, and reach a contradiction)
We now prove (2)

Since $v \leq Lv$ there exists a policy d such that

$$v \leq r_d + \lambda P_d v$$

By theorem 5.2

$$v \leq (I - \lambda P_d)^{-1} r_d = v_\lambda^d$$

Hence

$$v \leq \max_d \{v_\lambda^d\}$$

Part (3) follows immediately from parts (1) and (2). $\square$

5.3.2 The Solution of the Optimality Equations:

Operator T is called contracting if there exists $0 \leq \lambda < 1$ such that

$$\|T\vec{u} - T\vec{v}\| \leq \lambda \|\vec{u} - \vec{v}\| \quad \forall \vec{u}, \vec{v} \in R^n$$

Theorem 5.6 *Let $T : R^n \to R^n$ a contracting operator, then*

1. *there exists a unique $\vec{v^*}$ such that $T\vec{v^*} = \vec{v}^*$*
2. *For each starting point $\vec{v}_0$ the series $\vec{v}_{n+1} = T\vec{v}_n$ converges to $\vec{v}^*$*

Proof: We define $\vec{v}_{n+1} = T\vec{v}_n$

Existence of a limit $\vec{v^*}$

$$\begin{aligned}
\|\vec{v}_{n+m} - \vec{v}_n\| &= \|\sum_{k=0}^{n-1} \vec{v}_{n+k+1} - \vec{v}_{n+k}\| \\
&\leq \sum_{k=0}^{m-1} \|\vec{v}_{n+k+1} - \vec{v}_{n+k}\| \quad (\textit{according to the triangle inequality}) \\
&= \sum_{k=0}^{m-1} \|T^{n+k}\vec{v}_1 - T^{n+k}\vec{v}_0\| \\
&\leq \sum_{k=0}^{m-1} \lambda^{n+k} \|\vec{v}_1 - \vec{v}_0\| \qquad (\textit{contraction } n+k \textit{ times}) \\
&= \frac{\lambda^n(1-\lambda^m)}{1-\lambda} \|\vec{v}_1 - \vec{v}_0\|
\end{aligned}$$

Since the coefficient decreases as n increases, $\forall \epsilon > 0 \;\; \exists N > 0$ such that
$\forall n \geq N, \|\vec{v}_{n+m} - \vec{v}_n\| < \epsilon$
Thus the series $\vec{v}_n$ has a limit.

Let us call this limit $\vec{v^*}$ and show that $\vec{v^*}$ is a fixed point of the operator T.

$\vec{v}^{\,*}$ is a fixed point

$$\begin{aligned} 0 &\leq \|T\vec{v^*} - \vec{v}^{\,*}\| \\ &\leq \|T\vec{v^*} - \vec{v}_n\| \quad + \quad \|\vec{v}_n - \vec{v^*}\| \ (\textit{according to the triangle inequality}) \\ &= \|T\vec{v^*} - T\vec{v}_{n-1}\| \quad + \quad \|\vec{v}_n - \vec{v^*}\| \\ &\leq \lambda\|\underbrace{\vec{v^*} - \vec{v}_{n-1}}_{\to 0}\| \quad + \quad \|\underbrace{\vec{v^n} - \vec{v^*}}_{\to 0}\| \end{aligned}$$

Since $\vec{v^*}$ is a limit of $\vec{v}_n$,

$$\lim_{n\to\infty} \|\vec{v}_n - \vec{v^*}\| = 0$$

hence

$$\|T\vec{v^*} - \vec{v^*}\| = 0$$

thus $\vec{v^*}$ is a fixed point of the operator L.

Uniqueness of $\vec{v^*}$

If

$$T\vec{v}_1 = \vec{v}_1,\ T\vec{v}_2 = \vec{v}_2,\ \textit{and}\ \vec{v}_1 \neq \vec{v}_2$$

then

$$\|T\vec{v}_1 - T\vec{v}_2\| = \|\vec{v}_1 - \vec{v}_2\| \leq \lambda\|\vec{v}_1 - \vec{v}_2\|$$

Hence $\lambda > 1$ in contradiction to the premises.
Thus $\vec{v^*}$ is unique. □

Next we will show that the operator L is a contracting operator.

Claim 5.7 $\forall\ 0 \leq \lambda < 1$, *$L$ is a contracting operator.*

Proof: For all $\vec{u}$, $\vec{v}$, we choose $s \in S$, and assume $L\vec{v}(s) \geq L\vec{u}(s)$.
We define a_s^* such that:

$$a_s^* \in argmax_{a\in A_s}\{r(s,a) + \lambda \sum_{j\in S} P(j|s,a)\vec{v}(j)\}$$

$$\begin{aligned}
0 &\leq L\vec{v}(s) - L\vec{u}(s) \\
&\leq \underbrace{r(s,a_s^*) + \lambda \sum_{j\in S} P(j|s,a)\vec{v}(j)}_{L\vec{v}(s)} - \underbrace{r(s,a_s^*) - \lambda \sum_{j\in S} P(j|s,a)\vec{u}(j)}_{not\ neccessarily\ the\ optimal\ action\ for\ u} \\
&= \lambda \sum_{j\in S} P(j|s,a)[\vec{v}(j) - \vec{u}(j)] \\
&\leq \lambda \sum_{j\in S} P(j|s,a)\|\vec{v} - \vec{u}\| \\
&= \lambda\|\vec{v} - \vec{u}\|
\end{aligned}$$

We have shown that

$$L\vec{v}(s) - L\vec{u}(s) \leq \lambda\|\vec{v} - \vec{u}\|$$

(The same proof holds for $L\vec{v}(s) \leq L\vec{u}(s)$)

Thus, for *all* $s \in S$,

$$\|L\vec{v} - L\vec{u}\| \leq \lambda\|\vec{v} - \vec{u}\|$$

Hence L is a contracting operator. □

Theorem 5.8 *Let $0 \leq \lambda < 1$ and S a finite set, then*

1. *There exists a unique solution v^* such that Lv^* and $v_\lambda^* = v^*$*
2. *For all $\pi \in \Pi^{MR}$ there exists a unique v such that $L_\pi v = v$ and $v_\lambda^\pi = v$*

Proof:

1. As L has been shown to be a contracting operator there is a unique solution for the equation $Lv = v$ by theorem 5.6. This fixed point is v_λ^*.
2. Is true by the same argument.

□

5.3.3 Example:

Using the same example we calculate the optimal return value to be:

$$\begin{aligned}
V(S_1) &= max\{5 + \lambda[\frac{1}{2}V(S_1) + \frac{1}{2}V(S_2)], 10 + \lambda V(S_2)\} \\
V(S_2) &= -1 + \lambda V(S_2)
\end{aligned}$$

Thus

$$\begin{aligned} V(S_2) &= -\frac{1}{1-\lambda} \\ V(S_1) &= max\{5 + \lambda[\frac{1}{2}V(S_1) - \frac{1}{2}\frac{1}{1-\lambda}],\ 10 - \lambda\frac{1}{1-\lambda}\} \end{aligned}$$

If we examine different values of λ we get different optimal actions in S_1. For example:

$$\begin{aligned} \lambda = 0 &\quad : \quad V(S_1)^* = 10 \quad V(S_2)^* = -1 \\ \lambda = \frac{1}{2} &\quad : \quad V(S_1)^* = 9 \quad V(S_2)^* = -2 \\ \lambda = \frac{9}{10} &\quad : \quad V(S_1)^* = 1 \quad V(S_2)^* = -10 \end{aligned}$$

Note that as λ increases the optimal policy at S_1 changes from a_{12} to a_{11}.