Lecture Note 9

1 Explicit Explore or Exploit (E^3) Algorithm

Last lecture, we studied the Q-learning algorithm:

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \gamma_t \left[g_{a_t}(x_t) + \alpha \min_{a'} Q_t(x_{t+1}, a') - Q_t(x_t, a_t) \right].$$

An important characteristic of Q-learning is that it is a *model-free* approach to learning an optimal policy in an MDP with unknown parameters. In other words, there is explicit attempt to model or estimate costs and/or transition probabilities — the value of each action is estimated directly through the Q-factor.

Another approach to the same problem is to estimate the MDP parameters from the data and find a policy based on the estimated parameters. In this lecture, we will study one such algorithm — the Explicit Explore or Exploit (E^3) algorithm, proposed by Kearns and Singh [1].

The main ideas for E^3 are as follows:

- we divide states in two sets:

N	*known* states
N^C	*unknown* states

- known states have been visited sufficiently many times to ensure that $\hat{P}_a(x, y)$, $\hat{g}_a(x)$ are "accurate" with high probabilities
- an unknown state is moved to N when it has been visited at least m times for some number m

We introduce two MDPs $\hat{M}_N$ and M_N. The MDP $\hat{M}_N$ is presented in Fig. 1. Its main characteristic is that the unknown states from the original MDP are merged into a recurrent state x_0 with cost $g_a(x_0) = g_{\max}$, $\forall\, a$. The other MDP M_N has the same structure as $\hat{M}_N$ but the estimated transition probabilities and costs are replaced with their true values.

We now introduce the algorithm.

1.1 Algorithm

We will first consider a version of E^3 which assumes knowledge of J^*; the assumption will be lifted later. The E^3 algorithm proceeds as follows.

1. Let $N = \emptyset$. Pick arbitrary state x_0. Let $k = 0$.

2. If $x_k \notin N$, perform "balanced wandering:"

$$a_k = \text{ action chosen fewest times at state } x_k$$

If $x_k \in N$, then

attempt exploitation: If the optimal policy π^* for $\hat{M}_N$ has $\hat{J}_{\hat{M}_N}(x_k) \leq J^*(x_k) + \frac{\epsilon}{2}$, stop.
Return x_k and $\pi^*_{\hat{M}_N}$

attempt exploration: Follow policy $\hat{\pi}_{\mathcal{S}_0}$ for T steps where $T = \frac{1}{1-\alpha}$.

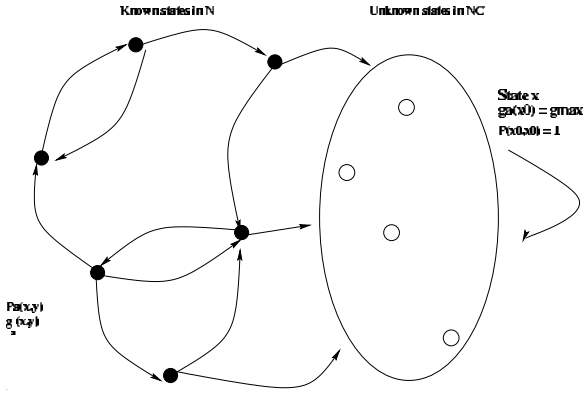

Figure 1: Markov Decision Process $\hat{M}_n$

Theorem 1 *With probability no less than $1-\delta$, E^3 will stop after a number of actions and computation time*

$$poly\left(\frac{1}{\epsilon}, \frac{1}{\delta}, |\mathcal{S}|, \frac{1}{1-\alpha}, g_{\max}\right)$$

and return a state x and policy u such that $J_u(x) \leq J^*(x) + \epsilon$.

1.2 Main Points

The main points used for proving Theorem 1 are as follows:

(i) There exists m that is polynomially bounded such that, if all states in N have been visited at least m times, then $\hat{M}_N$ is sufficiently close to M_N.

(ii) Balanced wandering can only happen finitely many times.

(iii) (a) $J_{u,M_N}(x) \geq J_u(x)$

(b) $\|J_{u,M_N} - J_{u,\hat{M}_N}\|_\infty \leq \frac{\epsilon}{2}$ with high probability

(iv) If exploitation is not possible, then there is an exploration policy that reaches an unknown state after T transitions with high probability.

To show the first main point, we consider the following lemma.

