R07/Sarthak Mittal/200050129

February 11, 2023

Artificial intelligence often requires formulating decision-making problems using **Markov Decision Processes (MDPs)**. Several algorithms are available for **planning** in MDPs, with upper bounds that are polynomial in the size of MDP representation and the discount factor γ. This paper provides the first set of non-trivial **upper bounds**, that are **independent** of parameter representation and γ, for the number of iterations of convergence of **policy iteration (PI)** in the worst case. They also introduce a **Randomized PI** that accepts a single state-state action improvement with probability **0.5**.

The MDP framework used is **infinite-horizon** with a **discounted** sum of rewards. They explain how general policy iteration works by changing action at each state and then deciding which changes to accept to improve the policy (better in terms of **partial ordering** over the policies using the value function). Using a **modification set** T that contains the (s, a) pairs that could improve the policy, they define an iteration in terms of two steps: (1) Improvement Selection (choose a subset U out of the modification set) (2) Policy Improvement (modify the policy according to the subset). The modification set is **"well-defined"** if each state appears only once. They also utilize the crucial fact that each iteration **strictly** improves the policy, thus **skipping** all policies between the current and new policies in partial ordering.

Their claims show that any subset of states appears **at most once** in the case of two-action MDP. The analysis of the **"modify"** operation (for policy improvement) shows that if a policy has a well-defined modification set, then the number of policies skipped by PI would be **at least** the size of that set. Using this analysis, they show that the **Greedy PI** (the subset selection always selects the whole set) considers at most $\mathcal{O}(2^n/n)$ policies for a two-action MDP by splitting the proof into cases on size of modification set.

The Random PI defines the **"select"** operation as choosing a random subset **uniformly** at random, hence accepting each local improvement with probability **0.5**. With the use of a set of properties related to partial orders and policy comparison in PI, they prove that the expected number of policies that are **skipped** at each iteration i is at least $2^{|T^{\pi_i}|-1}$. Considering the random behavior, they derive a **probabilistic bound** using binary entropy and a notion of "good" iteration and "typical" run, which says that **Random PI** considers at most $\mathcal{O}(2^{0.78n})$ different policies with probability $1 - 2^{-2^{\Omega(n)}}$.

To extend results to multi-action MDPs (action space of size k), they assume that there is a way to **reduce** a modification set T that is not well-defined to a subset L that is. Another important **fact** is that if there are two policies π_i and π_j in PI (where $i < j$) such that $U_i \subseteq U_j$ then the policies **cannot** be the same for all $s \in U_i$. Due to this, the number of iterations where $|L^{\pi_i}| \leq d$ is **bounded** by $\sum_{j=0}^{d} \binom{n}{j} k^j$.

These results lead to a bound of at most $\mathcal{O}(k^n/n)$ different policies for **Greedy PI**. For **Random PI**, the expected number of policies **skipped** becomes at least $2^{|L^{\pi_i}|-1}$, which gives a bound of at most $\mathcal{O}\left(\left(\left(1+\frac{2}{logk}\right)\frac{k}{2}\right)^n\right)$ different policies with probability $1-2^{-\Omega((k/2)^n)}$.

In concluding remarks, they compare their obtained upper bounds with the $\Omega(2^n)$ **lower bound** proven for **Sequential PI**. The few places where their analysis is lacking are: (1) consideration of **additional policies** to rule out, apart from the modification set (2) taking **advantage** of the modification set size (3) the **large gap** between the proven upper bound and the trivially known lower bound.