- Next: Foundations and KR for ML
  - Read Chapters 13 and 14
  - Uncertainty, Statistics, Probabilistic Reasoning
- VERY APPROXIMATE Grades are posted on Compass
  - G / UG curved separately (as announced)
  - HW2 is not included
  - Your position in distributions and honest assessment are more informative
  - Is your behavior predictive of the whole course?

Indirect / Direct RL

- Our TD RL is *indirect* RL
- Policy is constructed from world model
- Consider learning the policy *directly*
- Forgo learning the transition function T
- AKA Model based RL / Model free RL
- General distinctions in learning
  - Full or joint model / Conditional model
  - Generative model / Discriminative model

Q Learning: Direct RL

- Q function: Q: A x S $\rightarrow \Re$
- Q(a,s) - the expected utility of performing action a in state s
- The greedy policy is simpler:

$$\pi^*(s) = \arg\max_a Q^*(a,s)$$

- Recall model-based greedy policy:

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U^{\pi^*}(s')$$

- (Recall the need for exploration)

Q Learning
Off Policy Learner

- Definition of true Q

$$Q(a,s) = R(s) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q(a',s')$$

~Eqn 21.7

- Q update rule

$$Q(a,s) \leftarrow Q(a,s) + \alpha \cdot \left(R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s) \right)$$

Eqn 21.8

- Or

$$Q(a,s) \leftarrow (1-\alpha) \cdot Q(a,s) + \alpha \cdot \left(R(s) + \gamma \max_{a'} Q(a',s') \right)$$

We can avoid R in Q also

- Q update rule

$$Q(a,s) \leftarrow Q(a,s) + \alpha \cdot \left(R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s) \right)$$

Eqn 21.8

- Sampled r_s

$$Q(a,s) \leftarrow Q(a,s) + \alpha \cdot \left(r_s + \gamma \max_{a'} Q(a',s') - Q(a,s) \right)$$

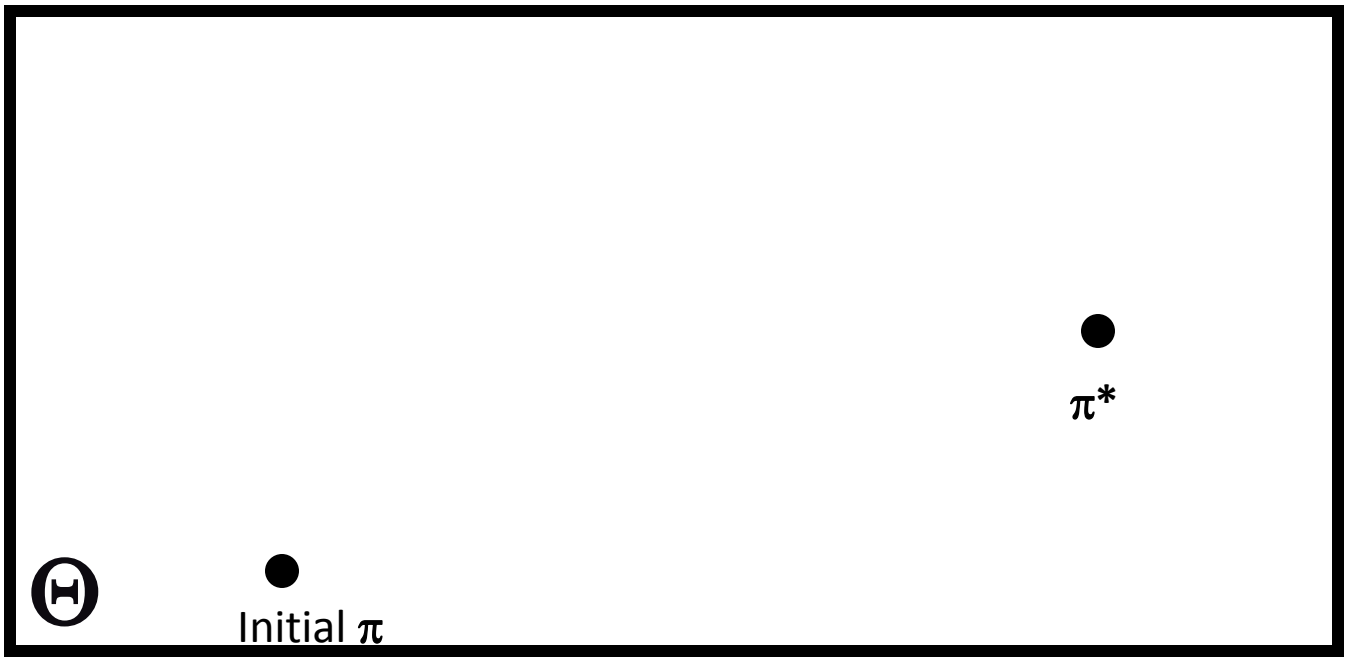
How Many Numbers for a Policy?

