Topics in machine learning: Naresh Manwani

Siddharth Bhat

Monsoon 2019

Contents

1 Policy iteration **5**
1.1 Value iteration algorithm 5

2 Monte carlo methods for MDP **7**
2.1 Naive 7

Chapter 1

Policy iteration

$$\pi_{k+1}(s) = \arg\max_{a \in A(s)} r(s,a) + \gamma \sum_{s} P(s'|s,a) v_{\pi_k}(s')$$

Theorem 1 *The policy iteration algorithm generates a sequecnce of policies with non-decreasing state values. That is,* $V^{\pi_{k+1}} \geqslant V^{\pi_k}$*,* $V^{\pi} \in \mathbb{R}^n$*, is the vector of state values for state* π

Proof 1 F^{π_k} *is the bellman expectation operator (?)*
Since V^{π_k} *is a fixed point of* F^{π_k}*,*

$$\begin{aligned}
V^{\pi_k} = F^{\pi_k}(V^{\pi_k}) \leqslant F(V^{\pi_k}) \qquad &\textit{(upper bounded by max value)} \\
F(V^{\pi_k}) = F^{\pi_{k+1}}(V^{\pi_k}) \qquad &\textit{(By defn of policy improvement step)} \\
V^{\pi_k} \leqslant F^{\pi_{k+1}}(V^{\pi_k}) \qquad &\textit{(eqn 1)} \\
F^{\pi_{k+1}}(V^{\pi_k}) \leqslant (F^{\pi_{k+1}})^2(V^{\pi_k}) \qquad &\textit{(Monotonicity of } F^{\pi_{k+1}}\textit{)} \\
\forall t \geqslant 1,\; F^{\pi_{k+1}}(V^{\pi_k}) \leqslant (F^{\pi_{k+1}})^t(V^{\pi_k}) \qquad &\textit{(Monotonicity of } F^{\pi_{k+1}}\textit{)} \\
F^{\pi_{k+1}}(V^{\pi_k}) \leqslant (F^{\pi_{k+1}})^t(V^{\pi_k}) \leqslant V^{\pi_{k+1}} \qquad &\textit{(Contraction mapping, } V^{\pi_{k+1}} \textit{ is fixed point)} \\
V^{\pi_k} = F^{\pi_{k+1}}(V^{\pi_k}) \leqslant V^{\pi_{k+1}} &
\end{aligned}$$

For a set of actions $\mathcal{A}$ and a set of states $\mathcal{S}$, the total number of policies is $|\mathcal{A}^{\mathcal{S}}|$. The number of computations per iteration is $O(|\mathcal{S}|^3)$. So the loose upper bound is be $O(|\mathcal{S}|^3 \times |\mathcal{A}^{\mathcal{S}}|)$.

1.1 Value iteration algorithm

```
let v n s = max [r s a + gamma * sum [(p s' s a) * v (n-1) s' | s' <- ss] | a <- as]
let vs = [v i | i <- [0..]]
-- | L infinity
let norm v v' = max [(v s - v' s) | s <- ss]
let out = head $
  dropWhile (\v v' -> norm (v' - v) < eps * (1 - gamma) / (2 * gamma)) $
  zip vs (tail vs)
let policy s = argmax as $ \a ->
  r s a + gamma * sum [ (p s' s a) * out s' | s' <- ss]
```

Theorem 2 *For the series* V_n *and the policy* π_ϵ *computed by the value iteration algorithm, then:*

$$\forall \epsilon > 0,\ \exists n_0 \in \mathbb{N}, \forall n \geqslant n_0, \quad \|V_{n+1} - V_n\|_\infty \leqslant \frac{\epsilon(1-\gamma)}{2\gamma}$$

Proof 2 *We need to show that the sequence* $\{V_n\}_{n=0}^{\infty}$ *is a Cauchy sequence. This has ben proven before by the use of contraction mapping. Thus, for a given* $\epsilon' \geqslant 0, \exists n_0 \in \mathbb{N}, \forall n \geqslant n_0, \|V_{n+1} - V_n\|_\infty \leqslant \epsilon'$ *by cauchy sequence. So, pick* $\epsilon' = \frac{\epsilon(1-\gamma)}{2\gamma}$*, and the proof immediately follows.*

Theorem 3 *If* $\|V_{n+1} - V_n\|_\infty \leqslant \frac{\epsilon(1-\gamma)}{2\gamma}$*, then* $\|V_{n+1} - V^\star\|_\infty < \epsilon/2$

Proof 3

$$\begin{aligned}
&\|V_{n+1} - V^\star\| = \|V_{n+1} - FV_{n+1} + FV_{n+1} - V^\star\| \leqslant \|V_{n+1} - FV^\star\| + \|FV_n - V_n\| \qquad \textit{(triangle inequality)}\\
&\leqslant \|V_{n+1} - FV^\star\| + \gamma\|V_{n+1} - V^\star\|\\
&\leqslant \gamma\|V_{n+1} - V_n\| + \gamma\|V_{n+1} - V^\star\|\\
&(1-\gamma)\|V_{n+1} - V^\star\| \leqslant \gamma\|V_n - V_{n+1}\| \qquad \textit{(how?)}\\
&\implies \ldots
\end{aligned}$$

It appears that V^{π_ϵ} is just V_{n+2}??

Theorem 4 *The policy* π_ϵ *is* ϵ*-optimal:* $\|V^\star - V^{\pi_\epsilon}\| \leqslant \epsilon$

Chapter 2

Monte carlo methods for MDP

For dynamic programming, we needed to know the transition probability distribution $P(s, a, s')$, nor the reward function $r(s, a)$.

In the monte carlo methods, we assume that we do not know the transition probability distribution. We rely only on simulations.

This samples over *episodes* for a fixed policy: sequences of states, actions, and rewards.

2.1 Naive

- For each $s \in S$, run π from s for m times, where the ith episode is T_i.
- Let r_i be the return of T_i
- Estimate the value of π starting from s as $\hat{v}_\pi(s) = \frac{1}{m} \sum_{i=1}^{m} r_i$.
- Show by chernoff bounds that this is an OK estimate. We can use Chernoff as $\{r_i\}$ are independent, since the $\{T_i\}$ are independent.