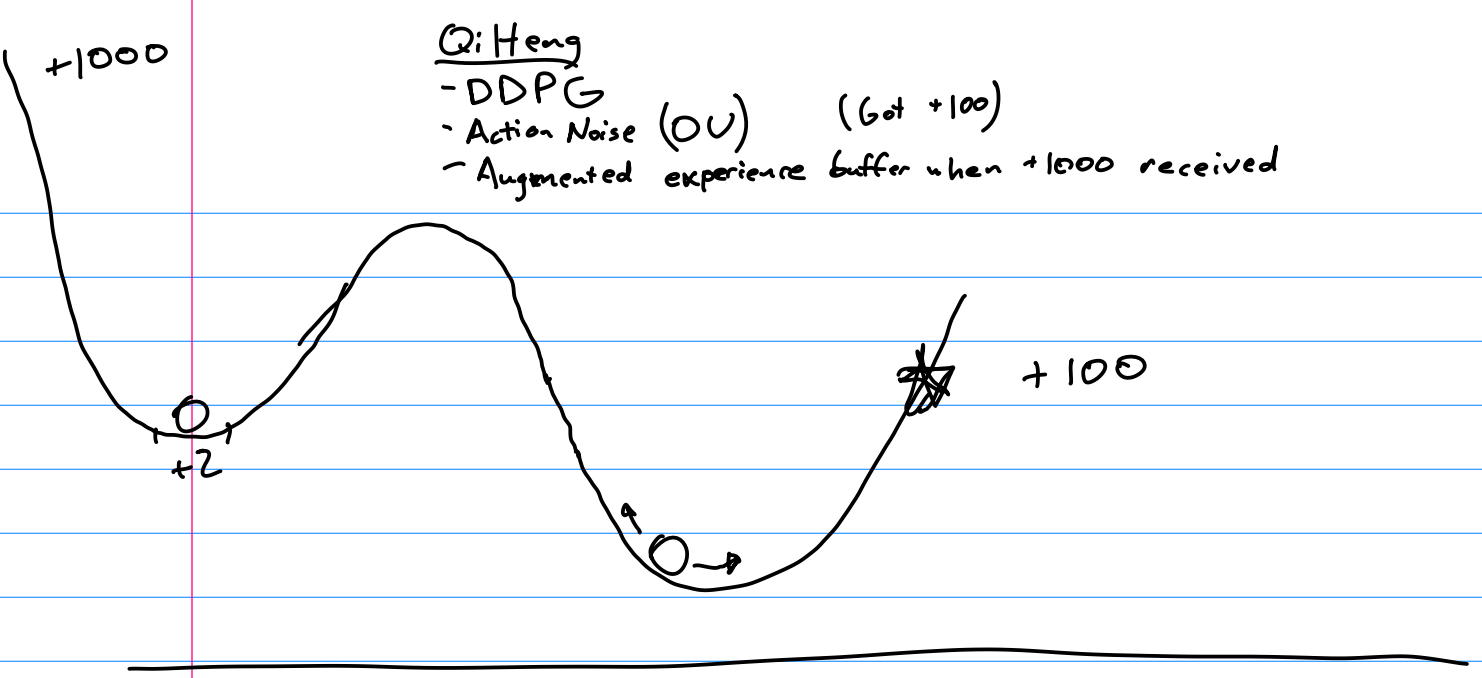


Last Time

The POMDP is PSPACE-Complete — exact algorithms (probably) will have exponential complexity

