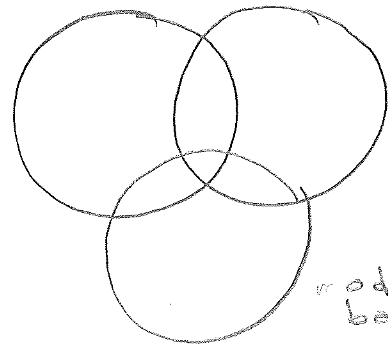
Fast (Sample Efficient) RL

New setting: acting online
every step get to make a decision (take an action)
and observe a next state & reward
Often important to learn "quickly"
multiple definitions of interest
quickly can mean minimizing # of samples
(actions) to learn an optimal policy
quickly may mean that empirical performance
(reward obtained) goes up rapidly as a
function of the # of data points
we will consider / discuss several objectives
but key idea is that care about impact of
amount of data on performance
sample efficiency
not computational efficiency
A lot of work on computationally efficient algorithms
When is fast learning / sample efficiency important
& when is computational efficiency key?
(of course, nice to have both)
- computational efficiency
robotics. consider self-driving vehicle driving
at 60 mph. by the time 1 second has
passed already moved over 26 meters!
quickly computing next action needed
(or, operating hierarchically, e.g. may
have meta-actions or mini-policies
and learn & plan over these)
simulated domains: video games
- sample efficiency
when obtaining data is hard / costly, avoiding
or limiting bad decisions is important
most applications that involve RL agents
interacting w/ people
education: what problem to give a student next
healthcare: what treatment to give a patient
consumer marketing: what ad (or set of ads)
to show
robotics in extreme / remote environments
e.g. Mars scientific exploration by robots

action space size

Types of algorithms



Evaluation criteria

how do we know if an algorithm enables fast learning?

Many possible objectives (& these can yield different algorithms)

1) Empirical performance
   - cumulative reward after fixed # of time steps (decisions)
   - how quickly (# time steps) can learn optimal policy (note: may or may not be executing that policy even if learned it, consider ϵ-greedy exploration approaches)
   - slope of increasing reward: rapidly increase average reward obtaining as a function of the # of timesteps (decisions)

2) PAC: on all but N steps/actions, take an action whose state-action value is near optimal
   N is a polynomial function of the MDP parameters
   Knows What It Knows (KWIK) is a related criteria

3) Regret. Compare total accumulated reward to expected reward from making optimal decisions. Can compare in terms of reward could have obtained if made optimal decisions from the start vs decisions made or for each state reached making actual decisions made, what's the difference in reward from the action taken vs the action on should have taken.
   → Why are these different?
   What is PAC guaranteeing?
   regret bounds tend to be in terms of rates
   e.g. is regret growing linearly, sublinearly as a func of T

time steps ↓ T

4) Bayesian optimality
given initial uncertainty over MDP parameters
maximize expected sum of cumulative rewards
if acting forever almost definitely will still want
to learn optimal policy for states encounter
for finite horizon H where H is short-ish, less clear
value of information
what is the value of information if only
making 1 more decision?

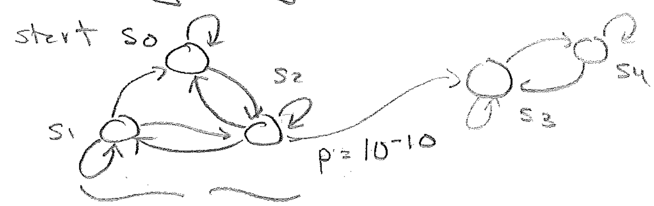
Today: Probably Approximately Correct (PAC) RL
Early ideas pioneered by Kearns & Singh (1998, 2002),
Brafman & Tennenholtz (2002), Sham Kakade (2003)
A lot of later work, esp by Michael Littman & his
group (Alex Strehl, Lihong Li, Tom Walsh, ...)
my group also active in this space
defn: a RL algorithm $\mathcal{A}$ is PAC-MDP if on all but
N steps, the algorithm's non-stationary policy
at time t, $\mathcal{A}_t$ (note: this policy is completely defined
by the algorithm $\mathcal{A}$ & history up to t) is at least
ϵ-optimal from the current state

$$V^{\mathcal{A}_t}(s_t) \geq V^*(s_t) - \epsilon$$

with probability at least $1-\delta$, where
$N =$ polynomial func of $(|S|, |A|, 1/\epsilon, 1/\delta, 1/1-\gamma)$.