Lemma 1 *Suppose a state x has been visited at least m times with each action $a \in A_x$ having been executed at least $\lfloor \frac{m}{|\mathcal{A}_x|} \rfloor$ times. Then, if*

$$m = poly\left(|\mathcal{S}|, \frac{1}{1-\alpha}, T, g_{\max}, \frac{1}{\epsilon}, \log\frac{1}{\delta}, var(g)\right)$$

we have, w.p. $\geq 1-\delta$,

$$\begin{aligned} |\hat{P}_a(x,y) - P_a(x,y)| &= O\left(\epsilon\left(\frac{1-\alpha}{|\mathcal{S}|g_{\max}}\right)^2\right) \\ |\hat{g}_a(x) - g_a(x)| &= O\left(\epsilon\left(\frac{1-\alpha}{|\mathcal{S}|g_{\max}}\right)^2\right) \end{aligned}$$

The proof of this lemma is a directly application of the Chernoff bound, which states that, if $z_1, z_2, \ldots$ are i.i.d. Bernoulli random variables, then

$$\frac{1}{n}\sum_{i=1}^{n} z_i \to \mathrm{E}z_1 \quad \text{(SLLN)}$$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathrm{E}z_1\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

The main point (ii) follows from pigeonhole principle:

after $(m-1)|\mathcal{S}|$ balanced wandering steps, at least one state will have to become known

The main point iii(a) follows from the next lemma.

Lemma 2 *For all policy* u,

$$J_{u,M_N}(x) \geq J_u(x), \forall\, x.$$

Proof: Trivial for $x \notin N$ since $J_{u,M_N}(x) = \frac{g_{\max}}{1-\alpha} \geq J_u(x)$. If $x \in N$, take $T = \inf\{t : x_t \notin N\}$. Then

$$\begin{aligned} J_u(x) &= \mathrm{E}\left[\sum_{t=0}^{T-1}\alpha^t g_u(x_t) + \sum_{t=T}^{\infty}\alpha^t g_u(x_t)\right] \\ &\leq \mathrm{E}\left[\sum_{t=0}^{T-1}\alpha^t g_u(x_t) + \alpha^T\frac{g_{\max}}{1-\alpha}\right] \\ &= J_{u,M_N}(x) \end{aligned}$$

□

To prove the main point iii(b), we first introduce the following definition.

Definition 1 *Let* M *and* $\hat{M}$ *be two MDPs. Then* $\hat{M}$ *is a* γ*-approximation to* M *if*

$$\begin{aligned} |\hat{P}_a(x,y) - P_a(x,y)| &\leq \gamma \\ |g_a(x) - \hat{g}_a(x)| &\leq \gamma. \end{aligned}$$

Lemma 3 *If* $T \geq \frac{1}{1-\alpha}\log\left(\frac{2g_{\max}}{\epsilon(1-\alpha)}\right)$ *and* $\hat{M}$ *is an* $O\left(\epsilon\left(\frac{1-\alpha}{|\mathcal{S}|g_{\max}}\right)^2\right)$ *approximation of* M*, then,* $\forall u$,

$$||J_{u,M} - J_{u,\hat{M}}||_\infty \leq \epsilon.$$

Sketch of proof: Take a policy u and a start state x. We consider paths of length T starting from x:

$$\mathbf{p} = x_0, x_1, x_2, \ldots, x_T$$

where $\mathbf{p}$ denotes the path. Note that

$$J_{u,M}(x) = \sum_{\mathbf{p}} P_{u,M}(\mathbf{p}) g_u(\mathbf{p}) + \mathrm{E}\left[\sum_{t=T+1}^{\infty} \alpha^t g_u(x_t)\right],$$

where

$$P_{u,M}(\mathbf{p}) = P_{u,M}(x_0, x_1) P_{u,M}(x_1, x_2) \ldots P_{u,M}(x_{T-1}, x_T)$$

is the probability of observing path $\mathbf{p}$ and

$$g_u(\mathbf{p}) = \sum_{t=0}^{T} \alpha^t g_u(x_t)$$

is the discounted cost associated with path $\mathbf{p}$.

By selecting T properly, we can have

$$\left|\mathrm{E}\left[\sum_{t=T+1}^{\infty} \alpha^t g_u(x_t)\right]\right| \leq \frac{\alpha^T g_{\max}}{1-\alpha} \leq \epsilon$$

Recall that $\left|P_a(x,y) - \hat{P}_a(x,y)\right| \leq \gamma$. We consider two kinds of paths:

(a) paths containing at least one transition x_t, x_{t+1} in the set R such that $P_u(x_t, x_{t+1}) \leq \beta$. Note that the total probability associated with such paths is less than or equal to $\beta|\mathcal{S}|T$, since the probability of any given path is less than or equal to β, starting with each state x in each transition there are at most $|\mathcal{S}|$ possible "small probability" transitions, and there are T transitions where this can occur. Therefore

$$\left|\sum_{\mathbf{p}\in R} P_u(\mathbf{p}) g_u(\mathbf{P})\right| \leq \sum_{\mathbf{p}\in R} P_u(\mathbf{p}) \frac{g_{\max}}{1-\alpha} \leq \beta|\mathcal{S}|T\frac{g_{\max}}{1-\alpha}.$$

