Curriculum Learning for Reinforcement Learning

Runxuan Jiang

1 Deterministic MDP

We first consider the case of using curriculum learning on an MDP with deterministic reward and transitions.

Proposition (Existence and bound for deterministic finite MDP.)**.** Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, R, \rho)$ 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 $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, H, P_i, R_i, \rho_i)$ 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 $k = |\mathcal{A}|$.

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 $P_i = P$ for $1 \leq i \leq H$.

Let $p_i(\epsilon)$ denote the probability that the ϵ-greedy algorithm will achieve the optimal policy for MDP $\mathcal{M}_i$. Then we have

$$p_i(\epsilon) \geq f(\epsilon) := (1 - \epsilon)^{i-1}(\frac{\epsilon}{k}).$$

This comes from the fact that the optimal trajectory can be achieved by playing the greedy option $i - 1$ times, and then choosing the exploratory option and choosing the action that reaches the correct state, which has probability $\frac{\epsilon}{k}$.

For $i \geq 2$, f is concave, so we can find the maximal probability by setting the derivative to 0.

Note. Will need a prove on why this is concave.

$$p_i'(\epsilon) = 0$$

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

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

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

By plugging in this value for ϵ in f 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:

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

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

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

Note. Proof of concavity needed.

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

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

We then plug this value of ϵ back into f to get the maximal point of f:

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

and take the reciporocal to get the expected number of episodes for $\mathcal{M}_i$, which is

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

.

Note that $x \leq \frac{x^2}{x-1} \leq 4x$, so since $2 \leq i \leq H$ we get that

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

which is $O(kH^2)$.

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.

Note. TODO: Prove this theorem (same as above but finding the number of episodes to get an optimal policy with high probability).

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 $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, \rho)$. The R-max algorithm works by encouraging exploration on "unknown" states by setting the reward for those states to a maximal value R_{max}.

2.1.1 Original R-Max Algorithm

We first describe the RMAX algorithm for MDP's:

Algorithm 1 R-max for MDP's

```
Require: MDP $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, \rho), \epsilon > 0, 0 \leq \delta \leq 1, R_{max} > 0$.
procedure R-MAX($M, \epsilon, \delta, R_{max}, K_1, K_2$)
    $Rewards[S][A][H]$ be a new array.
    $Transitions[S][A][H]$ be a new array.
    $Known[S][A][H]$ be a new array.
    $\hat{M} \leftarrow M$ with all states replaced by absorbing states with reward $R_{max}$ for any action.
    $Done \leftarrow False$
    while not $Done$ do
        $Done \leftarrow True$
        $\pi \leftarrow$ optimal policy for $\hat{M}$
        while $\pi$ has not been run for more than $K_1$ steps do
            obtain trajectories by running $\pi$ on $M$.
            if a state $(s, a, h)$ has been visited more than $K_2$ times and $Known[s][a][h]$ is false then
                $Known[s][a][h] \leftarrow True$
                set the estimated transition probabilities $Transitions[s][a][h]$ and rewards $Rewards[s][a][h]$ based on the average of the observed transitions and rewards for $(s, a, h)$
                replace the transition probabilities and reward in $(s, a, h)$ in $\hat{M}$ by the estimated values $Transitions[s][a][h]$ and $Rewards[s][a][h]$
                $Done \leftarrow False$
                break
            end if
        end while
    end while
    return The optimal policy for $\hat{M}$
end procedure
```

Note. In the original R-max algorithm, the optimal policy for each M_j is not used since M_j is not know fully. Instead, the optimal policy for the estimated MDP for M_j, which is based on the samples collected previously, is used.

Remark (Is this curriculum learning?)**.** Notice that this algorithm appears similar to curriculum learning. However there are some differences. In curriculum learning, the algorithm is given a sequence of tasks and trains an agent on the current task by using a (near)-optimal policy for the previous task. So given a sequence of MDP's $M_1, \ldots, M_n$ you explore M_j with the aid of the policy π_{j-1}. However, in the R-max algorithm we are exploring M directly using π_j.

2.1.2 R-max Algorithm with Curriculum Learning

We first introduce some definitions. Define $\mathcal{L} = \mathcal{P}(\mathcal{S} \times [H])$ to be the set of all possible sets of (nonstationery) states.

Definition. Let M be a nonstationery finite horizon MDP as defined above and let $L \in \mathcal{L}$.

Define the MDP $\underline{M_L}$ to be the MDP with the same states and actions as M, and with the same rewards for all actions for all states $(s, h) \notin L$. For $(s, h) \in L$, the state is an absorbing state and the reward is set to R_{max} for all actions.

