Curriculum Learning for Reinforcement Learning

Runxuan Jiang

1 Deterministic MDP

Proposition (Existence and bound for deterministic finite MDP.)**.** Let $\mathcal{M} = (S, A, H, P, R, \rho)$ be a deterministic MDP with sparse reward, and that $R(s, a) = \mathcal{X}(s = g)$ where for some $g \in A$.

Then there exists a finite sequence of H MDP's $\mathcal{M}_i = (S, A, H, P_i, R_i, \rho_i)$ such that the following algorithm

1. Initialize π_0 to be a random policy.
2. For $1 \leq i \leq H$, Follow π_{i-1} with an ϵ_i-greedy policy until an optimal policy is found.

will find an optimal policy for $\mathcal{M}$ in $O(H^2 k)$ episodes in expectation, where $k = A$.

Proof. Let $s_0, s_1, \ldots, s_H = g$ be a trajectory that the optimal policy for $\mathcal{M}$ takes. Construct $\mathcal{M}_i$ such that $R_i(s, a) = \mathcal{X}(s = s_i)$ and $P_i(s_i|s_i, a) = 1 \ \forall a \in A$, otherwise $P_i = P$.

Then for each $\mathcal{M}_i$, the probability that the ϵ-greedy algorithm will achieve the optimal policy is bounded below by

$$p(\epsilon) = (1 - \epsilon)^{i-1}(\frac{\epsilon}{k}).$$

For $i \geq 2$, p is concave (need a proof of this), so we can find the maximal probability by setting the derivative to 0.

$$\begin{gathered} p'(\epsilon) = 0 \\ \Rightarrow \frac{(1-\epsilon)^{i-1}}{k} = (i-1)(1-\epsilon)^{i-2}(\frac{\epsilon}{k}) \\ \frac{1-\epsilon}{k} = (i-1)(\frac{\epsilon}{k}) \Rightarrow \epsilon = \frac{1}{i}. \end{gathered}$$

So for $\mathcal{M}_i$, p is bounded below by

$$(1 - \frac{1}{i})^{i-1}\frac{1}{ki}.$$

So the expected number of episodes to reach the optimal policy for $\mathcal{M}_i$ is

$$\frac{1}{(1-\frac{1}{i})^{i-1}(\frac{1}{ki})}$$
$$= \frac{ki}{(1-\frac{1}{i})^{i-1}} = (i-1)k(\frac{i}{i-1})^i$$

Notice that since $i \geq 2$ and $x \mapsto (\frac{x}{x-1})^x$ is decreasing for $x > 0$, the above expression is bounded above by

$$4k(i-1).$$

So the number of episodes for the full curriculum is in expectation

$$\sum_{j=2}^{H} 4k(j-1) \leq \frac{H}{2}(4kH)$$

which is $O(kH^2)$. ■

Remark (Tighter bound)**.** Note that the above proposition only considers the case where for each $\mathcal{M}_i$, the ϵ-greedy algorithm is greedy for $i-1$ steps, and then gets the correct state for the single final exploratory step. However, the curriculum learning also helps in the situation where some of the steps are exploratory and "lucky" enough to get the same action as the optimal policy, while the remaining steps are greedy.

However, even using this the bound is still $O(kH^2)$, so this is the tightest bound we can achieve under these conditions (probably).

Proof. Using a similar setup as the above proof, we now consider using the probability for achieving an optimal policy:

$$p(\epsilon) = \binom{i-1}{0}(1-\epsilon)^{i-1}(\frac{\epsilon}{k}) + \binom{i-1}{1}(1-\epsilon)^{i-2}(\frac{\epsilon}{k})^2 + \ldots + \binom{i-1}{i-1}(\frac{\epsilon}{k})^j$$
$$= (\frac{\epsilon}{k})\sum_{j=0}^{i-1}\binom{i-1}{j}(1-\epsilon)^{i-1-j}(\frac{\epsilon}{k})^j$$
$$= (\frac{\epsilon}{k})(1-\epsilon+\frac{\epsilon}{k})^{i-1}.$$

Note that this is concave on $(0,1)$ (proof needed).

Since it is concave, we can set derivative to 0 to find the ϵ that maximizes the probability.

We get that

$$\epsilon = \frac{k}{i(k-1)}$$

We then plug back into p and take the reciporocal to get the expected number of episodes for $\mathcal{M}_i$, which is

$$\frac{i^2(k-1)}{i-1}$$

.