We can follow the same principle with the MDP $\hat{M}$ to conclude that

$$\left|\sum_{\mathbf{p}\in R} \hat{P}_u(\mathbf{p}) \hat{g}_u(\mathbf{P})\right| \leq \frac{(\beta+\gamma)|\mathcal{S}|T g_{\max}}{1-\alpha}.$$

Therefore, we have

$$\left|\sum_{\mathbf{p}\in R} P_u(\mathbf{p}) g_u(\mathbf{p}) - \sum_{\mathbf{p}\in R} \hat{P}_u(\mathbf{p}) \hat{g}_u(\mathbf{p})\right| \leq \frac{(\gamma+2\beta)\,|\mathcal{S}|\,T g_{\max}}{1-\alpha}$$

(b) For all other paths, we have

$$(1-\Delta) P_a(x_t, x_{t+1}) \leq \hat{P}_a(x_t, x_{t+1}) \leq (1+\Delta) P_a(x_t, x_{t+1})$$

where $\Delta = \frac{\gamma}{\beta}$. Therefore,

$$(1-\Delta)^T P_u(\mathbf{p}) \leq \hat{P}_u(\mathbf{p}) \leq (1+\Delta)^T P_u(\mathbf{p}).$$

Moreover, $|g_u(\mathbf{p}) - \hat{g}_u(\mathbf{p})| \leq T\gamma$, then

$$(1-\Delta)^T \left[J_{u,T} - \gamma T\right] - \frac{\epsilon}{4} \leq \hat{J}_{u,T} \leq (1+\Delta)^T \left[J_{u,T} + \gamma T\right] + \frac{\epsilon}{4}$$

The theorem follows by considering an appropriate choice of β. □

The main point (iv) says that: If exploitation is not possible, then exploration is. We show it by the following lemma.

Lemma 4 *For any $x \in N$, one of the following must hold.*

(a) there exists u in M_N such that $J^N_{u,T}(x) \leq J^*_T(x) + \gamma$, or

(b) there exists u such that the probability that a walk of T steps will terminate in N^C exceeds $\frac{\gamma(1-\alpha)}{g_{\max}}$.

Proof: Let u^* be the policy that attains J^*_T. If

$$J^N_{u^*,T}(x) \leq J^*_T(x) + \gamma$$

then we are done. Suppose that

$$J^N_{u^*,T}(x) > J^*_T(x) + \gamma.$$

Then we have

$$J^N_{u^*,T}(x) = \underbrace{\sum_{q \in N} P^N_{u^*}(q) g^N_u(q)}_{\text{path in } N} + \underbrace{\sum_r P^N_{u^*}(\mathbf{p}) g^N_u(\mathbf{p})}_{\text{path outside } N}$$

and

$$J^*_T(x) = \sum_q P_{u^*}(q) g_u(q) + \sum_r P_{u^*}(q) g_u(q).$$

Therefore

$$J^N_{u^*,T}(x) - J^*_{u^*,T}(x) = \sum_r \left[P^N_{u^*}(\mathbf{p}) \underbrace{g^N_{u^*}(\mathbf{p})}_{\leq \frac{g_{\max}}{1-\alpha}} - \underbrace{P_{u^*}(\mathbf{p}) g_{u^*}(\mathbf{p})}_{\leq 0} \right] > \gamma$$

which implies

$$\sum_r P^N_{u^*}(\mathbf{p}) \frac{g_{\max}}{1-\alpha} > \gamma \Rightarrow \sum_r P^N_{u^*}(\mathbf{p}) \geq \frac{\gamma(1-\alpha)}{g_{\max}}.$$

□

In order the complete the proof of Theorem 1 from the four lemmas above, we have to consider the probabilities from two forms of failure:

- failure to stop the algorithm with a near-optimal policy
- failure to perform enough exploration in a timely fashion

The first point is addressed by Lemmas 1, 2 and 3; which establish that, if the algorithm stops, with high probability the policy produced is near-optimal. The second point follows from Lemma 4, which shows that each attempt to explore is successful with some non negligible probability. By applying the Chernoff bound, it can be shown that, after a number of attempts that is polynomial in the quantities of interest, exploration will occur with high probability.

References

[1] M. Kearns and S. Singh, *Near-Optimal Reinforcement Learning in Polynomial Time*, Machine Learning, Volume 49, Issue 2, pp. 209-232, Nov 2002.