Numerical Approx

- Offline
  - SARSOP
- Online
  - This Time

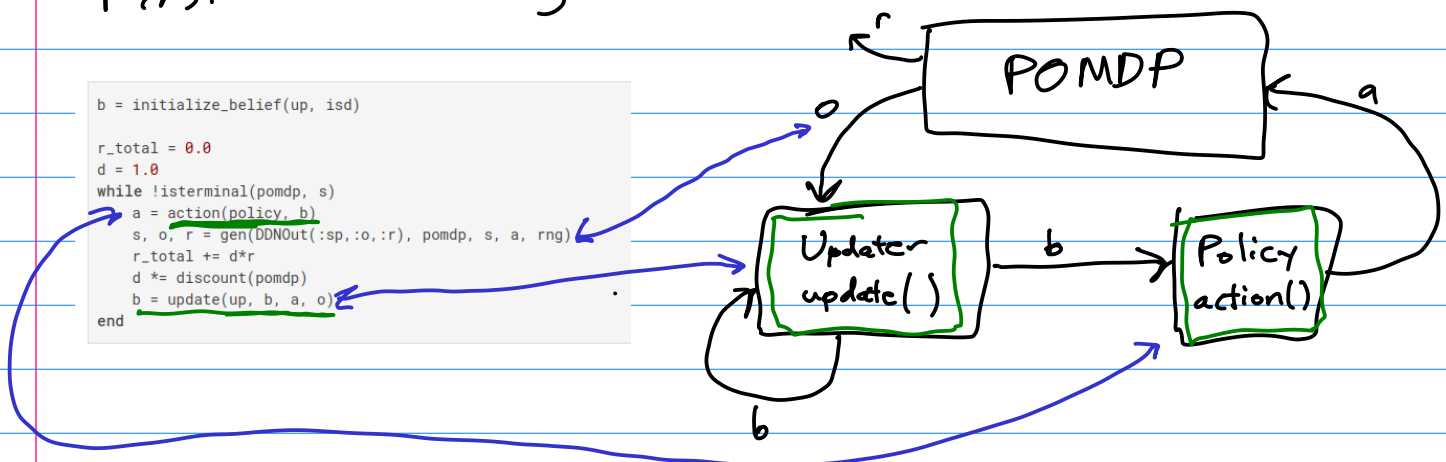
Formulation Approx.

eg. QMDP

First POMDPs.jl for HW 6

```
b = initialize_belief(up, isd)

r_total = 0.0
d = 1.0
while !isterminal(pomdp, s)
    a = action(policy, b)
    s, o, r = gen(DDNOut(:sp,:o,:r), pomdp, s, a, rng)
    r_total += d*r
    d *= discount(pomdp)
    b = update(up, b, a, o)
end
```



QMDP

$$\alpha_a^{(k+1)}[s] = R(s,a) + \gamma \sum_{s'} T(s'|s,a) \max_{a'} \alpha_{a'}^{(k)}[s']$$

$\max_{a'} \alpha_{a'}^{(k)}[s'] = Q_{MDP}(s',a')$

$= V_{MDP}(s')$

HW2 →

$$V_{MDP}^{(k+1)}(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} T(s'|s,a) V_{MDP}(s') \right)$$

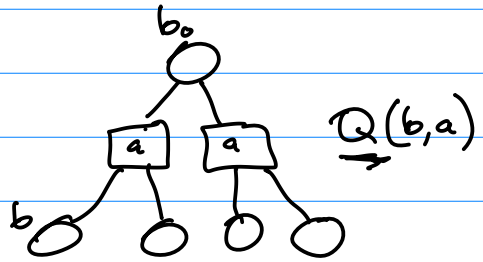
V_{MDP}^*

$$\alpha_a[s] = R(s,a) + \gamma \sum_{s'} T(s'|s,a) V_{MDP}^*(s')$$

$$\pi_{Q_{MDP}}(b) = \operatorname{argmax}_a \underset{s \sim b}{E}[Q_{MDP}(s,a)]$$

Online POMDP Methods

	Name	S	A	O	T	Z
2007	AEMS (Online Planning Algs for POMDPs)	D	D	D	E	E
2010	MCTS ↳ PO-UCT ↳ POMCP	C	D	D	G	G
2013	DESPOT	C	D	D	G	G
2018	POMCPOW	C	D(1-dC)	C	G	E
2019	DESPOT-α	C	D	C	G	E
2020	BOMCP	C	C	C	G	E
2021	VOMCPOW	C	C	C	G	E