We now introduce the curriculum learning algorithm utilizing the exploration principles of the R-max algorithm. For simplicity, we assume that the starting state for M, s_0 is fixed: $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, s_0)$. We also assume that we have access to the MDP M_L for all $L \in \mathcal{L}$.

Algorithm 2 Curriculum Learning with R-max

```
Require: MDP $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, s_0)$, $\epsilon > 0$, $\delta \leq 1$, $R_{max} > 0$ and $\mathcal{M} = \{M_L : L \in \mathcal{L}\}$.
 1: Initialize $L_0 = (\mathcal{S} \times [H]) \setminus \{(s_0, 1)\}$; $M_0 = M_{L_0}$; $\Phi_0 = \{\bar{\pi}_0\}$ for $\bar{\pi}_0$ being uniformly random
    policy; $(\bar{s}_0, \bar{h}_0) = (s_0, 1)$.
 2: for $t = 1, \ldots, S \times H$ do
 3:     Let $\pi_t$ = mixture of $\Phi_{t-1}$ with $\epsilon$-greedy exploration.
 4:     $\mathcal{H}_t = \{\}$.
 5:     for $i = 1, \ldots, K_1$ do
 6:         Run $\pi_t$ on $M_{t-1}$ and add the trajectory to $\mathcal{H}_t$.
 7:     end for
 8:     Call offline policy optimization on $\mathcal{H}_t$ that returns $\bar{\pi}_t$, a near-optimal policy for $M_{t-1}$.
 9:     if # $\bar{\pi}_t$ will visit $L_{t-1}$ with probability at least $p$. then
10:         Set $\bar{s}_t, \bar{h}_t$ to be the state in $L$ with the highest probability of being reached.
11:     else
12:         break
13:     end if
14:     $\Phi_t = \Phi_{t-1} \cup \{\bar{\pi}_{t-1}\}$.
15:     $L_t = L_{t-1}/(\bar{s}_t, \bar{h}_t)$; $M_t = M_{L_t}$.
16: end for
17: Output $\bar{\pi}_t$.
```

We now introduce the main theoretical guarantee for the above algorithm.

Theorem. Given a nonstationery finite horizon MDP M with fixed initial state, an $\epsilon > 0$, a $0 \leq \delta \leq 1$, the maximum reward for M, R_{max} and the $\mathcal{M}$ corresponding to M, then running algorithm 2 with $K_1 = \_$, $K_2 = \_$, the algorithm will output an ϵ-optimal policy with probability $1 - \delta$. Furthermore, the sample complexity of the algorithm is _.

We start off with some preliminary results about estimating and MDP, which will lead to a result justifying K_1 in line 5.

Theorem. Let M be a MDP and let M' be another MDP such that $||P_M - P_{M'}||_\infty < \frac{\epsilon}{SH^2 R_{max}}$. Then for all policies π,

$$|V_M(\pi) - V_{M'}(\pi)| \leq \epsilon$$

Proof. Lemma 4 of the R-max paper. ■

Lemma. Let M be an MDP. To approximate the transition probabilility for a state s, a, h in M with error less than ϵ with probability higher than $1-\delta$, we need $n \geq \frac{-\ln(\delta/2S)}{2\epsilon^2}$ samples.

Proof. Let n be the number of samples collected where the state (s, a, h) leads to the state s'. Let X_i be an indicator variable that is 1 if (s, a, h) leads to $(s', h+1)$ and 0 otherwise in ith sample. Let p be the actual transition probability of transitioning from (s, a, h) to $(s', h+1)$. Let $S = \frac{X_1 + \ldots + X_n}{n}$. Note that $0 \leq \frac{X_i}{n} \leq \frac{1}{n}$. Since there are S possibilities for (s', h), we want the probability that $|S - p| \geq \epsilon$ to be less than $\frac{\delta}{S}$. By Hoeffding's inequality we have

$$P(|S-p| \geq \epsilon) \leq 2\exp(\frac{-2\epsilon^2}{n(\frac{1}{n})^2}) = 2\exp(-2n\epsilon^2).$$

We want

$$2\exp(-2n\epsilon^2) \leq \delta/S$$

$$\Rightarrow -2n\epsilon^2 \leq \ln(\frac{\delta}{2S})$$

$$\Rightarrow n \geq \frac{-\ln(\frac{\delta}{2S})}{2\epsilon^2}.$$

■

Lemma (Justification for K_1 (loose bound))**.** Fix $0 \leq \delta \leq 1$. In order to collect n samples for each state (s, h) with probability $1-\delta$ or more, we need to set $K_1 = nSAH\frac{\ln(1-\delta^{\frac{1}{nSAH}})}{\ln(\frac{p\epsilon(1-\epsilon)^H}{SAH^2})}$.

