Reinforcement Learning **Spring Semester, 2018/9**

Lecture 5: March 27, 2019

Lecturer: Yishay Mansour *Scribe: ym*

Until now we looked at planning problems, where we are given a complete model of the MDP, and the goal is to either evaluate a given policy or compute the optimal policy. In this lecture we will start looking at learning problems, where we need to learn from interaction. This lecture will concentrate on *model based* learning, where the main goal is to learn an accurate model. Next lecture we will look at *model free* learning, where we learn a policy without recovering the actual underlying model.

5.1 Effective horizon of discounted return

Before we start looking a learning, we will show an "reduction" from discounted return to finite horizon return. The main issue will be to show that the discounted return has an *effective horizon* such that rewards beyond it have a negligible effect on the discounted return.

Theorem 5.1. *Given a discount factor γ, the discounted return in the first $T = \frac{1}{1-\gamma} \log \frac{R_{max}}{\epsilon(1-\gamma)}$, is within ϵ of the total discounted return.*

Proof. Recall that the rewards are $r_t \in [0, R_{max}]$. Fix an infinite sequence of rewards $(r_0, \ldots, r_t, \ldots)$. We would like to consider the following difference

$$\Delta = \sum_{t=1}^{\infty} r_t \gamma^t - \sum_{t=0}^{T-1} r_t \gamma^t = \sum_{t=T}^{\infty} r_t \gamma^t \leq \frac{\gamma^T}{1-\gamma} R_{max}$$

We like this difference Δ to be bounded by ϵ, hence

$$\frac{\gamma^T}{1-\gamma} R_{max} \leq \epsilon \ .$$

This implies that

$$T \log(1/\gamma) \geq \log \frac{R_{max}}{\epsilon(1-\gamma)} \ .$$

We can bound $\log(1/\gamma) = \log(1 + \frac{1-\gamma}{\gamma}) \leq \frac{1-\gamma}{\gamma}$. Since $\gamma < 1$, the theorem follows. □

5.2 Off-Policy Model-Based Learning

We consider the case that we are given as input a sequence of trajectories. Essentially, our input will be composed from quadruples:

$$(s, a, r, s')$$

where r is sampled from $R(s, a)$ and s' is sampled from $p(\cdot|s, a)$.

Our goal is to output an MDP $(S, A, \widehat{r}, \widehat{p})$, where S is the set of states, A is the set of actions, $\widehat{r}(s, a)$ is the approximate expected reward of $R(s, a) \in [0, R_{max}]$, and $\widehat{p}(s'|s, a)$ is the approximate probability of reaching state s' when we are in state s and doing action a.

5.2.1 Mean estimation

We start with a basic mean estimation problem (as you have seen in the Introduction to Machine Learning course). Suppose we are give a random variable $R \in [0, 1]$ and would like to approximate its mean $\mu = E[R]$. We observe m samples of R, $r_1, \ldots, r_m$, and compute their mean $\widehat{\mu} = \frac{1}{m} \sum_{i=1}^{m} r_i$.

By the law of large numbers we know that when m goes to infinity we have that $\widehat{\mu}$ converges to μ. We would like to have concrete finite convergence bounds, mainly to derive the value of m as a function of the desired accuracy ϵ.

For this we use concentration bounds (known as Chernoff-Hoffding bounds), which bounds the additive error as follows:

$$\Pr[|\mu - \widehat{\mu}| \geq \epsilon] \leq 2e^{-2\epsilon^2 m}$$

We can also derive relative error bounds, that guarantee for $\epsilon \in (0, 1)$,

$$\Pr[\widehat{\mu} \leq (1 - \epsilon)\mu] \leq e^{-\epsilon^2 m/2}$$

and

$$\Pr[\widehat{\mu} \geq (1 + \epsilon)\mu] \leq e^{-\epsilon^2 m/3}$$

Using the additive concentration bound, we have that for $m \geq \frac{1}{2\epsilon^2} \log(2/\delta)$ we have $|\mu - \widehat{\mu}| \leq \epsilon$, with probability $1 - \delta$.

