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

Lecture 7: December 9

Lecturer: Prof. Yishay Mansour — *Scribe: Boshy Shlomy, Berkman Omer*

7.1 Learning from experience

we consider the case in which the exact model is not known.
In this case we develop methods for learning the model using statistical methods based on sampling and experience.

7.1.1 Average Calculation

Let $V_1, \ldots, V_n$ be independent identical distributed (i.i.d) random variables.
We assume $|V_i| \le M$.
The average can be written as :

$\hat{V}_N = \frac{1}{N}\sum_{i=1}^{N} V_i$

or as a recursive online formula :

$\hat{V}_k = (1-\frac{1}{k})\hat{V}_{k-1} + \frac{1}{k}V_k$
(Since: $(1-\frac{1}{k})\hat{V}_{k-1} = \frac{k-1}{k}(\frac{1}{k-1}\sum_{i=1}^{k-1} V_i)$)

Convergence

According to the Law of Large Numbers the average converges to the expected value :

$\hat{V}_n \longrightarrow \bar{v} = E[V_i]$, note that clearly
$E[\hat{V}_n] = \bar{v}$.

The variance is :

$Var(\hat{V}_n) = \frac{1}{N^2}\sum_{i=1}^{N} Var(V_i) = \frac{\sigma^2}{N}$, where
$\sigma^2 = Var(V_i)$.

This implies that when N goes to infinity we have,

$\lim_{N\to\infty} \frac{\sigma^2}{N} = 0$

so as the sample size increases, the difference between the average and the expected value decreases.

Convergence rate

Using Chernoff's theorem :

- $Prob[|\hat{V_N} - \bar{v}| \geq \gamma] \leq 2e^{-2N\gamma^2}$
- $Prob[\hat{V_N} > (1+\gamma)\bar{v}] \leq e^{-N\bar{v}\gamma^2/3}$
- $Prob[\hat{V_N} < (1-\gamma)\bar{v}] \leq e^{-N\bar{v}\gamma^2/2}$

Note that convergence rate is exponential in the sample size.

7.1.2 Weighted average

The weighted average can be defined by:

$\hat{V} = \sum_{i=1}^{N} \alpha_i V_i$, where the constants α_i have the properties that,
$\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$

The expected value of the weighted average is,

$E[\hat{V}] = E[\sum_{i=1}^{N} \alpha_i V_i] = \sum_{i=1}^{N} \alpha_i E[V_i] = \bar{v}$

and the variance is :

$Var(\hat{V}) = E[(\sum_i \alpha_i V_i - \bar{v})^2] = E[(\sum_i \alpha_i (V_i - \bar{v}))^2]$
$= \sum_i \alpha_i^2 E(V_i - \bar{v})^2 + \sum_{i \neq j} \alpha_i \alpha_j E[(V_i - \bar{v})(V_j - \bar{v})]$

for $i \neq j$:
$E[(V_i - \bar{v})(V_j - \bar{v})] = E(V_i - \bar{v}) E(V_j - \bar{v})$
since sample i is independent of sample j. We know that $E(V_i - \bar{v}) = 0$ for all i, so :

$Var(\hat{V}) = \sum_i \alpha_i^2 E(V_i - \bar{v})^2 = \sum_i \alpha_i^2 \sigma^2$,

where σ^2 is the variance of a single variable.

If $\lim_{N\to\infty} \sum_i \alpha_i^2 = 0$ then $\lim_{N\to\infty} Var(\hat{V_N}) = 0$.

7.2 Stochastic Models

Generally, we would like to solve systems of the form $H\vec{r} = \vec{r}$ by finding a fix point of the operator H where $H : R^n \to R^n$.
One way of solving is to perform iterations of the form $r_{n+1} \leftarrow Hr_n$.
Note that if $\{r_n\}$ has a limit such that $r_n \longrightarrow r^*$ and H is continuous around r^* then $Hr^* = r^*$.

Remark : if H is a contracting operator then we showed that such a limit always exists.

An equivalent way for iteration is :

$$r_{n+1} \leftarrow (1-\alpha)r_n + \alpha Hr_n$$

which has the same convergence property.

Let us assume that H is not known, or hard to compute. We can replace H by a sample of the form $S = Hr + W$, where W is the sample's "noise" and $E(W) = 0$.

