Value Iteration

- Computing the optimal value function

$$V_H^*(s) = 0$$

$$V_t^*(s) = \max_{a \in \mathcal{A}} \left(\overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) V_{t+1}^*(s') \right)$$

- Computing an optimal policy

$$\mu_t(s) \in \arg\max_{a \in \mathcal{A}} \left(\overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) V_{t+1}^*(s') \right)$$

- Computing optimal state-action values

$$Q_{H-1}^*(s,a) = \overline{R}_{t,a}(s)$$

$$Q_t^*(s,a) = \overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) \max_{a' \in \mathcal{A}} Q_{t+1}^*(s',a')$$

Infinite-Horizon Problems

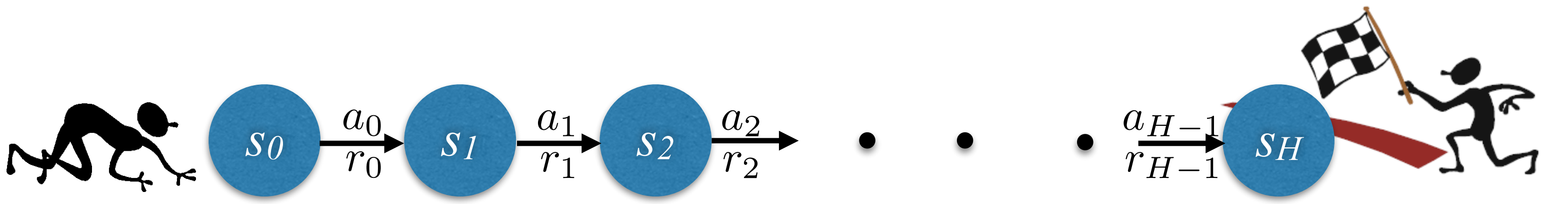
- Discounted reward MDP $(\mathcal{S}, \mathcal{A}, R, P, \alpha, \rho)$
  - Time-homogeneous
  - Discount factor $\alpha \in (0, 1)$
  - Policy $\mu = (\mu_0, \mu_1, \ldots)$
  - Objective

$$\max_{\mu} \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r_t \;\middle|\; a_t = \mu_t(s_t)\right]$$

- Average reward MDP $(\mathcal{S}, \mathcal{A}, R, P, \rho)$
  - Objective

$$\max_{\mu} \liminf_{H \to \infty} \frac{1}{H} \mathbb{E}\left[\sum_{t=0}^{H-1} r_t \;\middle|\; a_t = \mu_t(s_t)\right]$$

Episodic Learning



- Reinforcement learning algorithm
  - Given observations through episode $\ell - 1$
  - Select policy $\mu^\ell = (\mu_0^\ell, \ldots, \mu_{H-1}^\ell)$
  - Apply actions $a_t^\ell = \mu_t^\ell(s_t^\ell)$

- Cumulative reward over L episodes

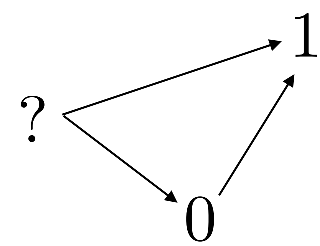
$$\sum_{\ell=1}^{L} \sum_{t=0}^{H-1} r_t^\ell$$

- Simple framework for designing algorithms and generating insight
  - Ideas extend more broadly
  - Overlapping episodes
  - Infinite horizon formulations

Episodic RL Example: Recommendation

- Consider recommending movies
  - N movies
  - Sequence of H recommendations for each customer
  - Customer accepts/rejects each
  - Goal: high acceptance rate

- MDP formulation
  - state: $s_t \in \{0, 1, ?\}^N$
  - action: $a_t \in \{1, \ldots, N\}$
  - reward: $r_t \in \{0, 1\}$



- Episodic learning
  - Customer = episode
  - Learn from customers how to deal with other customers
  - Episodes run in parallel
  - Important to update policy using data from incomplete episodes

- Netflix challenge: generalization
- Is exploration important?
- Are delayed consequences important?

Tabula Rasa Learning

- Model learning
  - Learn transition probabilities $P_{t,a}(s'|s)$
  - Learn expected rewards $\overline{R}_{t,a}(s)$
- Value function learning
  - Learn $Q_t^*(s,a)$
- Policy learning
  - Learn $\mu_t^*(s)$
- *Tabula rasa* learning
  - No generalization across state-action pairs
  - Data requirements grow with # states-action pairs
- Need for generalization
  - Curse of dimensionality
- Begin with *tabula rasa* learning
  - Study exploration with delayed consequences

Asynchronous Value Iteration

- Start with some Q^0
- For $k = 0, 1, 2, \ldots$
  - Select (t^k, s^k, a^k)
  - Obtain Q^{k+1} from Q^k
    - Apply value iteration update at (t^k, s^k, a^k)
- Note that
  - Q^{k+1} only differ at Q^k
  - For an appropriately selected sequence of updates, this is the same as regular value iteration
- Theorem: if the sequence samples each triple infinitely often then $Q^k \to Q^*$

Real-Time Value Iteration

- Asynchronous value iteration for a particular sequence of updates
- Episodic "learning"
  - Updates require knowledge of
- Start with some Q^0
- For $\ell = 0, 1, 2$
  - For $t = 0, 1, \ldots, H$
    - Select "greedy" action $a_t^\ell \in \arg\min_{a \in \mathcal{A}} Q_t^\ell(s_t^\ell, a)$
    - Update value at (t, s_t^ℓ, a_t^ℓ)
    - Sample next state $s_{t+1}^\ell \sim P_{t,a_t^\ell}(\cdot | s_t^\ell)$
- "greedy policy"
  - Does not necessarily sample all triples
- Theorem: if $Q^0 \geq Q^*$ then actions are optimal after some (random) finite time

DP Operators

- DP operators

$$(T_H V_{t+1})(s) = \max_{a \in \mathcal{A}} \left(\overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) V_{t+1}(s') \right)$$

$$(F_t Q_{t+1})(s,a) = \overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) \max_{a' \in \mathcal{A}} Q_{t+1}(s',a')$$

- Value iteration

$$V_t^* = T_t V_{t+1}^*$$

$$Q_t^* = F_t Q_{t+1}^*$$

- Monotonicity

$$V_{t+1} \geq V'_{t+1} \qquad \Longrightarrow \qquad T_t V_{t+1} \geq T_t V'_{t+1}$$

$$Q_{t+1} \geq Q'_{t+1} \qquad \Longrightarrow \qquad F_t Q_{t+1} \geq F_t Q'_{t+1}$$