We can now use this bound in order to estimate the expected rewards. For each state-action (s, a) let $\widehat{r}(s, a) = \frac{1}{m} \sum_{i=1}^{m} R_i(s, a)$ be the average of m samples. We can show the following:

Claim 5.1. *If we have $m \geq \frac{R_{max}}{2\epsilon^2} \log \frac{2|S|\,|A|}{\delta}$ samples for each state action (s, a), then with probability $1 - \delta$ we have for every (s, a) that $|r(s, a) - \widehat{r}(s, a)| \leq \epsilon$.*

Proof. We will need to scale the random variables to $[0, 1]$, which will be achieved by dividing by R_{max}. By the Chernoff-Hoffding bound, for each (s, a) we have that with probability $1 - \frac{\delta}{|S|\,|A|}$ that $|\frac{r(s,a)}{R_{max}} - \frac{\widehat{r}(s,a)}{R_{max}}| \leq \frac{\epsilon}{R_{max}}$.

We bound the probability over all state-action pais using a union bound over all state action pairs.

$$\Pr[\exists(s,a) : |\frac{r(s,a)}{R_{max}} - \frac{\widehat{r}(s,a)}{R_{max}}|| > \frac{\epsilon}{R_{max}}] \leq \sum_{(s,a)} \Pr[|\frac{r(s,a)}{R_{max}} - \frac{\widehat{r}(s,a)}{R_{max}}|| > \frac{\epsilon}{R_{max}}] \leq \sum_{(s,a)} \frac{\delta}{|S|\,|A|} = \delta$$

Therefore, we have that with probability $1-\delta$ for every (s, a) simultaneously we have $|r(s, a) - \widehat{r}(s, a)| \leq \epsilon$. □

5.2.2 Influence of reward estimation errors

We would like to quantify the influence of having inaccurate estimates of the rewards. We will look both at the finite horizon return and the discounted return. We start with the case of finite horizon.

Influence of reward estimation errors: Finite horizon

Fix a policy $\pi \in MD$. We want to compare the return using $r_t(s, a)$ versus $\widehat{r}(s, a)$ and $r_T(s)$ versus $\widehat{r}_T(s)$. We will assume that for every (s, a) and t we have $|r_t(s, a) - \widehat{r}_t(s, a)| \leq \epsilon$ and $|r_T(s) - \widehat{r}_T(s)| \leq \epsilon$. We will show that the difference in return is bounded by $\epsilon(T + 1)$, where T is the finite horizon.

Define the expected return of π with the true expectations

$$J_T^\pi(s_0) = E^{\pi,s_0}[\sum_{t=0}^{T} r_t(s_t, a_t) + r_T(s_T)].$$

and with the estimated expectations

$$\widehat{J}_T^\pi(s_0) = E^{\pi,s_0}[\sum_{t=0}^{T} \widehat{r}_t(s_t, a_t) + \widehat{r}_T(s_T)].$$

We are interested in bounding the difference between the two

$$error(\pi) = |J_T^\pi(s_0) - \widehat{J}_T^\pi(s_0)|.$$

Note that in both cases we use the true transition probability. For a given trajectory $\sigma = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$ we define

$$error(\pi, \sigma) = \sum_{t=0}^{T} r_t(s_t, a_t) + r_T(s_T)) - \sum_{t=0}^{T} \widehat{r}_t(s_t, a_t) + \widehat{r}_T(s_T))$$

taking the expercaction over trajectories we have

$$error(\pi) = |E^{\pi, s_0}[error(\pi, \sigma)]|$$

Theorem 5.2. *Assume that for every (s, a) and t we have $|r_t(s, a) - \widehat{r}_t(s, a)| \leq \epsilon$ and for every s we have $|r_T(s) - \widehat{r}_T(s)| \leq \epsilon$. Then, for any policy $\pi \in MD$ we have $error(\pi) \leq \epsilon(T + 1)$.*