Such an S can be given by simulation of the system or by a random experiment. We can use S instead of Hr and get the following iterative algorithm :

$$r_{n+1} \leftarrow (1-\alpha)r_n + \alpha(Hr_n + W_n)$$

Such an algorithm is called *Stochastic Approximation.*

7.2.1 Using α to control the convergence rate

Decreasing the parameter α causes the noise influence to decrease, but on the other hand the convergence is slower. Usually we will use α-values such that: $\lim_{i\to\infty} \alpha_i = 0$.

Let us look at one *Typical Example:*

- Let V be a random variable dependent of the parameter r.
- There exists a probability function : $P(V|r)$
- A function $g(r, V)$

We would like to solve : $E_V[g(r,V)] = r$, where the expectation regards $p(V|r)$.

We can write iterations of the form :

$r_{n+1} \leftarrow (1-\alpha)r_n + \alpha E[g(r_n, V)]$.

Computing the expected value can be complicated, but if $P(V|r)$ is known and can be simulated or sampled, we can get samples $\tilde{V}$ and use them to evaluate $E_V[g(r,V)]$. For example we can sample $\tilde{V_1},...,\tilde{V_k}$ and compute $\frac{1}{k}\sum_{i=1}^{k} g(r_n, \tilde{V_i})$ as the observed average.

If K is large, then the average will be close to its expected value $E_V[g(r,V)]$ and in such a case we should expect a similar behaviour. The other extrem is the case of $k=1$ (one sample), $\tilde{V_n}$.

$r_{n+1} \leftarrow (1-\alpha_n)r_n + \alpha_n g(r_n, \tilde{V_n})$ where $\tilde{V_n}$ is distributed by $p(v|r)$.

This algorithm is called the *Robbins-Morro algorithm* . The iterations can be written as:

$r_{n+1} \leftarrow (1-\alpha_n)r_n + \alpha_n[E(g(r,V) + g(r,\tilde{V}) - E(g(r,V)]$

where we can define :
$Hr_n = E[g(r_n, V)]$,
$W_n = g(r, \tilde{V}) - E(g(r,V)$. (and $E[W_n] = 0$ by definition of $\tilde{V}$)
This implies that for each $s \in S$ we have,

$r_{n+1}(s) \leftarrow (1-\alpha_n)r_n(s) + \alpha_n(s)[(Hr_n)(s) + W_n(s)]$.

7.2.2 Choosing $\alpha_n(s)$

Since :

$|r_n(s) - r_0(s)| \leq \sum_i |r_{i+1}(s) - r_i(s)| \leq \sum_i \alpha_(i)(s)|(Hr_i)(s) - r_i(s) + W_i(s)| \leq A\sum_{i=1}^{n} \alpha_i(s)$
(where A is a constant that bounds $|(Hr_i)(s) - r_i(s) + W_i(s)|$).

We need to require that $\sum_i \alpha_i(s) = \infty : \forall s \in S$. Otherwise, we have to assume that the distance $||r^* - r_0||$ is bounded. The above condition insures that any starting point converges to the optimal value.

We add a condition to insure that the converge rate is fast enough :

$\forall S \in S : \sum_i \alpha_i^2 < \infty$

One simple choice could be : $\alpha_i(s) = \frac{1}{i}$

7.2.3 Evaluating the average

Let $V_1, ..., V_n$ be i.i.d. random variables with expected value of $\bar{v} = E[V_i]$.
We would like to solve the system $E[V] = r$ (this is a special case of the earlier discussion in which $g(V, r) = V$). We will use a single sample V_n and get :

$r_{n+1} \leftarrow (1 - \alpha_n) r_n + \alpha_n V_n$

if $r_n \longrightarrow r^*$ then $\bar{v} = E[V] = r^*$.

We compute the variance as follows :

$Var(r_{n+1}) = (1 - \alpha_n)^2 Var(r_n) + \alpha_n^2 Var(V_n)$.

We show that if $\sum_n \alpha_n = \infty$ and $\alpha_n \longrightarrow 0$ then $Var(r_{n+1}) \longrightarrow 0$. The following is a general Lemma that is used and is shown later.