Note that $x \leq \frac{x^2}{x-1} \leq 4x$

so this also leads to a bound of $O(kH^2)$ if we sum across all $1 \leq i \leq H$.

To prove this is a tight bound, we can show that the probability given by $p(\epsilon)$ is a tight lower bound to the actual probability. ■

Proposition (Existence and bound for deterministic finite MDP)**.** Let $\mathcal{M} = (S, A, H, P, R, \rho)$ be a deterministic MDP with sparse reward, and that $R(s, a) = \mathcal{X}(s = g)$ for some $g \in A$.

Then there exists a finite sequence of H MDP's $\mathcal{M}_i = (S, A, H, P_i, R_i, \rho_i)$ such that the following algorithm

1. Initialize π_0 to be a random policy.
2. For $1 \leq i \leq H$, Follow π_{i-1} with an ϵ_i-greedy policy until an optimal policy is found.

will find an optimal policy with probability $1 - \delta$ in at least x episodes.

Proof. Let $s_0, s_1, \ldots, s_H = g$ be a trajectory that the optimal policy for $\mathcal{M}$ takes. Construct $\mathcal{M}_i$ such that $R_i(s, a) = \mathcal{X}(s = s_i)$ and $P_i(s_i|s_i, a) = 1 \; \forall a \in A$, otherwise $P_i = P$.

Then for each $\mathcal{M}_i$, the probability that the ϵ-greedy algorithm will achieve the optimal policy is

$$\begin{aligned} p(\epsilon) &= \binom{i-1}{0}(1-\epsilon)^{i-1}(\frac{\epsilon}{k}) + \binom{i-1}{1}(1-\epsilon)^{i-2}(\frac{\epsilon}{k})^2 + \ldots + \binom{i-1}{i-1}(\frac{\epsilon}{k})^j \\ &= (\frac{\epsilon}{k})\sum_{j=0}^{i-1}\binom{i-1}{j}(1-\epsilon)^{i-1-j}(\frac{\epsilon}{k})^j \\ &= (\frac{\epsilon}{k})(1-\epsilon+\frac{\epsilon}{k})^{i-1}. \end{aligned}$$

For $i \geq 2$, p is concave (need a proof of this), so we can find the maximal probability by setting the derivative to 0.

$$p'(\epsilon^*) = 0$$

$$\Rightarrow \frac{(1-\epsilon^* + \frac{\epsilon^*}{k})^{i-1}}{k} + (\frac{\epsilon^*}{k})(i-1)(1-\epsilon^* + \frac{\epsilon^*}{k})^{i-2}(\frac{1}{k}-1) = 0$$

$$\Rightarrow \epsilon^* = \frac{k}{i(k-1)}$$

So for $\mathcal{M}_i$, the probability of reaching the optimal policy in each episode is

$$p_i(\epsilon^*) = (\frac{\frac{k}{i(k-1)}}{k})(1 - \frac{k}{i(k-1)} + \frac{\frac{k}{i(k-1)}}{k})^{i-1}$$

$$= \frac{i-1}{i^2(k-1)}$$

■

2 Curriculum Learning with Model-Based algorithms

Given a model-based reinforcement learning algorithm (such as RMAX or UCB), we want to show that we can solve MDP's using the curriculum learning paradigm on the MDP's used by the algorithm to "model" the MDP to be solved. Formally, such algorithms generate a sequence of MDP's $M_1, \ldots, M_n$ with corresponding optimal policies $\pi_1^*, \ldots, \pi_n^*$. We want to show that by using π_{i-1}^* with ϵ-greedy on M_i we can achieve optimal regret bounds for M_i. Through induction, we therefore have a "curriculum" for learning the MDP.

2.1 RMAX

We start with an episodic nonstationery MDP $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, \rho)$. Let L be a collection of states and actions $L \subset \mathcal{S} \times \mathcal{A}$ corresponding to the set of states marked unknown in the RMAX algorithm.

Definition (M_L)**.** Let M and L be defined as above. We define M_L as a MDP with the same states and actions as M, and with the same rewards and transitions for $(s,a) \notin L$. For $(s,a) \in L$, the the reward will be equal to R_{max} and the transition will lead to an absorbing state that always gives reward R_{max}.

The RMAX algorithm works by starting with $L_0 = \mathcal{S} \times \mathcal{A}$, and using the optimal policy for M_{L_0} on M. Once enough samples have been collected from a state-action pair (s,a), it is removed from L_0 to get $L_1 = L_0 \setminus \{(s,a)\}$, and then using the optimal policy for M_{L_1}. This is done until there are no unknown states left (i.e., L is empty).

