Last Time

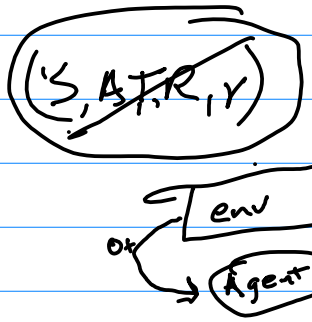
Continuous ADP MPC

Sparse Tree search

Progressive widening

Reinforcement Learning

Max-Likelihood Model-Based Tabular RL



This time

Julia tips

Exploration vs Exploitation : Bandits

Max-Likelihood Model-Based Tabular RL

$N \leftarrow 0$

$\rho \leftarrow 0$

$s \leftarrow s_0$

loop

choose a with π

$r = act!(env, a)$

$s' = observe(env)$

$N(s,a,s') \mathrel{+}= 1$

$\rho(s,a) \mathrel{+}= r$

estimate $\hat{T}$ from N

estimate $\hat{R}$ from ρ

$\pi = solve(S, A, \hat{T}, \hat{R}, \gamma)$ ← every time you get to a terminal state

$s \leftarrow s'$

$N(s,a,s')$

$\rho(s,a)$

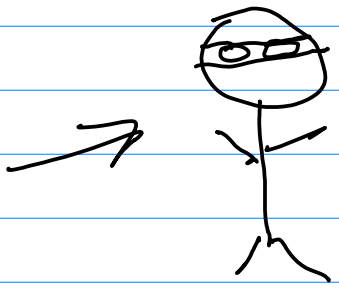
$N \rightarrow T$

$N, \rho \rightarrow R$

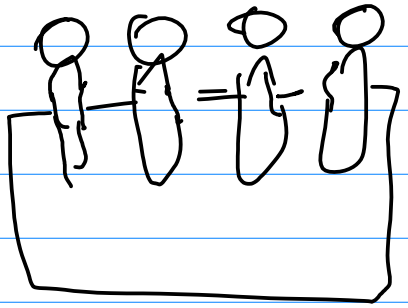
Solve with VI

$$T(s'|s,a) = \frac{N(s,a,s')}{\sum_{s'} N(s,a,s')}$$

Bandits



Multi-Armed Bandit



which arm to choose based on experience

Bournoulli Bandit

$A = \{1 \dots n\}$ / Θ_a = probability of $r_t = 1$

Breakout Rooms

4-arm Bernoulli Bandit: w_i, l_i

How do you choose next action.

B.

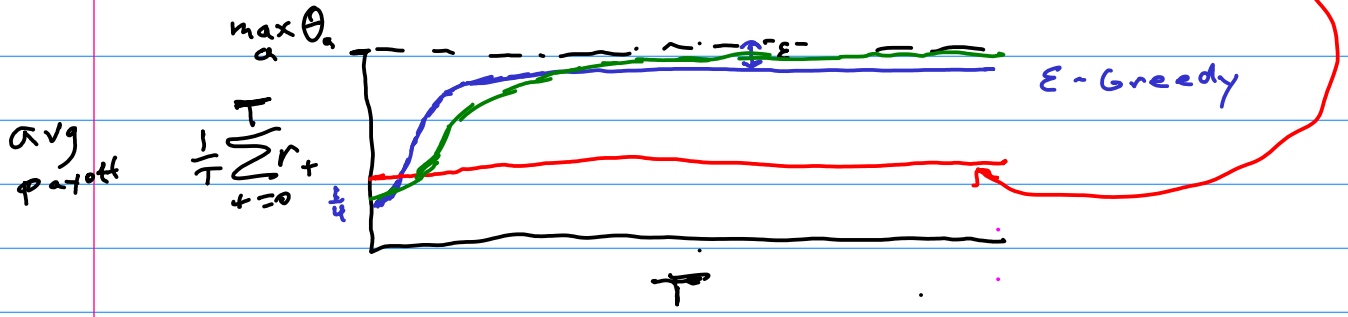
Team 2

if $r_t = 1$

$a_{t+1} = a_t$

else

$a_{t+1} = a_t + 1$



Bandit Algos

ρ_a = estimate of θ_a
$= \frac{w_a}{w_a + l_a}$

- Greedy $\quad \underset{a}{\text{argmax}}\ \rho_a \quad \varnothing$ no exploration

Pure Exploitation →

- Explore then commit: random for first k steps, $\underset{a}{\text{argmax}}\ \rho_a$ after that

- ε-Greedy $\quad \underset{a}{\text{argmax}}\ \rho_a$ w.p. $(1-\varepsilon)$, random otherwise
  Decay $\varepsilon_{t+1} \leftarrow \alpha \varepsilon_t$ where $\alpha \in [0, 1)$

less regret ↓

- softmax: choose a w.p. $\frac{e^{\lambda \rho_a}}{\sum_a e^{\lambda \rho_a}}$

- Interval Selection
  α

- UCB $\quad \underset{a}{\text{argmax}}\ \rho_a + c\sqrt{\frac{\log N}{N(a)}}$ logarithmic regret

Bayesian

- Thompson Sampling
  - keep track of $p(\theta_a)$
  - sample $\hat{\rho}_a$
  - choose $\underset{a}{\text{argmax}}\ \hat{\rho}_a$

Optimal

- Optimal DP solution
  - Usually not tractible
  - Gittin's Index

T

regret $= \theta^* N - \sum_{t=1}^{N} r_t$
for Bernoulli