Policy Iteration

Frederik Mallmann-Trenn
6CCS3AIN

Policy iteration

- Rather than compute optimal utility values, policy iteration looks through the space of possible policies.
- Starting from some initial policy π_0 we do:
  - Policy evaluation
    Given a policy π_i, calculate $U_i(s)$.
  - Policy improvement
    Given $U_i(s)$, compute π_{i+1}
- We will look at each of these steps in turn.
- But not in order.

Policy improvement

- Easy
- Calculate a new policy π_{i+1} by applying:

$$\pi_{i+1}(s) = \arg\max_{a \in A(s)} \sum_{s'} P(s'|s,a) U_i(s')$$

- For each state we do a one-step MEU lookahead.
- A simple decision.
- Use the values established by policy evaluation.

Policy evaluation

- How do we calculate the utility of each step given the policy π_i?
- Turns out not to be so hard.
- Given a policy, the choice of action in a given state is fixed (that is what a policy tells us) so:

$$U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$$

- Again there are lots of simultaneous equations, but now they are linear (no $\max$) and so standard linear algebra solutions will work.
- (This is because we use the same U_i on the lhs and rhs)

Policy iteration

- Put these together to get:
- Starting from some initial policy π_0 we do:
  1. Policy evaluation
     Compute:
     $$U_i(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$$
     for every state.
  2. Policy improvement
     Calculate a new policy π_{i+1} by applying:
     $$\pi_{i+1}(s) = \arg \max_{a \in A(s)} \sum_{s'} P(s'|s, a) U_i(s')$$
     for every state s.

  Until convergence.