Proof. The probability of each trajectory $\sigma = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$ is the same under $r_t(s, a)$ and $\widehat{r}_t(s, a)$, since $\pi \in MD$ implies that π depends only on the time t and state s_t. For each trajectory $\sigma = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$, we have,

$$\begin{aligned}
error(\pi, \sigma) &= |\sum_{t=0}^{T} r_t(s_t, a_t) + r_T(s_T) - \sum_{t=0}^{T} \widehat{r}_t(s_t, a_t) + \widehat{r}_T(s_T)| \\
&= |\sum_{t=0}^{T} (r_t(s_t, a_t) - \widehat{r}_t(s_t, a_t)) + (r_T(s_T) - \widehat{r}_T(s_T))| \\
&\leq \sum_{t=0}^{T} |r_t(s_t, a_t) - \widehat{r}_t(s_t, a_t)| + |r_T(s_T) - \widehat{r}_T(s_T)| \\
&\leq \epsilon T + \epsilon
\end{aligned}$$

The theorem follows since $error(\pi) = |E^{\pi, s_0}[error(\pi, \sigma)]| \leq \epsilon(T + 1)$. □

Computing approximate optimal policy: finite horizon

We can now describe how to compute a near optimal policy for the finite horizon case. We start with the sample requirement. We need a sample of size $m \geq \frac{1}{2\epsilon^2} \log \frac{2|S|\,|A|\,T}{\delta}$ for each $R_t(s, a)$ and $R_T(s)$. Given the sample, we compute $\widehat{r}_t(s, a)$ and $\widehat{r}_T(s)$. As we saw, with probability $1 - \delta$ we have $|r_t(s, a) - \widehat{r}_(s, a)| \leq \epsilon$ and $|r_T(s) - \widehat{r}_T(s)| \leq \epsilon$. Now we can compute the optimal policy $\widehat{\pi}^*$ for the estimated rewards $\widehat{r}_t(s, a)$ and $\widehat{r}_T(s)$. The main goal is to show that $\widehat{\pi}^*$ is a near optimal policy.

Claim 5.2. *Assume that for every (s,a) and t we have $|r_t(s,a) - \widehat{r}_t(s,a)| \leq \epsilon$ and for every s we have $|r_T(s) - \widehat{r}_T(s)| \leq \epsilon$. Then,*

$$J_T^{\pi^*}(s_0) - J_T^{\widehat{\pi}^*} \leq 2\epsilon(T+1)$$

Proof. From the definition of $error(\pi)$ and since for any $\pi \in MD$ we showed that $error(\pi) \leq \epsilon(T+1)$, we have,

$$J_T^{\pi^*}(s_0) - \widehat{J}_T^{\pi^*}(s_0) \leq error(\pi^*) \leq \epsilon(T+1)$$

and

$$\widehat{J}_T^{\widehat{\pi}^*}(s_0) - J_T^{\widehat{\pi}^*}(s_0) \leq error(\widehat{\pi}^*) \leq \epsilon(T+1)$$

Since $\widehat{\pi}^*$ is optimal for $\widehat{r}_t$ we have

$$\widehat{J}_T^{\pi^*}(s_0) \leq \widehat{J}_T^{\widehat{\pi}^*}(s_0)$$

The claim follows by adding the three inequalities. □

Influence of reward estimation errors: discounted return

Fix a policy $\pi \in SD$. Again, define the expected return of π with the true expectations

$$J_\gamma^\pi(s_0) = E^{\pi,s_0}[\sum_{t=0}^{\infty} r(s_t,a_t)\gamma^t]$$

and with the estimated expectations

$$\widehat{J}_\gamma^\pi(s_0) = E^{\pi,s_0}[\sum_{t=0}^{\infty} \widehat{r}(s_t,a_t)\gamma^t]$$

We are interested in bounding the difference between the two

$$error(\pi) = |J_\gamma^\pi(s_0) - \widehat{J}_\gamma^\pi(s_0)|$$

Note that as for the finite horizon, in both cases we use the true transition probability. For a given trajectory $\sigma = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$ we define

$$error(\pi,\sigma) = \sum_{t=0}^{\infty} \gamma^t r_t(s_t,a_t) - \sum_{t=0}^{\infty} \gamma^t \hat{r}_t(s_t,a_t)$$

Theorem 5.3. *Assume that for every (s,a) we have $|r_t(s,a) - \widehat{r}_t(s,a)| \leq \epsilon$. Then, for any policy $\pi \in SD$ we have $error(\pi) \leq \frac{\epsilon}{1-\gamma}$.*