Proof. The number of trajectories needed to be collected is upper bounded by the case where L is empty, so we consider this case only in this proof. We first calculate the probability of the policy π_t from line 6 for reaching any particular state in $(\mathcal{S} \times [H]) \setminus L_t$ and choosing some action in $\mathcal{A}$. We denote the state to be (s, h) and the action as a. In order to reach (s, h) we must first choose the policy in Φ which reaches (s, h) with guaranteed probability. From line 9 and pigeonhole principle, there exists such a policy that reaches (s, h) with probability at least $\frac{p}{SH}$. Furthermore we have $|\Phi| \leq SH$. So the probability of reaching (s, h) is at least $\frac{p(1-\epsilon)^H}{SH^2}$. Next, the probability of reaching a particular action from (s, h) is $\frac{\epsilon}{A}$. Thus, the probability of reaching a specific state and performing a specific action on that state is

$$\frac{\epsilon(1-\epsilon)^H}{SAH^2}.$$

We want to visit each state and action n times. In this case this is equivalent to visiting (s, h) and performing action a $nSAH$ times. Let k be the proposed number of trajectories to collect. Let $X_i, 1 \leq i \leq k$ be a Bernoulli random variable that corresponds to a trajectory visiting (s, h) and performing action a on it. Then we have

$$P(\sum_{i=1}^{k} X_i \geq nSAH) \geq \prod_{j=1}^{nSAH} P(\sum_{i=1}^{\frac{k}{nSAH}} X_{j*i} \geq 1)$$

$$= \prod_{j=1}^{nSAH} (1 - P(\sum_{i=1}^{\frac{k}{nSAH}} X_{j*i} = 0)$$

$$= (1 - (\frac{p\epsilon(1-\epsilon)^H}{SAH^2})^{\frac{k}{nSAH}})^{nSAH} \geq \delta$$

$$\Rightarrow 1 - (\frac{p\epsilon(1-\epsilon)^H}{SAH^2})^{\frac{k}{nSAH}} \geq \delta^{\frac{1}{nSAH}}$$

$$\Rightarrow 1 - \delta^{\frac{1}{nSAH}} \geq (\frac{p\epsilon(1-\epsilon)^H}{SAH^2})^{\frac{k}{nSAH}}$$

$$\Rightarrow \ln(1 - \delta^{\frac{1}{nSAH}}) \geq \frac{k}{nSAH} \ln(\frac{p\epsilon(1-\epsilon)^H}{SAH^2})$$

$$\Rightarrow k \geq nSAH \frac{\ln(1 - \delta^{\frac{1}{nSAH}})}{\ln(\frac{p\epsilon(1-\epsilon)^H}{SAH^2})}$$

■

Proof (tighter bound using induction). ■

Next we prove a lemma showing that if the policy in line 9 is not already an optimal policy for the MDP M, then it must have a guaranteed probability of exploring L.

Lemma. Let $\alpha > 0$ let $L \subset \mathcal{S} \times [H]$. Let π_L be an optimal policy for M_L. Let $V_M(\pi)$ denote the value function on MDP M of policy π, and let V_M^* be the value of any optimal policy on M. Then either

1. $V_M(\pi_L) > V_M^* - \alpha$
2. A state in L will be played with probability of at least $\frac{\alpha}{H*R_{max}}$ when playing π_L on M (or M_L).

Proof. This proof follows the proof of lemma 6 in the original R-max paper.

Suffices to show that if (2) does not hold then (1) holds. Suppose that (2) does not hold. Let T be the set of all possible trajectories on M and let T' be the set of all trajectories that pass through L. Let $P_M^\pi(t)$ denotes the probability that the trajectory t will be reached when playing policy π on the MDP M. Since (2) does not hold we have that

$$\sum_{t \in T'} P_M^{\pi_L}(t) < \frac{\alpha}{HR_{max}}.$$

We want to show that

$$V_M^* - V_M(\pi_L) < \alpha.$$

Note that the only differences between M and M_L is that in M_L, the states in L are absorbing states and any action from this state gives a reward of R_{max}, but in M the reward is replaced by the actual reward and the transition probabilities are also replaced by the actual transition probabilities from M. Thus we have

$$V_M^* \leq V_{M_L}^*.$$

$$\Rightarrow V_M^* - V_M(\pi_L) \leq V_{M_L}^* - V_M(\pi_L)$$

We now further decompose each of these value functions over all possible trajectories. Let $V_M(t)$ denote the total reward of going through trajectory t on MDP M. Then we have

