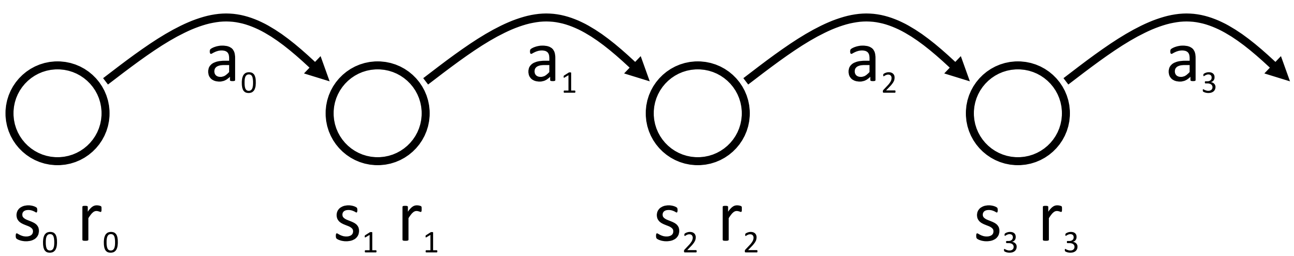
- Midterm Exam Thursday (here)
- Q/A session Wednesday 3PM 1310 DCL
- On Thursday spread out
- Don’t behave suspiciously
- Sit
  - Every other seat
  - Every third row

World Model as Functions

- Transition function
  - T: S x A x S $\rightarrow$ [0,1]
  - T(s,a,·) denotes a probability distribution over next states
  - P(· | s, a) with conditional probability notation
- Reward function
  - Rw: S x $\Re$ $\rightarrow$ [0,1]
  - Each Rw(s ,·) denotes a probability distribution over rewards
- What do we care about?
- R: S $\rightarrow$ $\Re$
- R maps states to expected rewards

If we know T and R…

- We know enough to act optimally (although the algorithm is inefficient)
- We can estimate T and R from data



T(i,j,k) can be estimated as the ratio:

times action j takes us from state i to state k
divided by # times action j is tried in state i

R(i) can be estimated as the sample average reward in state i

Why Inefficient?

- Rewards are local
- Prefer a global notion of state goodness including discounted future rewards
- This will be policy dependent (why?)
- Utility of a state s given a policy π with discount γ

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)\right]$$

Given that s_0=s and we follow policy π, R&N eqn 21.1 (also 17.2)

If we knew $U^{\pi*}$ and T…

then the optimal policy is obvious.

In state s choose the action with the highest expected utility:

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U^{\pi*}(s')$$

This is equation 17.4 in R&N
(they use conditional probability notation)

Can we estimate U^{π} ?

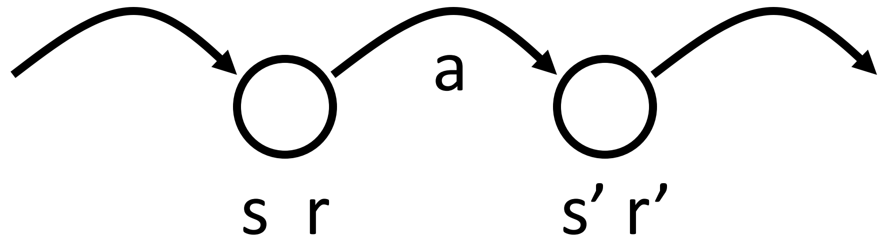
Recall we can already estimate T (how?)

- Initialize U(s) arbitrarily
- Iteratively improve it:

$$\underset{new}{U(s)} \leftarrow \underset{old}{U(s)} + \alpha \cdot error$$

- α is the learning rate $0 < \alpha < 1$
- What is the error?

Utility error



- Assuming policy π chooses action a in s
- Relate $U^{\pi}(s)$ and $U^{\pi}(s')$
- $U^{\pi}(s) = \gamma U^{\pi}(s') + R(s)$
- $U^{\pi}(s) - \gamma U^{\pi}(s') = R(s)$
- If not equal then there is an error
- So error = $R(s) + \gamma U^{\pi}(s') - U^{\pi}(s)$

Temporal Difference Learning

Since error = R(s) + γU(s') – U(s)

$$\underset{new}{U(s)} \leftarrow \underset{old}{U(s)} + \alpha \cdot error$$

Becomes

$$U^{\pi}(s) \leftarrow U^{\pi}(s) + \alpha\left(R(s) + \gamma \cdot U^{\pi}(s') - U^{\pi}(s)\right)$$

This is the TD update equation 21.3 in R&N

A Different Perspective

- TD Update

$$U^{\pi}(s) \leftarrow U^{\pi}(s) + \alpha\left(R(s) + \gamma \cdot U^{\pi}(s') - U^{\pi}(s)\right)$$

