Topics in machine learning: Naresh Manwani

Siddharth Bhat

Monsoon 2019

Contents

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},

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

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..]]
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]
```