Proof. The probability of each trajectory $\sigma = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$ is the same under $r(s,a)$ and $\widehat{r}(s,a)$, since $\pi \in SD$. For each trajectory $\sigma = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$, we have,

$$\begin{aligned} error(\pi, \sigma) &= |\sum_{t=0}^{\infty} r(s_t, a_t)\gamma^t - \sum_{t=0}^{\infty} \widehat{r}(s_t, a_t)\gamma^t| \\ &= |\sum_{t=0}^{\infty} (r(s_t, a_t) - \widehat{r}(s_t, a_t))\gamma^t| \\ &\leq \sum_{t=0}^{\infty} |r(s_t, a_t) - \widehat{r}(s_t, a_t)|\gamma^t \\ &\leq \frac{\epsilon}{1-\gamma} \end{aligned}$$

The theorem follows since $error(\pi) = |E^{\pi, s_0}[error(\pi, \sigma)]| \leq \frac{\epsilon}{1-\gamma}$. □

Computing approximate optimal policy: discounted return

We can now describe how to compute a near optimal policy for the discounted return. We need a sample of size $m \geq \frac{R_{max}}{2\epsilon^2} \log \frac{2|S|\,|A|}{\delta}$ for each $R(s,a)$. Given the sample, we compute $\widehat{r}(s,a)$. As we saw, with probability $1-\delta$ we have for every (s,a) that $|r(s,a) - \widehat{r}(s,a)| \leq \epsilon$. Now we can compute the policy $\widehat{\pi}^*$ for the estimated rewards $\widehat{r}_t(s,a)$. Again, the main goal is to show that $\widehat{\pi}^*$ is a near optimal policy.

Claim 5.3. *Assume that for every (s,a) we have $|r(s,a) - \widehat{r}(s,a)| \leq \epsilon$. Then,*

$$J_\gamma^{\pi^*}(s_0) - J_\gamma^{\widehat{\pi}^*} \leq \frac{2\epsilon}{1-\gamma}$$

Proof. From the definition of $error(\pi)$ and since for any $\pi \in SD$ we showed that $error(\pi) \leq \frac{\epsilon}{1-\gamma}$, we have,

$$J_\gamma^{\pi^*}(s_0) - \widehat{J}_\gamma^{\pi^*}(s_0) \leq error(\pi^*) \leq \frac{\epsilon}{1-\gamma}$$

and

$$\widehat{J}_\gamma^{\widehat{\pi}^*}(s_0) - J_\gamma^{\widehat{\pi}^*}(s_0) \leq error(\widehat{\pi}^*) \leq \frac{\epsilon}{1-\gamma}$$

Since $\widehat{\pi}^*$ is optimal for $\widehat{r}$ we have

$$\widehat{J}_\gamma^{\pi^*}(s_0) \leq \widehat{J}_\gamma^{\widehat{\pi}^*}(s_0)$$

The claim follows by adding the three inequalities. □

5.2.3 Estimating the transition probabilities

We now estimate the transition probabilities. Again, we will look at the observed model. Namely, for a given state-action (s, a), we consider the transitions (s, a, s'_i), for $1 \leq i \leq m$. We define the observed transition distribution,

$$\widehat{p}(s'|s, a) = \frac{|\{i : s'_i = s'\}|}{m}$$

Our main goal would be to evaluate the observed model as a function of the sample size m.

We start by considering two Markov chains which differ by at most α for each transition. Namely, we have Markov chains M_1 and M_2, where $M_1[i, j] = \Pr[i \to j] = p_1(j|i)$ and $M_2[i, j] = \Pr[i \to j] = p_2(j|i)$. We assume that for every i, j we have $|M_1[i, j] - M_2[i, j]| \leq \alpha$. We would like to relate α to the statistical distance between the state distributions generated by the two Markov chains. Clearly if $\alpha \approx 0$ then the probabilities will be almost identical, but we like to get a quantitative bound on the difference, which will allow us to derive the sample size m.

We start with a general well-known observation about distributions.

Theorem 5.4. *Let q_1 and q_2 be two distributions over S. Let $f : S \to [0, F_{max}]$. Then,*

$$|E_{s\sim q_1}[f(s)] - E_{s\sim q_2}[f(s)]| \leq F_{max}\|q_1 - q_2\|_1$$

where $\|q_1 - q_2\|_1 = \sum_{s\in S} |q_1(s) - q_2(s)|$.

