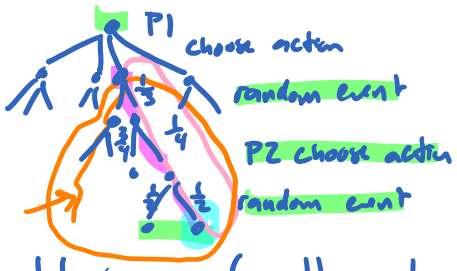
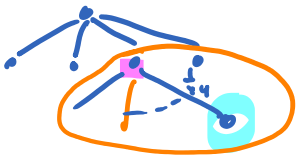
$p(s', r \mid s, a)$ satisfies $p(s', r \mid s, a) = 0$ if s' nonterminal
or s' is P1 win and $r \neq 1$
s' is P2 win and $r \neq -1$
s' is draw and $r \neq 0$

Can you model 2nd player? and P1 won't adjust to P1

If so:





If not: find equilibrium (neither player has incentive to change policy)

$$v^*(s) = \max_a \sum_{s',r} p(s', r \mid s, a) \cdot [r + \gamma \cdot v^*_{opt}(s')]$$

$q(s,a)$

for zero-sum if s is P1 turn

$$v^*(s) = \min_a \sum_{s',r} p(s', r \mid s, a) \cdot [r + \gamma \cdot v^*(s')]$$ if s is P2 turn

Two-player Yahtzee

anchors:

P1

2^{12}
3
106

used/unused Yahtzee unused/0/50 upper total $1, \ldots, 6$

Score difference $-1500 \ldots +1500$

P2

used/unused Yahtzee unused/0/50 upper total

2^{12}
3
106
3000

3 billion times longer than solitaire

Evaluating a Policy

$$v_\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s'} \sum_r p(s', r | s, a) \cdot (r + \gamma v_\pi(s'))$$

$q_\pi(s,a)$

$$= \sum_{s'} \sum_r p(s', r | s, \pi(s)) \cdot (r + \gamma v_\pi(s'))$$

for deterministic π

substitute $v[s']$ for $v_\pi(s')$

Initialize $v[s] \leftarrow 0$ for terminal and arbitrarily for other

$\Delta \leftarrow 0$

repeat

for each state s

$v_{old} \leftarrow v[s]$

let $v[s] \leftarrow$

$\Delta \leftarrow \max(\Delta, |v[s] - v_{old}|)$

until Δ small enough

Policy Iteration

Initialize $\pi[s]$ arbitrarily

repeat
  evaluate π to get v_π
  change ← false
  for each state s
    $a_{old} \leftarrow \pi[s]$
    $\pi[s] \leftarrow \underset{a}{\text{argmax}} \sum_{s',r} p(s',r|s,a) \cdot (r + \gamma v_\pi[s'])$
    if change ← false and $\pi[s] \neq a_{old}$
      change ← true
until change = false