Policy: State $\rightarrow$ Action

- Policy for TD Value Iteration?
  - T, U
  - T: $\mathfrak{R}^{|S|\cdot|A|\cdot|S|}$
  - U: $\mathfrak{R}^{|S|}$
  - $|S|^2\cdot|A| + |S|$ real numbers
- Policy for Q?
  - Q
  - Q: $\mathfrak{R}^{|S|\cdot|A|}$
  - $|S|\cdot|A|$ real numbers

Learning in Parametric Policy Space

- Parameters Θ determine the policy



- Policy / Value iteration as following gradients

Can we ever stop learning / exploring?

Can we ever stop trying action a_2 in state s_7

Two-arm bandit problem

- Arm1 pays with probability p1
- Arm2 pays with probability p2

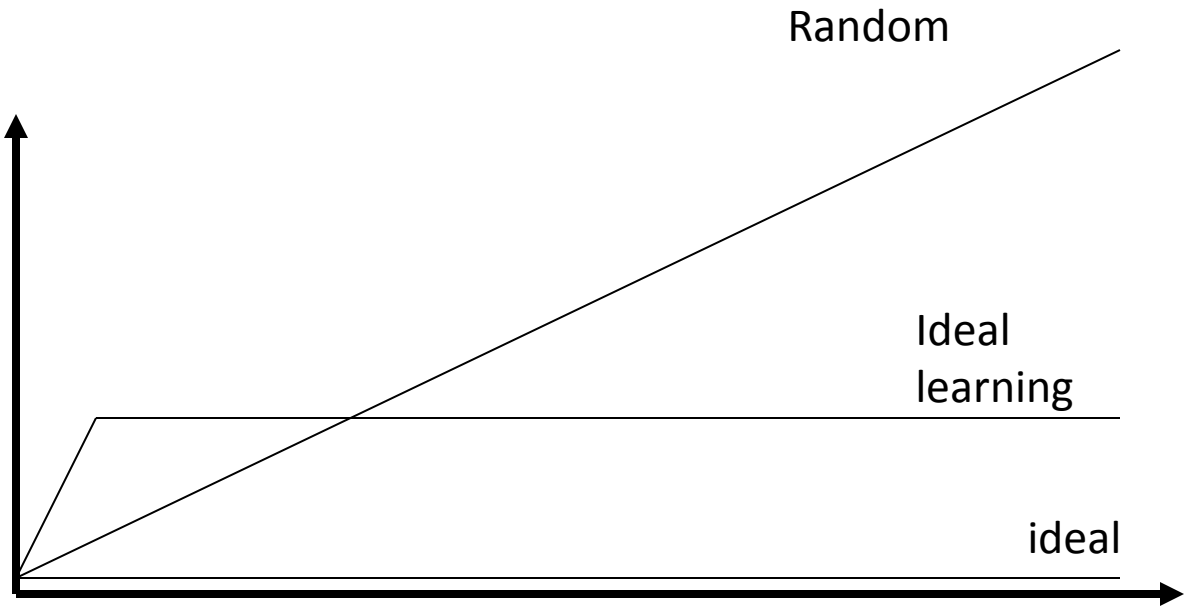
What's the optimal policy?

Pull Arm1 if p1 > p2

Pull Arm2 otherwise

Can we learn the optimal policy for the two-arm bandit?

- Regret function
  - Accumulated deficit of not following the optimal policy
- Regret curves: Ideal, Random, Ideal learning



Theorem

- Regret grows as Sqrt (N)
- In fact

  $0.264*\text{Sqrt}(N) < \text{Regret} < 0.376*\text{Sqrt}(N)$
- This function grows without bound
- We can never stop trying the non-preferred arm
- We can never stop trying action a_2 in state s_7

What is On / Off Policy?

