Understanding Monotonicity of the Bellman Operator and the Implications Thereof

Ron Parr
CPS 570

The Bellman Operator

\[ V^{i+1}(s) = \max_a R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^i(s') \]

- Let’s simplify things call this \( T \)
- \( V^{i+1} = TV^i \)
- We say \( V \) is **pessimistic** if \( V^{i+1} \geq TV^i \)
- Call \( k \) applications of \( T = T^k \)

Observations about Pessimistic \( V \)

- Suppose \( V \) is pessimistic
  - \( T^k V \) is also pessimistic for all \( k \)
  - \( T^{k+1} V \geq T^k V \) for all \( k \)
- Why (sketch): Consider state \( s \), if the states reachable by \( s \) have increased in value at iteration \( i \), then state \( s \) cannot decrease in value at iteration \( i+1 \)

Bellman Operator for a Specific Policy

\[ V^{i+1}_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s)) V^i_\pi(s') \]

- Call this \( T_\pi \)
- This is also monotonic under the same assumptions as for \( T \)
Modified Policy Iteration

- Guess $V^0$
- $\pi^0 = \text{greedy}(V^0)$
- $i=1$
- $V^i = T^i V^{i-1}$
- $i=i+1$
- $\pi^i = \text{greedy}(V^{i-1})$
- For $k=\infty$, MPI = PI (Policy Iteration)
- For $k=1$, MPI = VI (Value Iteration)

Implications of This

- MPI is monotone once $V^i$ is pessimistic
- Implies that PI is monotone in general

- Why?
  - $V_\pi$ (the true value function for any policy) will be pessimistic
  - For PI ($k=\infty$), $V^1$ will be pessimistic