Why do we consider only a high probability bound?
given finite data/experience in a stochastic
environment, could always yield experience
that makes it impossible to identify the optimal
π w/perfect confidence

number of mistakes (times take potentially poor
decisions) is bounded
* note: doesn't say when (on what time step)
these mistakes will occur
Why might not all mistakes occur at the start?



can learn to act near optimally here, likely long before visit s_3, s_4

example PAC algorithm: R-max

maintain set of "known" state-action pairs
"unknown" " "

initially all s-a pairs are unknown
intuitively s-a pair becomes known when have a good estimate of the model parameters for that s-a pairs
(can also formulate related intuition for model free algorithms)

define a MDP $\hat{M}$ w/same state & action spaces & γ

for all s, a pairs in K_t (known set)

set $\hat{R}(s,a) = \sum_i r_i(s,a) / \#$ times seen s,a (e.g. empirical avg reward)

$\hat{T}(\cdot \mid s,a) =$ empirical estim of trans model

for all $(s,a) \notin K_t$ (unknown)

$\hat{R}(s,a) = R_{max}$

$\hat{T}(s \mid s,a) = 1$ (self loop)

compute $\hat{\pi}^*$ for $\hat{M}$
act using $\hat{\pi}^*$
keep counts of # visits to each (s,a) pair
if counts$(s,a) \geq m$, add (s,a) to known set K_t

intuitively what does this algorithm do?
optimism under uncertainty

with correct setting of m, R-max is a PAC algorithm

proof approach (Strehl et al UAI 2006)

generic proof that can be used for related algorithms
assume RL alg $\mathcal{A}$ maintains $Q(s,a)$
let $Q_t(s,a)$ be $\mathcal{A}$'s estimate immediately before taking t^{th} action
$\mathcal{A}$ is a "greedy" alg if $a_t = \arg\max_a Q_t(s_t, a)$

- define known s-a MDP $M_K = \langle S \cup \{s_0\}, A, T_K, R_K, \gamma \rangle$
related to a MDP $M = \langle S, A, T, R, \gamma \rangle$. Let K be the set of known s-a pairs and Q be a set of state-action values (not necessarily related to values of M or M_K).

$\forall s,a \in K$, $R_K(s,a) = R(s,a)$ $T_K(\cdot \mid s,a) = T(\cdot \mid s,a)$ $\Big\}$ true model parameters of M

$R(s_0, \cdot) = 0$ $T(s_0 \mid s_0, \cdot) = 1$ (sink state)

$\forall s,a \notin K$ $R_K(s,a) = Q(s,a)$ $T_K(s_0 \mid s,a) = 1$

Proposition 1. Let $A(\epsilon, \delta)$ be any greedy learning alg s.t. for every time step t there exist a set of K_t state-action pairs. $K_t = K_{t+1}$ unless during timestep t an unknown (s,a) pair is visited or a Q value is updated. M_{K_t} is the known s-a MDP and $\pi_t(s) = \arg\max_a Q_t(s,a)$.
Suppose that for any inputs ϵ & δ, w/prob $\geq 1-\delta$ the following holds $\forall s, a, t$

1) Optimism: $Q_t(s,a) \geq Q^*(s,a) - \epsilon$
2) Accuracy: $V_t(s) - V_{M_{K_t}}^{\pi_t}(s) \leq \epsilon$
3) Learning complexity bounded: total # of updates of Q estimates and the # of times visit an unknown (s,a) pair is bounded by $\zeta(\epsilon, \delta)$.

Then when $A(\epsilon, \delta)$ is executed on any MDP M, it will follow a 4ϵ-optimal π from its current state on all but

$O\left(\frac{\zeta(\epsilon,\delta)}{\epsilon(1-\gamma)^2} \ln \frac{1}{\delta} \ln \frac{1}{\epsilon(1-\gamma)}\right)$ timesteps w/prob $\geq 1 - 2\delta$.

Proof sketch. Assume $R \in [0,1]$

Define $D = \frac{1}{1-\gamma} \ln \frac{1}{\epsilon(1-\gamma)}$

$|V_{M_{K_t}}^{\pi}(s, D) - V_{M_{K_t}}^{\pi}(s)| \leq \epsilon$ } finite horizon value ϵ-close to ∞ horizon value if horizon long enough (Kearns & Singh 2002)

let A_t be the current π of the agent (nonstationary)
let ω be the event that by executing A_t starting in state s_t for D steps, one of the 2 occurs
1) the algorithm successfully updates $Q(s,a)$ for some (s,a)
2) visit an unknown (s,a) pair

over D steps

can bound value of A_t in true MDP by prob of ω

$V_M^{A_t}(s_t, D) \geq V_{M_{K_t}}^{\pi_t}(s_t, D)(1 - Pr(\omega)) + Pr(\omega) \cdot$ value when ω occurs

note $A_t = \pi_t$ unless update $Q_t(s,a)$
M and M_{K_t} are identical on all $(s,a) \in K_t$
so get exact same behavior & rewards unless event ω happens