- Q learns how to perform optimally even when we are following a non-optimal policy
- In ε–greedy, ε leaves no trace in Q
- SARSA is on-policy
- Learns the best policy given our systematic departures from true optimal
- In ε–greedy, ε is reflected within SARSA's Q values

On Policy vs Off Policy

- Q, an off-policy learner:

$$Q(a,s) \leftarrow Q(a,s) + \alpha \cdot \left(r_s + \gamma \max_{a'} Q(a',s') - Q(a,s) \right)$$

- SARSA, an on-policy learner:

$$Q(a,s) \leftarrow Q(a,s) + \alpha \cdot \left(r_s + \gamma \cdot Q(a',s') - Q(a,s) \right)$$

- How do they differ?
  - Remember the persistent need for exploration
  - Consider cliff walking

Can we use more Information?

- Efficient = use all information
- No information until rewards
- Information is propagated one step back
- Can we do better?

Eligibility traces

Another parameter: λ

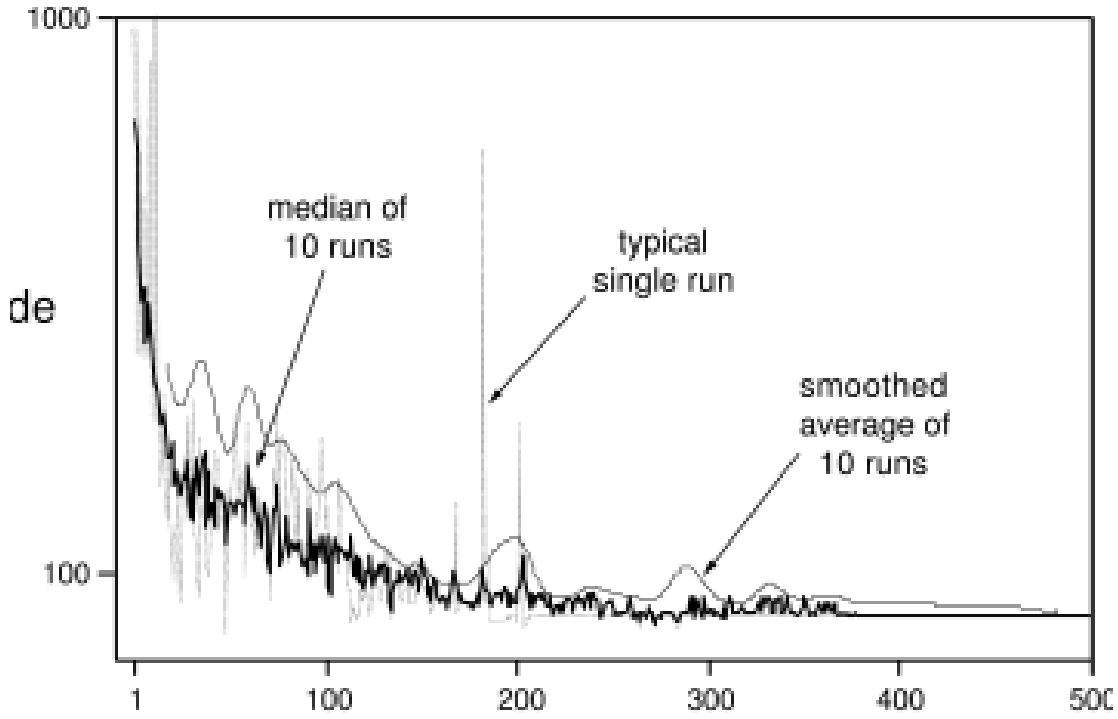
Between 0 and 1 (close to 1)

TD(λ), Q(λ), SARSA(λ)

Reinforcement Learning vs. Classical Planning

- More robust
  - Fewer & weak *a priori* assumptions (esp. actions)
  - Empirical model
  - Fit (via parameter adjustment) to the *observed* world
- Scaling difficulties
  - Propositional expressiveness
  - Space complexity
  - States / Features (e.g., block positions)
  - Time complexity
  - Planning vs. Learning
  - Recognizing convergence
- Markov assumption
  - Our world?
  - Discretizing may not respect Markov

Typical Learning Curve



RL Links

- Eligibility Traces
  http://www.cs.ualberta.ca/~sutton/book/7/node1.html
- RoboCup at AAAI conferences
- Cobot.research.att.com

many others