Note that

$$\mathcal{S} \times \mathcal{A} = L_0 \supset L_1 \supset \ldots \supset L_{SA} = \emptyset$$

Denote M_{L_j} as M_j. The following describes the main theorem.

Theorem. Fix $0 < \delta < 1, \epsilon > 0$. Then we can find an ϵ-optimal policy with probability higher than $1 - \delta$ by performing the following steps on the sequence of MDP's $M_1, \ldots, M_{SA}$ with corresponding optimal policies π_1^*, π_{SA}^*:

- Initialize π_0 to be a random policy.
- For $t = 1, \ldots, n$, follow π_{t-1} with ϵ-greedy exploration to collect K_2 trajectories. Compute the optimal policy π_t from the model learned by the trajectories.
- Output π_{SA}.

Lemma. Let $\alpha > 0$ and $1 \leq j \leq SA$. Denote (s, a) to be the entry in L_{j-1} but not in L_j. Consider playing the policy π^*_{j-1} with ϵ-greedy on M_j. Then either

1. $V_{M_j}^{\pi^*_{j-1}} > V_{M_j}^{\pi^*_j} - \alpha$
2. (s, a) will be played with probability of at least $\frac{\alpha}{H*R_{max}}$.

Proof. Let T be the set of possible trajectories in M_j. Then

$$|V_{M_{j-1}}^{\pi^*_{j-1}} - V_{M_j}^{\pi^*_j}| \leq |\sum_{t \in T} Pr_{M_{j-1},\pi_{j-1}}[t] V_{M_{j-1}}(t) - \sum_{t \in T} Pr_{M_j,\pi_j}[t] V_{M_j}(t)|$$

Let T' be the set of trajectories that passes through (s, a). Then the above expression can be reduced to

$$|\sum_{t \in T'} Pr_{M_{j-1},\pi_{j-1}}[t] V_{M_{j-1}}(t) - \sum_{t \in T'} Pr_{M_j,\pi_j}[t] V_{M_j}(t)|$$

$$\leq H * R_{max} \sum_{t \in T'} Pr_{M_{j-1},\pi_{j-1}}[t] - Pr_{M_j,\pi_j}[t] \leq H * R_{max} * \sum_{t \in T'} Pr_{M_j,\pi_{j-1}}[t]$$

Note that $\sum_{t \in T'} Pr_{M_j,\pi_{j-1}}[t]$ is the probability of π_{j-1} of reaching (s, a). So if it is higher than $\frac{\alpha}{H*R_{max}}$ then (1) holds and we are done. Otherwise, we have that $|V_{M_{j-1}}^{\pi^*_{j-1}} - V_{M_j}^{\pi^*_j}| \leq \alpha$ which means that (2) holds. ■

Lemma. Fix $0 < \delta < 1$. Then if we obtain $K_1 = max((\frac{4SHR_max}{\epsilon})^3, -6\ln^3(\frac{\delta}{6SA^2})) + 1$ samples for (s, a) then the estimate for the transition probability for (s, a) is within $\frac{\epsilon}{2*SHR_{max}}$ of the actual probabilities.

Proof. Use Chernoff bound and pigeonhole principle. ■

Lemma. Fix $0 < \delta < 1$ and $\epsilon > 0$. Then we can find an ϵ-optimal policy with probability higher than $1 - \delta$ for MDP M_j after collecting trajectories by running policy π_{j-1} with ϵ-greedy exploration for K_2 episodes where $K_2^{\frac{1}{3}} + K_2 \frac{\alpha}{H*R_{max}} > A^2 S K_1$ and computing the optimal policy for that model.

Note. We may require less than this number since we only have a single unknown state.

Proof. By the previous two lemmas, this number of steps ensures that K_1 samples are obtained for (s, a) with high (greater than $1 - \delta$) probability. Thus, we can accurately estimate the MDP M_j with a TV distance of less than $\frac{\epsilon}{2*SHR_{max}}$. By lemma 4 in the R-max paper, this means that the optimal value function of the estimated MDP and M_j are ϵ close to each other. So calculate the optimal policy for the estimated MDP to get the resulting policy. ■

Note. In this case we don't really use ϵ-greedy (we set ϵ to zero).

Proof (of Theorem). Follows from previous lemma. ■

2.2 UCRL