Proof.

$$\begin{aligned}
|E_{s\sim q_1}[f(s)] - E_{s\sim q_2}[f(s)]| &= |\sum_{s\in S} f(s)q_1(s) - \sum_{s\in S} f(s)q_2(s)| \\
&\leq \sum_{s\in S} f(s)|q_1(s) - q_2(s)| \\
&\leq F_{max}\|q_1 - q_2\|_1
\end{aligned}$$

□

Therefore, we would like to bound the L_1-norm between the state distributions generated by M_1 and M_2.

Theorem 5.5. *Let q_1^t and q_2^t be the distribution over states after trajectories of length t of M_1 and M_2, respectively. Then,*

$$\|q_1^t - q_2^t\|_1 \leq \alpha|S|t$$

Proof. Let x_0 be the distribution of the start state. Then $q_1^t = x_0 M_1^t$ and $q_2^t = x_0 M_2^t$. The proof is by induction on t. Clearly, for $t = 0$ we have $q_1^0 = q_2^0 = x_0$.

We start with a basic facts about matrix norms. Let $\|M\|_\infty = max_i \sum_j |M[i,j]|$. Then

$$\|zM\|_1 = \sum_j |\sum_i z[i]M[i,j]| \leq \sum_{i,j} |z[i]| \cdot |M[i,j]| \leq \sum_i |z[i]| \cdot \|M\|_\infty \leq \|z\|_1 \|M\|_\infty \tag{5.1}$$

This implies the following two simple facts. First, let q be a distribution, i.e., $\|q\|_1 = 1$, and M a matrix with all the entries at most α, i.e., $|M[i,j]| \leq \alpha$ which implies $\|M\|_\infty \leq \alpha|S|$. Then,

$$\|qM\|_1 \leq \|q\|_1 \|M\|_\infty \leq \alpha|S| \tag{5.2}$$

Second, let M be a row-stochastic matrix, implies that $\|M\|_\infty = 1$. Then,

$$\|zM\|_1 \leq \|z\|_1 \|M\|_\infty \leq \|z\|_1 \tag{5.3}$$

For the induction step, let $z^t = q_1^t - q_2^t$.

$$\begin{aligned}
\|q_1^t - q_2^t\|_1 &= \|x_0 M_1^t - x_0 M_2^t\|_1 \\
&= \|q_1^{t-1} M_1 - (q_1^{t-1} + z^{t-1}) M_2\|_1 \\
&\leq \|q_1(M_1 - M_2)\|_1 + \|z^{t-1} M_2\|_1 \\
&\leq \alpha|S| + \alpha|S|(t-1) = \alpha|S|t
\end{aligned}$$

where in the last inequality, for the first term we used the first fact and for the second term we used the second fact with the inductive claim. □

Approximate model and simulation lemma

We define an *α-approximate model* as follows. A model $\widehat{M}$ is an α-approximate model of M if for every state-action (s,a) we have: (1) $|\widehat{r}(s,a) - r(s,a)| \leq \alpha$ and (2) for every s' we have $|\widehat{p}(s'|s,a) - p(s'|s,a)| \leq \alpha$.

The following simulation lemma, for the finite horizon case, guarantees that approximate models have similar return.

Lemma 5.1. *Assume that model $\widehat{M}$ is an α-approximate model of M. For the finite horizon return, for any policy $\pi \in MD$, we have*

$$|J_T^\pi(s_0; M) - J_T^\pi(s_0; \widehat{M})| \leq \epsilon$$

for $\alpha \leq \frac{\epsilon}{R_{max}|S|T^2}$

Proof. By Theorem 5.5 the distance between the state distributions of M and $\widehat{M}$ at time t s bounded by $\alpha|S|t$. Since the maximum reward is R_{max}, by Theorem 5.4 the difference is bounded by $\sum_{t=0}^{T} \alpha|S|tR_{max} \leq \alpha|S|T^2R_{max}$. For $\alpha \leq \frac{\epsilon}{R_{max}|S|T^2}$ implies that the difference is at most ϵ. □

We now switch to the simulation lemma, for the discounted return case, which also guarantees that approximate models have similar return.