Lemma 7.1 *Let δ_t be a series such that $\delta_t \longrightarrow 0, \sum_t \delta_t = \infty, 0 \leq \delta_t \leq 1$. and Let e_t be a series such that $e_t \geq 0$ and : $e_{t+1} \leq (1 - \delta_t) e_t + c\delta_t^2$. Then $e_t \longrightarrow 0$.*

Proof: *step 1:* we show that for every constant ϵ, $e_t < \epsilon$ for an infinite number of t's.
Let us assume (by contradiction) that there exists T such that for all $t > T$, $e_t > \epsilon$ and $c\delta_t \leq \epsilon/2$. (since $\delta_t \longrightarrow 0$). In such a case we have

$e_{t+1} \leq (1 - \delta_t) e_t + c\delta_t^2 = e_t - \delta_t e_t + c\delta_t^2 \leq e_t - \delta_t \epsilon + \delta_t \epsilon/2 = e_t - \delta_t \epsilon/2$.

Hence :

$e_m \leq e_T - \epsilon/2 * \sum_{t=1}^{m-1} \delta_t$.

Since for any m we have $e_m \geq 0$,then it has to be the case that $\sum_{i=0}^{\infty} \delta_t < \infty$, which is a contradiction to the assumption in the lemma.
Therefore, for any $\epsilon > 0$, there exists T such that $\forall t > T : c\delta_t < \epsilon$ and $e_T \leq \epsilon$.

step 2: we show that $\forall t > T : e_t \leq \epsilon$.
$e_{T+1} \leq (1 - \delta_T) e_T + c\delta_T^2 \leq (1 - \delta_T)\epsilon + \epsilon\delta_T \leq \epsilon$

Under the hypothesis that $e_T \leq \epsilon$ we showed that $e_{T+1} \leq \epsilon$, this implies that
$\forall t > T : e_t \leq \epsilon$. □
So the variance could be expressed as :

$$Var(r_{n+1}) = (1-\alpha_n)^2 Var(r_n) + \alpha_n^2 Var(V_n) \leq (1-\alpha_n) Var(r_n) + \alpha_n^2 Var(V_n)$$

we can substitute :

- $\alpha_n \equiv \delta_n$
- $C \equiv Var(V_n)$
- $e_n \equiv Var(r_n)$

Since $\sum_n \alpha_n = \infty$ and $\alpha_n \longrightarrow 0$, we have that $Var(V_n) \longrightarrow 0$.

Remark : the condition $\sum_n \alpha_n^2 \longrightarrow 0$ will insure convergence with probability one.

We say that $\lim_{n\to\infty} X_n = X$ with probability one if and only if
$lim_{n\to\infty} \sup_{n\leq m} |X_m = X| = 0$.

7.3 Evaluating Policy Reward

Let π be a policy. We would like to calculate the reward of policy π : $V^\pi(s)$.

For simplicity, we assume that there exist a state s_0 in the MDP, such that s_0 has a reward of 0, and $Prob(s_0|s_0, a) = 1 : \forall a \in A_{s_0}$.
Also, we assume that each policy reachs state s_0 within a finite number of steps with probability 1.
Under these assumptions, we can assume each run is finite.

7.3.1 The Naive Approach

for each $s \in S$:
Run π from s for m times, where the i-th run is T_i.
Let r_i be the reward of T_i.
Estimate the reward of policy π starting from s, by :

$$\hat{V^\pi}(s) = Avg(r_i) = \frac{1}{m}\sum_{i=1}^{m} r_i$$

The variables r_i are independent since the runs T_i are independent. By Chernoff's theorem :

$Prob[|\hat{V^{\pi}}(s) - V^{\pi}(s)| \geq \gamma] \leq 2e^{-2m\gamma^2}$

This implies that :

$Prob[\exists s \in S : |\hat{V^{\pi}}(s) - V^{\pi}(s)| \geq \gamma] \leq \sum_s Prob[|\hat{V^{\pi}}(s) - V^{\pi}(s)| \geq \gamma] \leq |S| 2e^{-2m\gamma^2} = \epsilon.$

for $\gamma = \sqrt{\frac{ln(|S|)/\epsilon}{2m}}$ the above holds.

