Iterative Policy Evaluation

Input $\pi$, the policy to be evaluated
Initialize $V(s) = 0$, for all $s \in S^+$
Repeat
  $\Delta \leftarrow 0$
  For each $s \in S$:
  $v \leftarrow V(s)$
  $V(s) \leftarrow \sum_a \pi(s, a) \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V(s')]$
  $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)
Output $V \approx V^\pi$
Policy Iteration

1. Initialization
   \[ V(s) \in \mathbb{R} \text{ and } \pi(s) \in \mathcal{A}(s) \text{ arbitrarily for all } s \in \mathcal{S} \]

2. Policy Evaluation
   
   Repeat
   \[ \Delta \leftarrow 0 \]
   For each \( s \in \mathcal{S} \):
   \[ u \leftarrow V(s) \]
   \[ V(s) \leftarrow \sum_{s'} \mathcal{P}^{\pi(s)}_{ss'} [R^{\pi(s)}_{ss'} + \gamma V(s')] \]
   \[ \Delta \leftarrow \max(\Delta, |u - V(s)|) \]
   until \( \Delta < \theta \) (a small positive number)

3. Policy Improvement
   
   \textit{policy-stable} \leftarrow true
   
   For each \( s \in \mathcal{S} \):
   \[ b \leftarrow \pi(s) \]
   \[ \pi(s) \leftarrow \arg \max_a \sum_{s'} \mathcal{P}^a_{ss'} [R^a_{ss'} + \gamma V(s')] \]
   If \( b \neq \pi(s) \), then \textit{policy-stable} \leftarrow false
   
   If \textit{policy-stable}, then stop; else go to 2
Value Iteration

Initialize $V$ arbitrarily, e.g., $V(s) = 0$, for all $s \in S^+$

Repeat

\[ \Delta \leftarrow 0 \]

For each $s \in S$:

\[ \nu \leftarrow V(s) \]

\[ V(s) \leftarrow \max_a \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V(s')] \]

\[ \Delta \leftarrow \max(\Delta, |\nu - V(s)|) \]

until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, $\pi$, such that

\[ \pi(s) = \arg \max_a \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V(s')] \]
Gambler’s Problem

- Gambler can repeatedly bet $ on a coin flip
- Heads he wins his stake, tails he loses it
- Initial capital $1, $2, … $99
- Gambler wins if his capital becomes $100
  loses if it becomes $0
- Coin is unfair
  - Heads (gambler wins) with probability $p = .4$

- States, Actions, Rewards?
Gambler’s Problem Solution

Value estimates

Sweep 1
Sweep 2
Sweep 3
Sweep 32

Final policy (stake)
Herd Management

- You are a consultant to a farmer managing a herd of cows
- Herd consists of 5 kinds of cows:
  - Young
  - Milking
  - Breeding
  - Old
  - Sick
You are a consultant to a farmer managing a herd of cows

Herd consists of 5 kinds of cows:
- Young
- Milking
- Breeding
- Old
- Sick

Number of each kind is the State
Number sold of each kind is the Action
Cows transition from one kind to another
Young cows can be born
Asynchronous DP

- All the DP methods described so far require exhaustive sweeps of the entire state set.
- Asynchronous DP does not use sweeps. Instead it works like this:
  - Repeat until convergence criterion is met:
    - Pick a state at random and apply the appropriate backup
- Still need lots of computation, but does not get locked into hopelessly long sweeps
- Can you select states to backup intelligently? YES: an agent’s experience can act as a guide.
Generalized Policy Iteration (GPI):
any interaction of policy evaluation and policy improvement, independent of their granularity.

A geometric metaphor for convergence of GPI:
To find an optimal policy is polynomial in the number of states…

BUT, the number of states is often astronomical, e.g., often growing exponentially with the number of state variables (what Bellman called “the curse of dimensionality”).

In practice, classical DP can be applied to problems with a few millions of states.

Asynchronous DP can be applied to larger problems, and appropriate for parallel computation.

It is surprisingly easy to come up with MDPs for which DP methods are not practical.
Summary

- Policy evaluation: backups without a max
- Policy improvement: form a greedy policy, if only locally
- Policy iteration: alternate the above two processes
- Value iteration: backups with a max
- Full backups (to be contrasted later with sample backups)
- Generalized Policy Iteration (GPI)
- Asynchronous DP: a way to avoid exhaustive sweeps
- **Bootstrapping**: updating estimates based on other estimates