Lemma 5.2. *Assume that model $\widehat{M}$ is an α-approximate model of M. For the discounted return, for any policy $\pi \in MD$, we have*

$$|J_\gamma^\pi(s_0; M) - J_\gamma^\pi(s_0; \widehat{M})| \leq \epsilon$$

for $\alpha \leq \frac{c\epsilon(1-\gamma)^2}{R_{max}|S|\log^2(R_{max}/(\epsilon(1-\gamma)))}$, for some constant $c > 0$.

Proof. We briefly sketch the proof. We will use the proof for the finite horizon case. which has $\alpha = \frac{\epsilon/2}{R_{max}|S|T^2}$ to get error at most $\epsilon/2$.

We can now use the effective horizon of the discounted case which is $T = \frac{1}{1-\gamma}\log\frac{R_{max}}{\epsilon(1-\gamma)/2}$ which guarantees that the error increases by at most $\epsilon/2$. We can use this is the finite horizon return bound to derive the lemma. □

putting it all together

We want with high probability $(1-\delta)$ to have accuracy α. To get accuracy α we need $m = O(\frac{1}{\alpha^2}\log(|S|^2|A|/\delta))$ samples for each state-action pair (s, a). Plugging in the value of α we have, for the finite horizon,

$$m = O(\frac{R_{max}^2}{\epsilon^2}|S|^2T^4\log(|S|\ |A|/\delta))$$

and for the discounted return

$$m = O(\frac{R_{max}^2}{\epsilon^2}|S|^2\frac{1}{(1-\gamma)^4}\log(|S|\ |A|/\delta)\log^2\frac{R_{max}}{\epsilon(1-\gamma)})$$

Assume we have a sample of m for each (s, a). Then with probability $1 - \delta$ we have an α-approximate model $\widehat{M}$. We find an optimal policy $\widehat{\pi}^*$ for $\widehat{M}$. This implies that $\widehat{\pi}^*$ is a 2ϵ-optimal policy. Namely,

$$|J^*(s_0) - J^{\widehat{\pi}^*}(s_0)| \leq 2\epsilon$$

We can now look on the dependency of the sampling bounds on the parameters.

1. The error scales like $\frac{R_{max}^2}{\epsilon^2}$ which looks like the right bound, even for estimation of random variables expectations.
2. The dependency on the horizon is necessary, although it is probably not optimal.
3. The dependency on the number on the number of states $|S|$, is due to the fact that we like a very high approximation of the next state distribution. Simply we need to approximate $|S|^2$ entries, so for this the bound is reasonable. However, we will show that for approximation the optimal policy we can do better.

5.2.4 Improved sample bound: using approximate value iteration

Recall, that the value iteration works as follows. Initially, we set the values arbitrarily,

$$V_0 = \{V_0(s)\}_{s \in S}$$

In iteration n we compute

$$\begin{aligned} V_{n+1}(s) &= \max_{a \in A}\{r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V_n(s')\} \\ &= \max_{a \in A}\{r(s, a) + \gamma E_{s' \sim p(\cdot|s,a)}[V_n(s')]\} \end{aligned}$$

We showed that $\lim_{n \to \infty} V_n = V^*$, and that the convergence rate is $O(\frac{\gamma^n}{1-\gamma} R_{max})$. This implies that if we run for N, where $N = \frac{1}{1-\gamma} \log \frac{R_{max}}{\epsilon(1-\gamma)}$, we have an error of at most ϵ.

We would like to approximate the value iteration algorithm using a sample. Namely, for each (s, a) we have a sample of size m, i.e., $\{(s, a, s'_i)\}_{i \in [1,m]}$ The value iteration using the sample would be,

$$\widehat{V}_{n+1}(s) = \max_{a \in A}\{r(s, a) + \gamma \frac{1}{m} \sum_{i=1}^{m} \widehat{V}_n(s'_i)\}$$

The intuition is that if we have a large enough sample, it will approximate the value iteration. We set m such that for every (s, a) and any iteration $n \in [1, N]$ we have:

$$|E[\widehat{V}_n(s')] - \frac{1}{m}\sum_{i=1}^{m} \widehat{V}_n(s'_i)| \leq \epsilon$$

and also

$$|\widehat{r}(s, a) - r(s, a)| \leq \epsilon$$

We can have this for $m = O(\frac{V_{max}^2}{\epsilon^2})$ where V_{max} bound the maximum value. I.e., for finite horizon $V_{max} = TR_{max}$ and for discounted return $V_{max} = \frac{R_{max}}{1-\gamma}$.

