Computational Learning Theory	Fall Semester, 1999/00

Lecture 11: January 6

Lecturer: Yishay Mansour *Scribe: Eyal Even-dar, Doron Jacoby*

11.1 Large State Space

When we have a large state space and we can not compute $V(s,r)$, we would like to build an approximation function $\tilde{(}s,r)$. Let

$$\epsilon = \min_r \{||\tilde{V}(s,r) - V^*||\}$$

be the minimal distance between $\tilde{V}(s,r)$ and V^*. It is clear that ϵ is the upper bound for any approximation algorithm.(if ϵ is large, we can not expect a good approximation regardless the learning process). We will show later on error bounds from the type: $\frac{\epsilon}{(1-\lambda)^2}$ or $\frac{\epsilon}{(1-\lambda)}$
Theses bounds might seem disappointing since when $\lambda \to 1$ we will have a large number as our bound. On the other hand if we enrich our architecture (enlarge the family of r) $\epsilon \to 0$, and when λ is a constant the bound will also $\to 0$.
If we approximate $Q^*(s,a)$ by $\tilde{Q}^*(s,a)$ and $\tilde{Q}(s,a,r)$ is given then

$$\pi(s,r) = argmax_{a \in A_s}\{\tilde{Q}(s,a,r)\}$$

If we have $\tilde{V}(s,r)$, then

$$\pi(s,r) = argmax_{a \in A_s}\{r(s,a) + \lambda E_{s'}[\tilde{V}(s',r)]\}$$

We have to approximate $E_{s'}[\tilde{V}(s',r)]$, so we have more mistakes from the approximation too.
If we only few states S', than we compute exactly, otherwise we will approximate by taking samples. We will get a stochastic policy since every time we will get other sample, not like the case of Q where we get deterministic policy.

Theorem 11.1 *Consider a discounted problem, with parameter λ. If V satisfies*

$$\epsilon = ||V^* - V||,$$

and π is a greedy policy based on V, then

$$||V^\pi - V^*|| \leq \frac{2\lambda\epsilon}{1-\lambda}$$

Furthermore there exists δ s.t for every $\epsilon \leq \delta$ π is the optimal policy.

Proof:
Let

$$L_\pi V = r_\pi + \lambda P_\pi V$$

$$LV = \max_\pi \{r_\pi + \lambda P_\pi r\}$$

Then

$$\begin{aligned}
||V^\pi - V^*|| &= ||L_\pi V^\pi - V^*|| \\
&\leq ||L_\pi V^\pi - L_\pi V|| + ||L_\pi V - V^*|| \\
&\leq \lambda ||V^\pi - V|| + ||LV - LV^*|| \\
&\leq \lambda ||V^\pi - V|| + \lambda ||V^* - V|| + \lambda ||V - V^*||
\end{aligned}$$

$$\Longrightarrow ||V^\pi - V^*|| \geq || \frac{2\lambda\epsilon}{1-\lambda}$$

Second part:
Since we have finite number of policies, then there exist δ s.t.

$$\delta = \min_{\pi \neq \pi^*} \{||V^\pi - V^*||\}$$

For ϵ s.t.

$$\delta < \frac{2\lambda\epsilon}{1-\lambda},$$

It is

$$\pi = \pi^*$$

. ∎

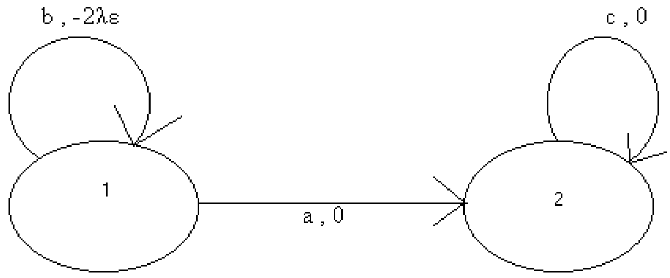

Figure 11.1: Example Diagram

11.1.1 Example of a tied bound

- The optimal policy is : $V^*(1) = V^*(2) = 0$
- Let $V(1) = +\epsilon, V(2) = -\epsilon$, than $||V - V^*|| = 2\epsilon$
- Greedy policy of V will give:

$$-2\epsilon\lambda + \lambda V(1) = -2\epsilon\lambda + \lambda\epsilon = -\epsilon\lambda$$

$$V^\pi(1) = \sum_i \lambda^i(-2\epsilon\lambda) = \frac{-2\epsilon\lambda}{1-\lambda}$$

11.1.2 Approximate Policy Iteration

The general structure is the same as in the Policy Iteration, except the following differences:

- We will not use V^π, instead we use $\tilde{V}^\pi$ (or $\tilde{Q}^\pi$), which is only an approximation of V^π. The reasons of using approximations are the architecture that may not be strong enough and the noise caused by the simulations.
- Let $\tilde{\pi}$ be the greedy policy of $\tilde{V}^\pi$. We will take π, which is close to $\tilde{\pi}$.