- Can be written

$$U^{\pi}(s) \leftarrow (1-\alpha)U^{\pi}(s) + \alpha\left(R(s) + \gamma \cdot U^{\pi}(s')\right)$$

- Or (1-α) (old estimate) + α (new estimate)

R(s) can be avoided

R(s) is E(r_s) so

$$U^{\pi}(s) \leftarrow U^{\pi}(s) + \alpha\big(R(s) + \gamma \cdot U^{\pi}(s') - U^{\pi}(s)\big)$$

can be replaced with

$$U^{\pi}(s) \leftarrow U^{\pi}(s) + \alpha\big(r_s + \gamma \cdot U^{\pi}(s') - U^{\pi}(s)\big)$$

relying on repeated updates to average r_s
and eliminating the explicit estimate of R(s)

This is Value Iteration

- Update the utility of the experienced state
- Take a step to improve $U^{\pi}(s)$
- Rely on repetition
  - Follow π
  - R(s) emerges
  - T(s,a,s’) emerges
  - $U^{\pi}(s)$ emerges
- Note we are neglecting information…

Exploration & Policy Improvement

- Can we change the policy?
- New / better $U^{\pi}(s)$ for some states
- Optimal policy

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U^{\pi^*}(s')$$

- Danger of greedy behavior
- Need exploration

Exploration

- ε–greedy exploration
  - Decaying ε
  - Theoretical vs. practical concerns
  - GLIE – greedy in the limit of infinite exploration
- Optimistic initialization (optimism under uncertainty)
- More principled ways of balancing exploration with exploitation

Policy Iteration

- Fix a policy S $\rightarrow$ A, initially can be arbitrary
- Exercise the policy (Policy Evaluation)
- Gather statistics to estimate U(s) and T(s,a,s')
  - Better T & U estimates expose policy flaws
  - Note T estimates are OK
  - But U are specific to this policy
- Calculate a new policy (Policy Improvement)
  - Maximize the expected discounted utility
  - Use one-step lookahead with new U & T estimates
- Repeat
- Convergence: in probability, utility estimates improve
- We are still neglecting information…

Policy Iteration

estimate T and R

$$U^{\pi}(s) = \underset{\pi}{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)\right]$$

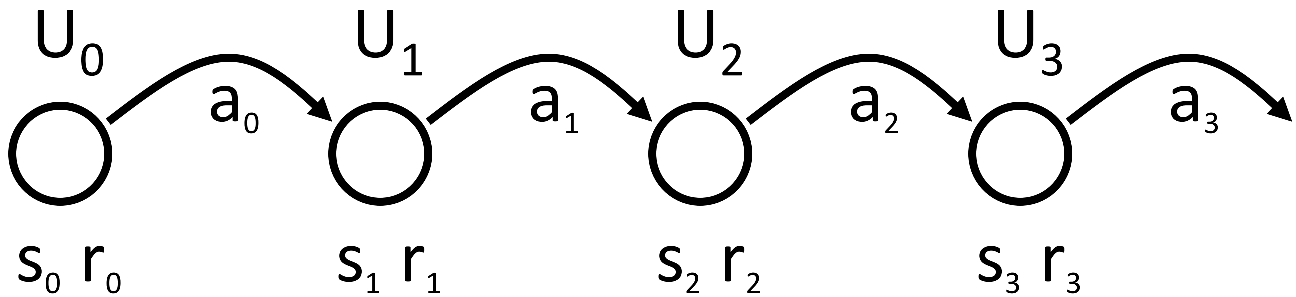
Given that s_0=s and we follow policy π, R&N eqn 21.1 (also 17.2)

$$U^{\pi}(s) = R(s) + \underset{\pi}{E}\left[\sum_{t=1}^{\infty} \gamma^t R(s_t)\right]$$

$$U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s, \pi(s), s') \cdot U^{\pi}(s')$$

Eqn 21.2 also 17.10, assuming we are at policy iteration i

Adaptive Dynamic Programming



Imagine successive value iteration…on s_0 … on s_1 … on s_2:

Perform a_2, update U_2 with r_2 and U_3

U_2 is now updated to a better value

We used the old U_2 to update U_1, shouldn't it be changed as well?

What about U_0?

Fully appreciate each r

Adaptive Dynamic Programming

$$U^{\pi}(s) = R(s) + \gamma \sum_{s'} T(s, \pi(s), s') \cdot U^{\pi}(s')$$

- We estimated R and T
- Consider $U^{\pi}(s)$ as unknowns
- We have n states (n=|S|)
- This is just a system of n linear equations
- Solve numerically or use modified policy iteration