Assume that we have, for every $s \in S$,

$$|\widehat{V}_n(s) - V_n(s)| \leq \lambda$$

Then

$$|\widehat{V}_{n+1}(s) - V_{n+1}(s)| \leq \epsilon + \gamma\lambda \leq \lambda$$

where the last inequality holds for $\lambda \geq \frac{\epsilon}{1-\gamma}$.

Therefore, if we sample $m = O(\frac{1}{\epsilon^2} \log \frac{N|S|\,|A|}{\delta})$, we have that with probability $1 - \delta$ for every (s, a) the approximation is at most ϵ. This implies that the approximate value iteration has error at most $\lambda = \frac{\epsilon}{1-\gamma}$.

The main result is that we can run Value Iteration algorithm for N iterations and approximate well the optimal value function and policy. The implicit drawback is that we are approximating only the optimal policy, and cannot evaluate an arbitrary policy.

5.3 On-Policy Learning

5.3.1 Learning Deterministic Decision Process

Recall that a Deterministic Decision Process is modeled by a directed graph, where the states are the vertices, and each action is associate with an edge.

We first define the observed model. Given an observation set $\{(s_t, a_t, r_t, s_{t+1})\}$, we define an observed model $\widehat{M}$, where $\widehat{f}(s_t, a_t) = s_{t+1}$ and $\widehat{r}(s, a) = r_t$. For (s, a) which do not appear in the observation set, we define $\widehat{f}(s, a) = s$ and $\widehat{r}(s, a) = R_{max}$.

First we claim that for any $\pi \in SD$ the completion can only increase the value.

$$\widehat{V}^{\pi}(s; \widehat{M}) \geq V^{\pi}(s; M)$$

The increase holds for any trajectory, and note that once π reaches (s, a) that was not observed, its reward will be R_{max} forever. (This is since $\pi \in SD$.)

We can now present the on-policy learning algorithm.

1. At time T let $\{(s_t, a_t, r_t, s_{t+1}) : 0 \leq t \leq T-1\}$ be the previous observations.
2. Build the observed model $\widehat{M}_T$.
3. Compute $\widehat{\pi}^*_T$, the optimal policy for $\widehat{M}_T$.
4. Use $a_T = \widehat{\pi}^*_T(s_T)$.
5. Observe the reward r_T and the next state s_{T+1}.

We first show that the model $\widehat{M}_T$ cannot change too often.

Lemma 5.3. *The observed model $\widehat{M}_T$ can change at most $|S|\ |A|$ times.*

Proof. Each time we change the observed model, we add one (s, a) to be known. Since there are $|S|\ |A|$ such pairs, this bounds the number of changes. □

Next we show that we either make progress in the next $|S|$ steps or we never make any more changes.

Lemma 5.4. *Either we change $\widehat{M}_T$ in some time $t \in [T, T+|S|]$ or we never change $\widehat{M}_T$.*

Proof. The model is deterministic. If we do not make any update in the next $|S|$ steps, the policy $\widehat{\pi}^*_T \in MD$ will close a cycle and continue on this cycle forever. Hence, the model will never change. □

We can now state the convergence of the algorithm to the optimal policy.

Theorem 5.6. *After $T = |S|^2|A|$ time steps $\widehat{\pi}^*_T$ is optimal.*

Proof. We showed that the number of changes is at most $|S|\ |A|$, and the time between changes is at most $|S|$. This implies that after time $T = |S|^2|A|$ we never change.

The return of $\widehat{\pi}^*_T$ after time T is identical in $\widehat{M}_T$ and M, since all the edges it traverses are known. Let $\widehat{V}^*$ be its return. Assume that the policy π^* has a strictly higher return in M, say $V^* > \widehat{V}^*$. This implies that the return of π^* is at least $V^* > \widehat{V}^*$ in $\widehat{M}_T$, since the rewards in $\widehat{M}_T$ are always at least those in M. This contradicts the fact that $\widehat{\pi}^*_T$ is optimal for $\widehat{M}_T$. □

5.3.2 On-policy learning MDP

Similar to the DDP, we will use the principle of *Optimism in face of uncertainty*. Namely, we substitute the unknown quantities by the maximum possible values.