$$|V_M(\pi_L) - V_{M_L}^*| = |\sum_{t \in T} P_M^{\pi_L}(t) V_M(t) - \sum_{t \in T} P_{M_L}^{\pi_L}(t) V_{M_L}(t)|$$

$$= |\sum_{t \in T'} P_M^{\pi_L}(t) V_M(t) + \sum_{t \in T \setminus T'} P_M^{\pi_L}(t) V_M(t) - \sum_{t \in T'} P_{M_L}^{\pi_L}(t) V_{M_L}(t) - \sum_{t \in T \setminus T'} P_{M_L}^{\pi_L}(t) V_{M_L}(t)|$$

$$\leq |\sum_{t \in T'} P_M^{\pi_L}(t) V_M(t) - \sum_{t \in T'} P_{M_L}^{\pi_L}(t) V_{M_L}(t)| + |\sum_{t \in T \setminus T'} P_M^{\pi_L}(t) V_M(t) - \sum_{t \in T \setminus T'} P_{M_L}^{\pi_L}(t) V_{M_L}(t)|$$

Since the rewards and transition probabilities for $t \in T \setminus T'$ are the same for MDPs M and M_L, the second expression above is equal to 0. Thus we are left with the first expression

$$= |\sum_{t \in T'} P_M^{\pi_L}(t) V_M(t) - \sum_{t \in T'} P_{M_L}^{\pi_L}(t) V_{M_L}(t)|.$$

Finally note that since the probabilities for all states before reaching something in L have the same transition probability for M and M_L, we have that $\sum_{t \in T'} P_M^{\pi_L}(t) = \sum_{t \in T'} P_{M_L}^{\pi_L}(t)$. We also have that $\forall t, 0 \leq V_M(t) \leq HR_{max}$ and $0 \leq V_{M_L}(t) \leq HR_{max}$. Thus we have

$$|\sum_{t \in T'} P_M^{\pi_L}(t) V_M(t) - \sum_{t \in T'} P_{M_L}^{\pi_L}(t) V_{M_L}(t)| \leq \sum_{t \in T'} P_M^{\pi_L} HR_{max}.$$

However from our assumption we have that

$$\sum_{t \in T'} P_M^{\pi_L}(t) < \frac{\alpha}{HR_{max}}.$$

Substituting this into the previous expression we have that

$$V_M^* - V_M(\pi_L) \leq V_{M_L}^* - V_M(\pi_L) < \alpha$$

Which is the same statement as (1). ■

Proof (of main theorem). By the above lemma, for any $L \subset \mathcal{S} \times [H]$ we either have that the optimal policy for M_L is ϵ-optimal for M or a state in L will be played with probability of at least $\frac{\epsilon}{HR_{max}}$. So we set $p = \frac{\epsilon}{HR_{max}}$.

Next, let M'_L be our estimated MDP for M_L. We need $||P_{M_L} - P_{M'_L}||_\infty < \frac{\epsilon}{SH^2 R_{max}}$ which means that for each state we need (here we use σ-greedy)

$$n \geq \frac{-\ln(\frac{\delta}{2S})}{2(\frac{\epsilon}{SH^2 R_{max}})^2}$$

samples. In order to satisfy this amount for every state in M_L that is not in L, and for every action played on each state, we would need

$$K_1 = nSAH \frac{\ln(1 - \delta^{\frac{1}{nSAH}})}{\ln(\frac{p\sigma(1-\sigma)^H}{SAH^2})} = \frac{-\ln(\frac{\delta}{2S})}{2(\frac{\epsilon}{SH^2 R_{max}})^2} SAH \frac{\ln(1 - \delta^{\frac{1}{\frac{-\ln(\frac{\delta}{2S})}{2(\frac{\epsilon}{SH^2 R_{max}})^2} SAH}})}{\ln(\frac{\frac{\epsilon}{HR_{max}}\sigma(1-\sigma)^H}{SAH^2})}$$

total samples. Since we need this number of samples for each iteration, the total sample complexity of the algorithm is

$$\frac{-\ln(\frac{\delta}{2S})}{2(\frac{\epsilon}{SH^2 R_{max}})^2} S^2 AH^2 \frac{\ln(1 - \delta^{\frac{1}{\frac{-\ln(\frac{\delta}{2S})}{2(\frac{\epsilon}{SH^2 R_{max}})^2} SAH}})}{\ln(\frac{\frac{\epsilon}{HR_{max}}\sigma(1-\sigma)^H}{SAH^2})}$$

The algorithm is optimal because the resulting M_L would have a value function ϵ-close to M's value function, and the estimated M'_L would have a value function ϵ-close to M_L's value function, all with probability $1 - \delta$. ■