7.3.2 First Visit

We can improve the approximations by updating rewards of many states in each single run.
for each $s \in S$:

1. Run π from s for m times, where the i-th run is T_i.
2. For each run T_i and state s in it, let $r(s, T_i)$ be the reward of π in run T_i from the first appearance of s in T_i until the run ends (reaching state s_0).
3. Let the reward of policy π starting from s be : $\hat{V^{\pi}}(s) = Avg_{s \in T_i}(r(s, T_i))$

The random variables $r(s, T_i)$ for a given state s and different T_i's, are independent since different runs are independent. The improvement is based on increasing the number of samples without increasing the number of runs, so we have a smaller estimation-error.

7.3.3 Every Visit

We can hope to improve the approximations by using every appearance of state s in each single run.
for each $s \in S$:

1. Run π from s for m times, where the i-th run is T_i.
2. For each run T_i and state s in it, let $r(s, T_i, j)$ be the reward of π in run T_i from the j-th appearance of s in T_i until the run ends (reaching state s_0).
3. Let the reward of policy π starting from s be : $\hat{V^{\pi}}(s) = Avg(r(s, T_i, j))$

The problem is that the random variables $r(s, T_i, j)$ are *dependent* for different j's.

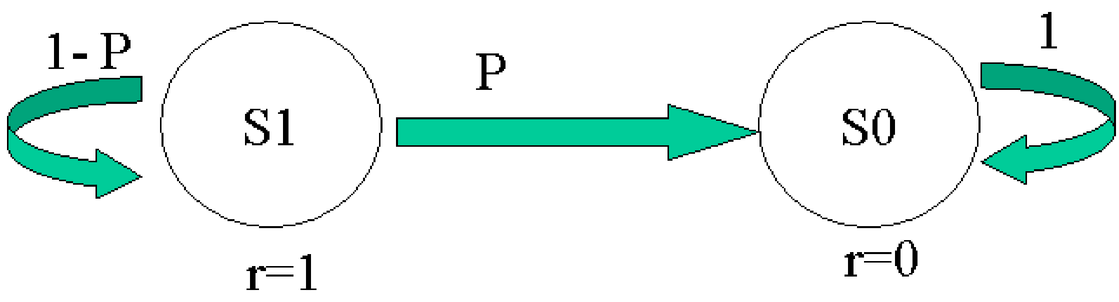

Figure 7.1: Example Diagram

7.3.4 Example

Consider the MDP in the Figure 7.1. Let N be the length of the run. We choose $\gamma = 1$ (undiscounted). It is easy to see that,

$V^{\pi}(s_1) = \frac{1}{p} = E[N]$,

since the distribution is geometric ("success" with probability p when moving to s_0).
However note that $r(s_1, T, j) \geq r(s_1, T, j+1)$, which implies that $r(s_1, T, j)$ are independent.
The return from the k-th appearance of S_1 until the end of the run is $N - k$ (there are N steps in the run, each of them with reward 1 and the run ends the first time we get to S_0).
The average is $\frac{1}{N}\sum_{k=0}^{N-1}(N-k) = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$
Hence $E[\frac{N+1}{2}] = \frac{\frac{1}{p}+1}{2} = \frac{1+p}{2p}$.
Note that this is different from $E[V^{\pi}(s_1)] = \frac{1}{p}$.
The difference (bias) is because the random variables are dependent.

Error Calculations

The fact that *Every Visit* is biased does not implies that it is inferier to *First Visit* , as an estimate. Consider using a single run to estimate the return in the above example.

- First Visit :

  $Var(\hat{V_F}(S_1)) = E[N^2] - E^2[N] = \frac{1}{p^2} - \frac{1}{p}$

  Note that the variance in this case can be bigger than the expected value when p is small. for example if p=0.1 then the expected value is 10 and variance is 90.

- Every Visit :

$E[(\frac{N+1}{2} - \frac{1}{p})^2] = \frac{1}{2p^2} - \frac{3}{4p} + \frac{1}{4}$

For each $p < 1$, the expected square error in *Every Visit* is smaller than in *First Visit* , meaning that Every Visit creates a biased estimation but it's variance is much smaller. The reason is that *Every Visit* contains many more samples. Note that in *Every Visit* the average is computed on all the runs, and not within the runs.