Similar to DDP, we will partition the state-action pairs (s, a) `known` and `unknown`. The main difference is that in a DDP it is sufficient to have a single sample to move (s, a) from `unknown` to `known`. In a general MDP we need have a larger sample to move (s, a) from `unknown` to `known`. Otherwise, the algorithm would be very similar.

We describe algorithm `R-MAX`, which performs on-policy learning of MDPs.

Initialization: Initially we set for each (s, a) a next state distribution which always return to s, i.e., $p(s|s, a) = 1$ and $p(s'|s, a) = 0$ for $s' \neq s$. We set the reward to be maximal, i.e., $r(s, a) = R_{max}$. We mark (s, a) to be `unknown`.

Execution: At time t. (1) Build a model $\widehat{M}_t$, explained later. (2) Compute $\widehat{\pi}^*_t$ the optimal policy for $\widehat{M}_t$, and (3) Execute $a_t = \widehat{\pi}^*_t(s_t)$ and observe r_t and s_{t+1}.

Building a model At time t, if the number of samples of (s, a) is *exactly* m for the *first time*, then: modify $p(\cdot|s, a)$ to the observed transition distribution, $r(s, a)$ to the average observed reward for (s, a) and mark (s, a) as `known`. Note that we update each (s, a) only once, when it moves from `unknown` to `known`.

Here is the basic intuition for algorithm `R-MAX`. We consider the discounted return. Fix a large enough horizon T, at least the effective horizon. Assume we run $\widehat{\pi}^*_t$ for T time steps. Either, we explore a state-action (s, a) which is `unknown`, in this case we make progress on the exploration. Also this can happen at most $m|S|\ |A|$ times. Alternatively, we did not reach any state-action (s, a) which is `unknown`, in which case we are optimal on the observed model, and near optimal on the true model.

For the analysis consider the event W, which is that event that we visit a state-action (s, a) which is `unknown` during the block of T time steps. For the return of $\widehat{\pi}^*_t$ we have,

$$V^{\widehat{\pi}^*_t} \geq V^* - \Pr[W]\frac{R_{max}}{1-\gamma} - \lambda$$

where λ is the approximation error, and we can bound it by $\epsilon/2$ by setting m large enough.

We consider two cases, depending on the probability of W. First, we consider the case that the probability of W is small. If $\Pr[W] \leq \frac{\epsilon(1-\gamma)/2}{R_{max}}$, then $V^{\widehat{\pi}^*_t} \geq V^* - \epsilon/2 - \epsilon/2$, since we assume that $\lambda \leq \epsilon/2$.

Second, we consider the case that the probability of W is large. If $\Pr[W] > \frac{\epsilon(1-\gamma)/2}{R_{max}}$. Then, there is a good probability to visit an `unknown` (s, a), but this can happen at most $m|S|\ |A|$. Therefore, the expected number of such blocks is at most $m|S|\ |A|\frac{R_{max}}{\epsilon(1-\gamma)/2}$.

This implies that only in

$$m|S|\ |A|\frac{R_{max}}{\epsilon(1-\gamma)/2}$$

blocks, the algorithm $\mathtt{R}-\mathtt{MAX}$ will be more than ϵ sub-optimal, i.e., have an expected return less than $V^*-\epsilon$.

5.4 Bibliography Remarks

The first polynomial time model based learning algorithm is E^3 of [3]. While we did not outline the E^3 algorithm, we did describe the effective horizon and the simulation lemma from there.

The improved sampling bounds for the optimal policy using approximate value iteration is following [2].

The `R-MAX` algorithm is due to [1].

Analysis of related models, especially the `PAC-MDP` model appears in [5, 4].

Bibliography

[1] Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[2] Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998]*, pages 996–1002, 1998.

[3] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[4] Lihong Li. *Sample Complexity Bounds of Exploration*, pages 175–204. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[5] Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite mdps: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.