Those differences are a source for an error.

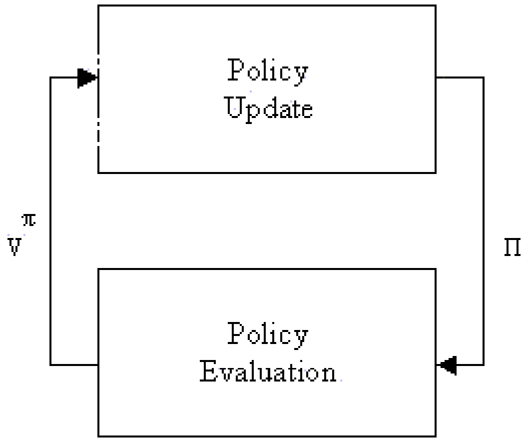

Figure 11.2: Regular Policy Iteration

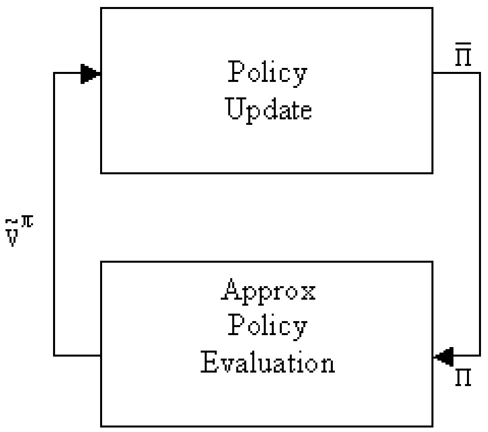

Figure 11.3: Approximate Policy Iteration

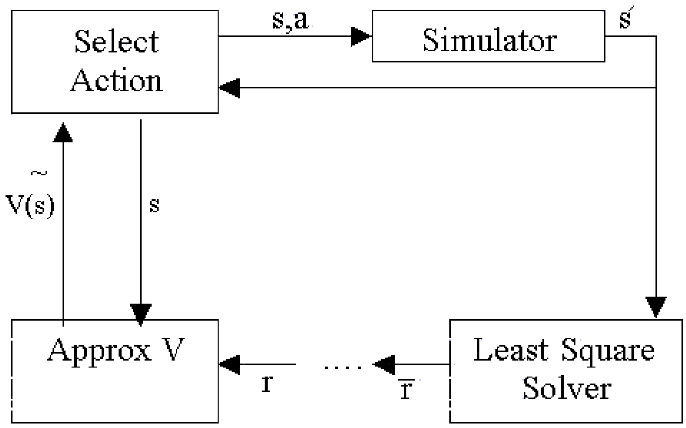

Figure 11.4: Diagram for a mechnism that produce Approximate Policy Iteration

The algorithm using MonteCarlo method

- Since we have too much states, lets take only subset of the states - $\tilde{S}$.
- $\forall s \in \hat{s}$, there are M(s) runs : c(s,1) ... c(s,M(s)).
- We look for r s.t.

$$\sum_{s\in\tilde{s}} \sum_{i=1}^{M(s)} (\tilde{V}^{\pi}(s) - c(s,i))^2$$

  will be minimal.

Solving the Least-Squares Problem

Let $\tilde{(S)}$ be a set of representive states, M(s) the samples of the cost V^{π}, the mth such sampled is denoted by $C(s,m)$ and r is the vector parameter upon which the following optimization problem is solved.

$$\min_r \sum_{s\in\tilde{S}} \sum_{m=1}^{M(s)} (\tilde{V}(s,r) - C(s,m))$$

The solution can be obtained by an incremental algorithm, which performs steps in the gradient direction.
We will have the following equation for a certain run $(s_1, a_1, ..., s_n)$.

$$\vec{r} = \vec{r} - \alpha \sum_{k=0}^{|\tilde{S}|} \nabla_r \tilde{V}(s,r)(\tilde{V}(s,r) - C(s,k))$$

Evaluation Of Approximate Policy Iteration

In this section will make two assumptions on our approximation. (Both of them are pretty strong.) We will create a sequence of $V_1, \pi_1, V_2, \ldots$

1. $\forall k ||V_k - V^{\pi_k}|| < \epsilon$
2. $\forall k ||L_{\pi_{k+1}} V_k - LV_k|| < \delta$

where ϵ and δ are positive scalars.

