CS 188 SECTION 5

ABOUT ME

- Daylen Yang
- daylen@berkeley.edu (but use Piazza!)
- Sections MW 4-5pm in 310 Soda
- Office Hours Thursdays 4-6pm in 411 Soda

UPCOMING DEADLINES

- HW 2 due **tonight** @ 11:59
- Project 2 due **Friday** 7/8 @ 5pm
- Practice Midterm 1 (1 pt EC) due **Saturday** 7/9 @ 11:59
- Midterm 1 on **Monday** 7/11, 12–3pm

MDPs vs. Reinforcement Learning

- If you **know** the transition and reward functions…
  - you have a **Markov Decision Process**, which you can solve offline using value or policy iteration
  - See the Discussion 4 worksheet
- If you **don't know** the transition or reward functions…
  - you must do **reinforcement learning**!

REINFORCEMENT LEARNING

- **Model-Based**: learn the MDP, then solve it
- **Model-Free**:
  - Direct Evaluation: learns $V^{\pi}(s)$ inefficiently
  - Temporal Difference Learning: learn $V^{\pi}(s)$

Sample of V(s): $sample = R(s, \pi(s), s') + \gamma V^{\pi}(s')$

Update to V(s): $V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + (\alpha)sample$

  - Q-Learning: learn Q(s, a)

Consider your new sample estimate:

$$sample = R(s, a, s') + \gamma \max_{a'} Q(s', a')$$

Incorporate the new estimate into a running average:

$$Q(s, a) \leftarrow (1-\alpha)Q(s, a) + (\alpha)\,[sample]$$

EXPLORATION VS EXPLOITATION

- Epsilon-Greedy Exploration: With probability epsilon, act randomly. Else, act on current policy.
- Exploration Function: takes a value estimate u and a visit count n, and returns utility $f(u, n) = u + k / n$

PROBLEM 1

FEATURE BASED REPRESENTATIONS

- Describe a state using a vector of features
- Adjust the weight of features. If something bad happens, blame the features and don't prefer states with that state's features

$$\text{transition} = (s, a, r, s')$$

$$\text{difference} = \left[r + \gamma \max_{a'} Q(s', a')\right] - Q(s, a)$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha\,[\text{difference}]$$

$$w_i \leftarrow w_i + \alpha\,[\text{difference}]\, f_i(s, a)$$

PROBLEM 2