Online

Only trying to find a good action for current belief

DESPOT

$Q(b,a)$ only

Offline

Good action for all beliefs

SARSOP

α-vectors

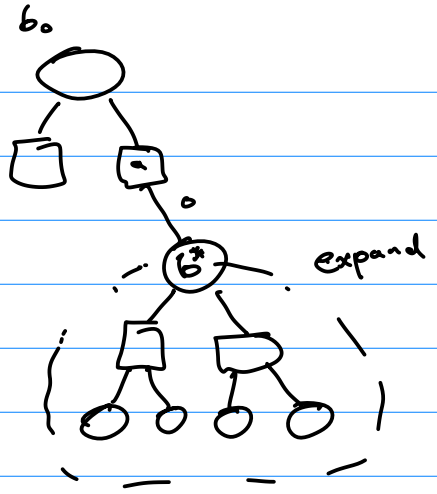
AEMS

while time remains

$b^* = \underset{b \in Fringe(G)}{\text{argmax}}\ E(b)$

expand (b^*)

backup (b^*)



$$E(b) = \gamma^d P(b^d)\, \hat{\varepsilon}(b^d)$$

$$\hat{\varepsilon}(b) = U(b) - L(b)$$

$$P(b^d) = \prod_{i=0}^{d-1} P(o^i \mid b^i, a^i)\, P(a^i \mid b^i)$$

$$P(a \mid b) = \frac{U(a,b) - L(b)}{U(b) - L(b)}$$

$$P(a \mid b) = \begin{cases} 1 & \text{if } a = \underset{a' \in A}{\text{argmax}}\ U(a', b) \\ 0 & \text{otherwise} \end{cases}$$

AEMS 1

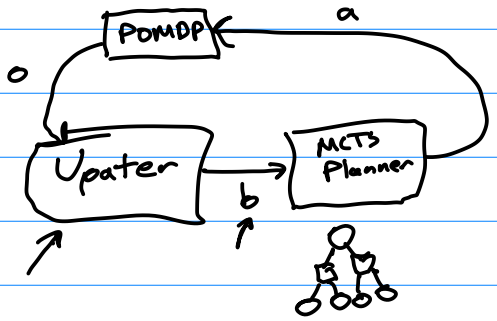
AEMS 2

Breakout Rooms

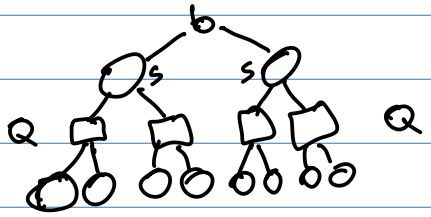
```
function simulate!(π::MonteCarloTreeSearch, s, d=π.d)
    if d ≤ 0
        return 0.0
    end
    𝒫, N, Q, c = π.𝒫, π.N, π.Q, π.c
    𝒜, TR, γ = 𝒫.𝒜, 𝒫.TR, 𝒫.γ
    if !haskey(N, (s, first(𝒜)))
        for a in 𝒜
            N[(h,a)] = 0
            Q[(h,a)] = 0.0
        end
        return rollout(𝒫, s, π.π, d)
    end
    a = explore(π, s)
    s′, r = TR(s,a)
    q = r + γ*simulate!(π, s′, d-1)
    N[(h,a)] += 1
    Q[(h,a)] += (q-Q[(h,a)])/N[(h,a)]
    return q
end
```

$(b, a_1, o_1, a_2, o_2 \ldots o_n)$

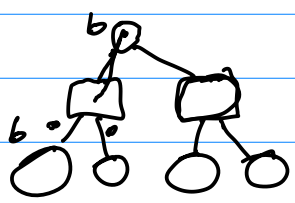
How can we modify MCTS to work for POMDPs avoiding explicit Bayesian Belief Updates?