$\geq V_{M_{K_t}}^{\pi_t}(s_t, D) - Pr(\omega)\, ^1/_{1-\gamma}$ (last term pos, $V \leq \frac{1}{1-\gamma}$)

$\geq V_{M_{K_t}}^{\pi}(s_t) - \epsilon - Pr(\omega)\, ^1/_{1-\gamma}$ defn of D

$\geq V(s_t) - 2\epsilon - Pr(\omega)\, ^1/_{1-\gamma}$ by accuracy condition

$\geq V^*(s_t) - 3\epsilon - Pr(\omega)\, ^1/_{1-\gamma}$ by optimism "

all hold w/prob $\geq 1 - \delta$

If $Pr(\omega) < \epsilon(1-\gamma)$

$V_M^{A_t}(s_t) \geq V_M^{A_t}(s_t, D) \geq V^*(s_t) - 4\epsilon$

proof of Prop. 1 continued

if $Pr(\omega) > \epsilon(1-\gamma)$

want to know # of D-intervals until # of ω events $= \zeta(\epsilon,\delta)$ (since know from cond 3 this bounds the # of ω events)

to get upper bound, treat each D-interval as iid opportunity for ω to occur w/prob $\geq \epsilon(1-\gamma)$

lemma 56 (Lihong Li, PhD thesis, 2009): let $X_1, \ldots X_m$ be a seq of m indep Bernoulli trials, each w/a success prob of at least μ: $E[X_i] \geq \mu > 0$. Then for any $k \in \mathbb{N}$ and $\delta \in (0,1)$, with prob $\geq 1-\delta$,

$$X_1 + X_2 + \ldots X_m \geq k$$

if

$$m \geq \frac{2}{\mu}\left(k + \ln\frac{1}{\delta}\right).$$

applying lemma 56 we get that if the # of intervals is

$$\geq \frac{2}{\epsilon(1-\gamma)}\left(\zeta(\epsilon,\delta) + \ln 1/\delta\right)$$

then w/high prob all $\zeta(\epsilon,\delta)$ ω events will have occured.

this yields since each interval is D-steps

$$\geq \frac{2D}{\epsilon(1-\gamma)}\zeta(\epsilon,\delta)\ln(1/\delta) \leftarrow \text{since } \ln(1/\delta) > 1$$

$$D = \frac{1}{1-\gamma}\ln\frac{1}{\epsilon(1-\gamma)}$$

so # time steps on which $Pr(\omega) > \epsilon(1-\gamma)$ is upper bounded by

$$\frac{2D}{\epsilon(1-\gamma)^2}\zeta(\epsilon,\delta)\ln(1/\delta)\ln\left(\frac{1}{\epsilon(1-\gamma)}\right) \qquad \Box$$

how do we achieve the required preconditions?

simulation lemma & bounds on how far empirical model param estimates can be from true parameters (see earlier lecture)

yield accuracy requirement

also yield criteria to make (s,a) pair known

pigeonhole principle bounds # times can visit unknown (s,a) pair

optimism comes from simul lemma plus using optimistic estimate for all unknown (s,a) pairs

what is final sample complexity (# of steps not necessarily near optimal?) for R-max*

$$\text{ignore log} \rightarrow \tilde{O}\left(\frac{S^2 A}{\epsilon^3(1-\gamma)^6}\right)$$

Are these practical? No
Can we do better? Yes.

- Lattimore & Hutter (ALT 2012). discounted MDP

$$\tilde{O}\left(\frac{N_{ssa}}{\epsilon^2 (1-\gamma)^3}\right) \qquad N_{ssa} = \#\text{ non-zero } s,a \text{ trans}$$

$$\leq \tilde{O}\left(\frac{|S|^2|A|}{\epsilon^2 (1-\gamma)^3}\right)$$

much better (by $\frac{1}{\epsilon} \frac{1}{(1-\gamma)^3}$) than prior bounds in terms of ϵ & γ

can be worse in state ~~dep~~

- Dann & Brunskill (NIPS 2015). finite horizon episodic MDP

$$\leq \tilde{O}\left(\frac{|S|^2|A|}{\epsilon^2} H^3\right) \text{ time steps}$$

lower bound

$$\tilde{O}\left(\frac{|S||A|H^3}{\epsilon^2}\right) \text{ time steps}$$

1 of key insights of the above approaches: not all (s,a) pairs are equally likely to be visited nor influence value func equally

Bounds still impractical
but ideas are useful
optimism under uncertainty (goes back, sans theory, at least to early 1990s Kaelbling)
quantifying "enough" info
many extensions, alternatives
don't have to be binary in "knowness" but still use info about state-action experience
build in rep of uncertainty
e.g. confidence bounds