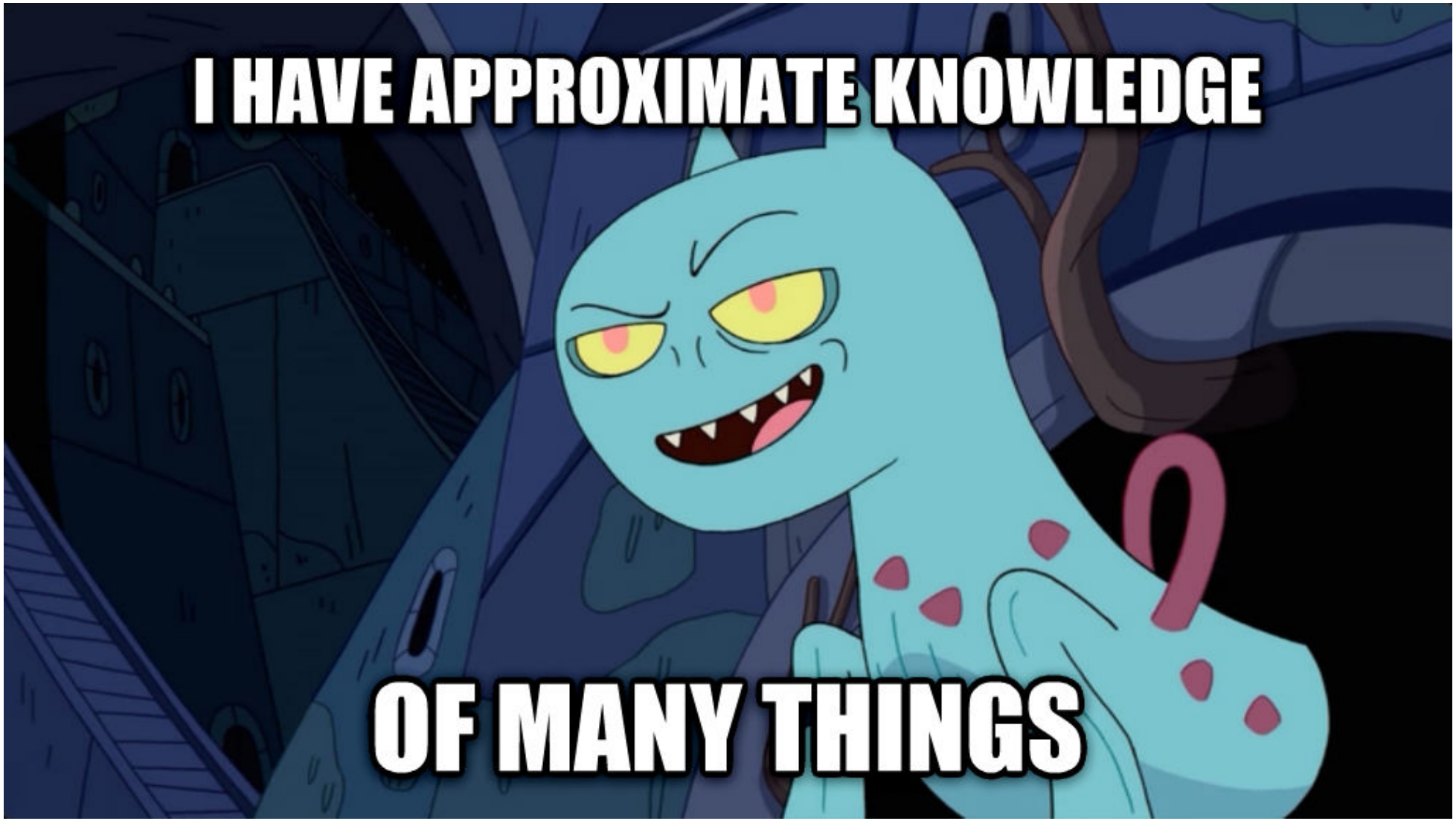
Policy iteration

- The iteration will terminate when there is no improvement in utility from one iteration to the next.
- At this point the utility U_i is a fixed point of the Bellman update and so π_i must be optimal.

Policy evaluation

- There is a problem with the policy evaluation stage of the policy iteration approach.
- If we have n states, we have n linear equations with n unknowns in the evaluation stage.
- Solution in $O(n^3)$ (there is also an impractical solution with $O(n^{2.376})$).
- For large n, can be a problem.
- So, an approximate solution.

Approximate?



(Pendleton Ward/Cartoon Network)

Approximate policy evaluation

- Run a simplified value iteration.
- Policy is fixed, so we know what action to do in each state.
- Repeat:

$$U_{i+1}(s) \leftarrow R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$$

  a fixed number of times.

Modified policy iteration

- Starting from some initial policy π_0 we do:
  1. Approximate policy evaluation
     Repeat
     $$U_{i+1}(s) \leftarrow R(s) + \gamma \sum_{s'} P(s'|s, \pi_i(s)) U_i(s')$$
     a fixed number of times. The difference to policy iteration is the arrow.
  2. Policy improvement
     $$\pi_{i+1}(s) = \arg \max_{a \in A(s)} \sum_{s'} P(s'|s, a) U_i(s')$$
     for every state s.

  Until convergance
- Often more efficient than policy iteration or value iteration.

Solving MDPs

- Have covered three methods for solving MDPs
  - Value iteration
    (Exact)
  - Policy iteration
    (Exact)
  - Modified policy iteration
    (Approximate)
- Which to use is somewhat problem specific.

Bellman redux

- The Bellman equation(s)/update are widely used.

- D. Romer, It's Fourth Down and What Does the Bellman Equation Say? A Dynamic Programming Analysis of Football Strategy, NBER Working Paper No. 9024, June 2002

Bellman redux

This paper uses play-by-play accounts of virtually all regular season National Football League games for 1998-2000 to analyze teams' choices on fourth down between trying for a first down and kicking. Dynamic programming is used to estimate the values of possessing the ball at different points on the field. These estimates are combined with data on the results of kicks and conventional plays to estimate the average payoffs to kicking and going for it under different circumstances. Examination of teams' actual decisions shows systematic, overwhelmingly statistically significant, and quantitatively large departures from the decisions the dynamic-programming analysis implies are preferable.