Curriculum Learning for Reinforcement Learning

Runxuan Jiang

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.

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

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.

$$\begin{aligned} p'(\epsilon^*) &= 0 \\ \Rightarrow \frac{(1-\epsilon^*+\frac{\epsilon^*}{k})^{i-1}}{k} &+ (\frac{\epsilon^*}{k})(i-1)(1-\epsilon^*+\frac{\epsilon^*}{k})^{i-2}(\frac{1}{k}-1) = 0 \end{aligned}$$

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

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

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

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

■