Theorem 11.2 *The sequence of Policies π_k generated by the approximate policy iteration algorithm satisfies*

$$\lim_{k \to \infty} sup||V^{\pi_k} - V^*|| \leq \frac{\delta + 2\lambda\epsilon}{(1-\lambda)^2}$$

Proof: We will first show that a policy generated by the policy update can not be much worse than the current policy. We will prove it by using the following lemmas

Lemma 11.3 *Let π be some policy and V is a value function satisfies $||V - V^\pi|| \leq \epsilon$ for some $\epsilon \geq 0$.*
Let $\bar{\pi}$ be a policy that satisfies:

$$(L_{\bar{\pi}} V)(s) \geq (LV)(s) - \delta$$

for some $\delta \geq 0$ then

$$V^{\bar{\pi}}(s) \geq V^{\pi}(s) - \frac{\delta + 2\lambda\epsilon}{(1-\lambda)}$$

Proof: Let

$$\beta = \max_s \{V^\pi(s) - V^{\bar{\pi}}(s)\}$$

and

$$\vec{1} = (1, 1, \ldots.1)^t$$

so

$$V^{\bar{\pi}}(s) \geq V^\pi - \beta * \vec{1}$$

Hence we will have

$$V^{\bar{\pi}} = L_{\bar{\pi}} V^{\bar{\pi}}$$

$$\leq L_V^{\pi} - \beta * \vec{1} = L_{\bar{\pi}} V - (\lambda\beta) * \vec{1}$$

1. $V^{\bar{\pi}} - L_{\bar{\pi}} V + (\lambda\beta) * \vec{1} \geq 0$

2. $0 \geq -L_{\bar{\pi}} V + LV - \delta * \vec{1} \geq -L_{\bar{\pi}} V + L_{\pi} V - \delta * \vec{1}$

Using inequality 1, we obtain

$$V_{\pi} - V_{\bar{\pi}} \leq V_{\pi} - L_{\bar{\pi}} V^{\pi} + (\lambda\beta) * \vec{1}$$

Using inequality 2, we obtain

$$\leq -L_{\bar{\pi}} V^{\pi} + (\lambda\beta) * \vec{1} + L_{\bar{\pi}} V - L_{\pi} V + \delta * \vec{1}$$

$$= (L_{\bar{\pi}} V - L_{\bar{\pi}} V^{\pi}) + (V_{\pi} - L_{\pi} V) + (\delta + \beta\lambda) * \vec{1}$$

$$\leq \lambda ||V^{\pi} - V|| * \vec{1} + \lambda ||V - V^{\pi}|| * \vec{1} + (\delta + \beta\lambda) * \vec{1}$$

$$= 2 * \lambda * \epsilon + \delta + \beta\lambda$$

Thus we conclude the following:

$$||V^{\pi} - V\bar{\pi}|| = \beta \leq 2 * \lambda * \epsilon + \delta + \beta\lambda$$

$$\beta \leq \frac{2\lambda\epsilon + \delta}{1 - \lambda}$$

and the desired result follows
□

Let β_k be

$$\max_s (V^{\pi_{k+1}} - V^{\pi_k})$$

We apply this lemma with $\pi = \pi_k$ and $\bar{\pi} = \pi_{k+1}$
As a result we have

$$\beta_k \leq \frac{2\lambda\epsilon + \delta}{1 - \lambda}$$

We now let γ_k to be distance from the optimum. $\gamma_k = \max_s (V^*(s) - V^{\pi_k}(s))$

Lemma 11.4 $\forall k \gamma_{k+1} \leq \lambda\gamma_k + \lambda\beta_k + \delta + 2\lambda\epsilon$

$$V_{\pi_k} \geq V^* - \gamma_k$$

$$LV_{\pi_k} \geq L(V^* - \gamma_k)$$

we then have (assumption 1)

$$L_{\pi_k} V^{\pi_{k+1}} \geq L_{\pi_{k+1}}(V_k - \epsilon)$$

$$= L_{\pi_{k+1}} V_k - LV_k - \lambda\epsilon * \vec{1}$$

(assumption 2)

$$\geq LV_k - \delta * \vec{1} - \lambda\epsilon * \vec{1}$$

(assumption 1)

$$\geq L(V^{\pi_k} - \epsilon) + (-\delta - \lambda\epsilon) * \vec{1}$$

$$= LV^{pi_k} + (-\delta - 2\lambda\epsilon) * \vec{1}$$

(definition of γ_k)

$$\geq L(V^* - \gamma_k) + (-\delta - 2\lambda\epsilon) * \vec{1}$$

$$= V^* + (-\delta - 2\lambda\epsilon + \lambda\gamma_k) * \vec{1}$$

Thus
(fixed pint of the operator)

$$V^{\pi_{k+1}} = L_{\pi_{k+1}} V^{\pi_{k+1}}$$

(definition of β_k)

$$\geq L_{\pi_{k+1}}(V^{\pi_k} - \beta_k)$$

$$= L_{\pi_{k+1}} V^{\pi_k} - \lambda\beta_k * vec1)$$

(using the previous)

$$\geq V^* + (-\delta - 2\lambda\epsilon + \lambda\gamma_k - \lambda\beta_k) * \vec{1}$$

finally

$$\Longrightarrow V^* - V^{\pi_{k+1}} \leq (\delta + 2\lambda\epsilon + \lambda\gamma_k + \lambda\beta_k) * \vec{1}$$

