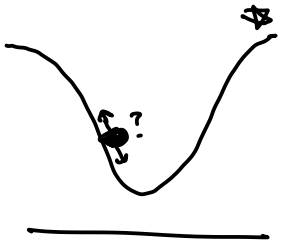
4

HW ← Traditional (10000 episodes)
← Deep Q Learning.jl

give expert
expert 5 episode
1 episode

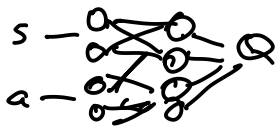
full credit
above 30
45/50 above 25

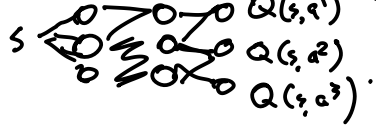


5 actions
[-1, -0.5, 0, 0.5, 1]

DQN

$Q(s,a)$





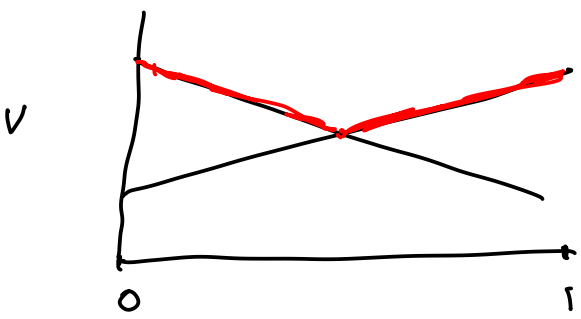
~~Wrong~~ Corrections

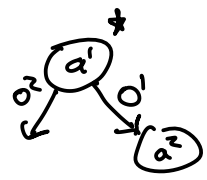
$Q(\lambda)$ Off-policy learning with eligibility traces is unstable

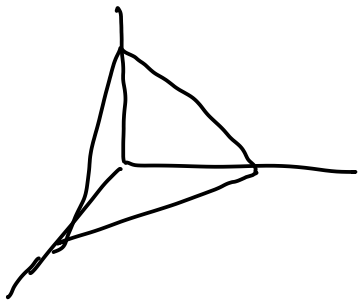
decaying α is a good idea.

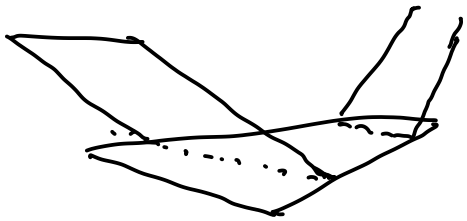
POMDPs

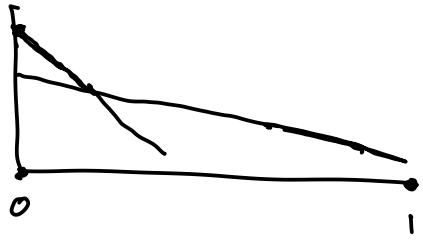
Last Time:
- Value Function ↔ belief
- What do POMDP Policies Look like?







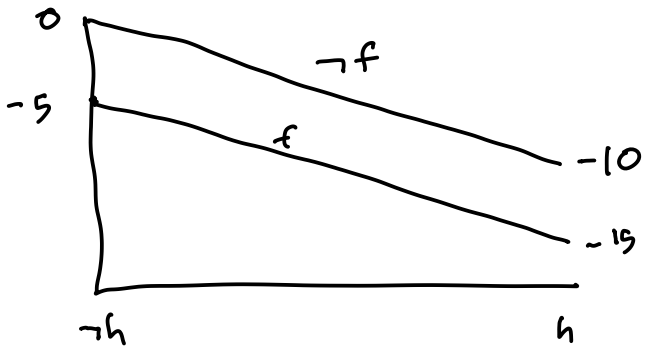




$S = \{h, \neg h\}$ -10 h

$A = \{f, \neg f\}$ -5 f

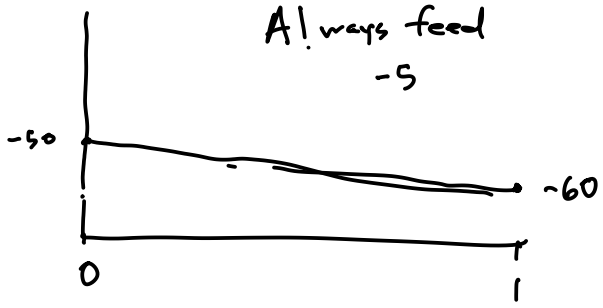
$O = \{c, \neg c\}$



0.9

$\frac{-5}{1-\gamma} = 50$

(f) (¬f)



$$\alpha^{p}[s] = R(s, p_0) + \gamma \sum_{s'} T(s'|s,a) \sum_{o} Z(o|a,s') \alpha'^{p(o)}_{[s]}$$

Γ = set of alpha vectors; p_0 = initial action; $p(o)$ = action after o

Bellman Backup (b, Γ)

for $a \in A$

for $o \in O$

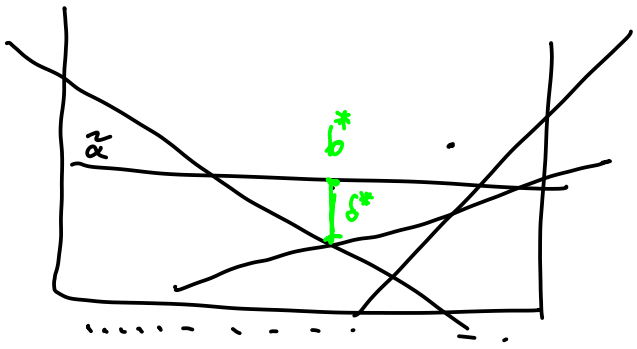
$b' = \tau(b, a, o)$

$\alpha_{ao} = \underset{\alpha \in \Gamma}{\operatorname{argmax}}\ b'^{T}\alpha$

for $s \in S$

$\alpha_a[s] = R(s,a) + \gamma \sum_{s'o} Z(o|s',a)\, T(s'|s,a)\, \alpha_{ao}(s')$

$\Gamma = \Gamma \cup \{\alpha\}$



LP

maximize δ (over δ, b)

subject to

$b^T\alpha \geq \delta + b^T\alpha' \quad \forall \alpha' \in \Gamma$

b is prob: $-\ b^T 1 = 1$

$-\ b \geq 0$

if $\delta^* > 0$

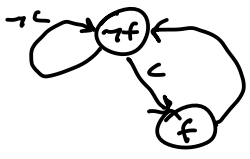
b^* is a "witness"

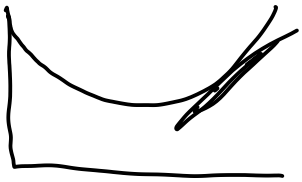
Witness algorithm (1996)

Incremental Pruning (1997)

Baby POMDP

Feed When Crying

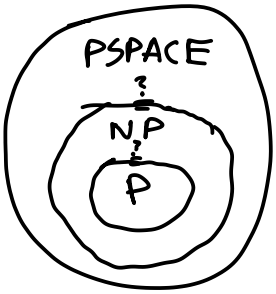




Computational Complexity of POMDPs

"NP-Hard" - at least as hard as all problems in NP

"NP-Complete" - NP-Hard and in NP



PSPACE - can be solved with polynomial memory space

QSAT — known to be PSPACE Complete

$\forall x \exists y \exists z \left((x \vee z) \wedge y\right)$

QSAT can be transformed into a finite horizon POMDP

$\therefore$ POMDPs are PSPACE complete

Approximations

- Numerical
- Optimization objective / Formulation