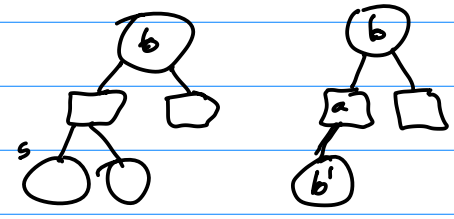
Group 2



Group 1



Real Life Group

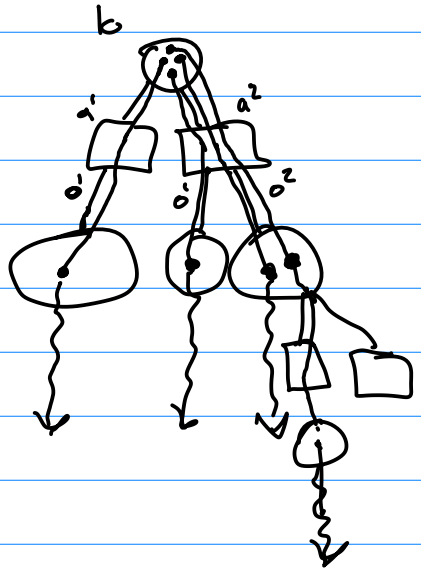


QMDP

Do Work { MCTS, PO-UCT

need $h' \leftarrow G(h,a)$ have $s',o,r \leftarrow G(s,a)$

```
function simulate(π::HistoryMonteCarloTreeSearch, s, h, d)
    if d ≤ 0
        return 0.0
    end
    𝒫, N, Q, c = π.𝒫, π.N, π.Q, π.c
    𝒮, 𝒜, TRO, γ = 𝒫.𝒮, 𝒫.𝒜, 𝒫.TRO, 𝒫.γ
    if !haskey(N, (h, first(𝒜)))
        for a in 𝒜
            N[(h,a)] = 0
            Q[(h,a)] = 0.0
        end
        b = [s == s' ? 1.0 : 0.0 for s' in 𝒮]
        return rollout(𝒫, b, π.π, d)
    end
    a = explore(π, h)
    s', r, o = TRO(s,a)
    q = r + γ*simulate(π, s', vcat(h, (a,o)), d-1)
    N[(h,a)] += 1
    Q[(h,a)] += (q-Q[(h,a)])/N[(h,a)]
    return q
end
```



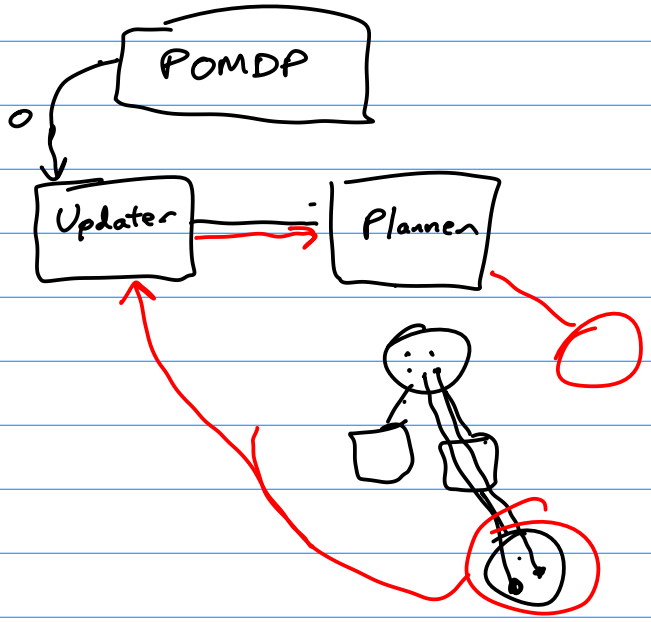
10 actions

10 observations

POMCP

Re-use planning simulations for interaction loop belief updates

Does this work in practice? No



DESPOT not MCTS; Heuristic Search

K scenarios

ϕ_i generate beforehand

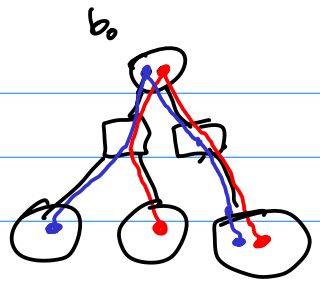
$s' \leftarrow G(s, a, \phi_i[i])$

Example $s' = s + a + w$ $\phi_1[1] = 2$

$G(1,1,\phi_1[1]) = 1+1+2 = 4$

$G(1,2,\phi_1[1]) = 1+2+2 = 5$

weighted, normalized



$U(b)$

$L(b)$

μ

$$a^* = \arg\max_a \mu(b,a)$$

$$\delta^* = \arg\max_\delta \underline{E}(\tau(b,a,o))$$

↑ weighted excess