We will use the former lemma and the β_k definition to prove the theorem Since one can easily see from the equation

$$\beta_k \leq \frac{\delta + 2\epsilon\lambda}{1-\lambda}$$

that bounding is not dependent on K. Therefore, we can define the following

$$\alpha = \frac{\delta + 2\epsilon\lambda}{1-\lambda}\lambda + \delta + 2\epsilon\lambda = \frac{\delta + 2\epsilon\lambda}{1-\lambda}$$

while α is constant, which is not dependent on K. Then we have

$$\gamma_{k+1} \leq \lambda\gamma_k + \alpha$$

By opening the recursion we will get

$$\lambda^k\gamma_1 + \sum_{i=0}^{k-1} \lambda^i\alpha$$

BY taking the limit superior of the equation as $k \to \infty$ to obtain

$$\lim_{k\to\infty} \gamma_k = \frac{\delta + 2\epsilon\lambda}{(1-\lambda)^2}$$

which proves the theorem □

11.1.3 Approximate Value Iteration

We will make the following assumption.

$$||V_{k+1} - LV_k|| \leq \epsilon$$

$$LV_0 - \epsilon * \vec{1} \leq V_1 \leq LV_0 - \epsilon * \vec{1}$$

Activating L (operator) on the inequality

$$LV_0^2 - \lambda\epsilon * \vec{1} \leq LV_1 \leq LV_0^2 - \lambda\epsilon * \vec{1}$$

We have also the next inequality

$$LV_1 - \epsilon * \vec{1} \leq V_2 \leq LV_1 - \epsilon * \vec{1}$$

Using both inequalities

$$LV_0^2 - (\lambda\epsilon + \epsilon) * \vec{1} \leq LV_1 \leq LV_0^2 - (\lambda\epsilon + \epsilon) * \vec{1}$$

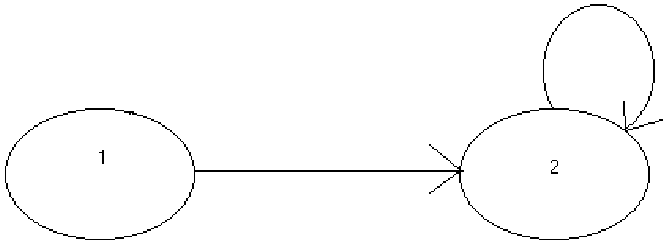

Figure 11.5: Example 2 Diagram

Thus for each k we will have

$$V_k - L^k V_0|| \leq \epsilon \sum_i = 0^k - 1\lambda^I \leq \frac{\epsilon}{1-\lambda}$$

If we look at Therefore:

$$||\tilde{V} - V^*| \leq \frac{\epsilon}{1-\lambda}$$

Although calculations are much simpler than in PI. The method is less natural.

11.1.4 Example

We will show a MDP, where the approximate value iteration does not converge.
All the rewards equal zero.V(1) = V(2) = 0

$$\tilde{V}(1,r) = r\tilde{V}(2,r) = 2r$$

One can see that fro r = 0 we have the value function.
We will calculate the square error.

$$\min_r [\tilde{V}(1,r) - \lambda\tilde{V}(2,r)]^2 + \tilde{V}(2,r) - \lambda\tilde{V}(2,r)]^2$$

In such simple case the minimum can be easily found

$$\min_r [(r - 2\lambda r_k)^2 + (2r - 2\lambda r_k)^2]$$

$$2(r - 2\lambda r_k) + 4(2r - 2\lambda r_k)$$

Hence

$$r = \frac{6}{5}\lambda r_k$$

Since $r_k = (\frac{6}{5}\lambda)^k$ For $\lambda > \frac{5}{6}r_k \rightarrow \infty$ We have shon an example for a value function, which does not converge. We will look to see if our assumption was not satisfied

$$||V_{k+1} - LV_k]|| = \max\{|r_{k+1} - 2\lambda r_k|, |2r_{k+1} - 2\lambda r_k|\} = \max\{frac65\lambda r_k, frac125\lambda r_k\}$$

The error is a function of r_k and therefore we do not have an upper bound and the assumption is not staisfied.