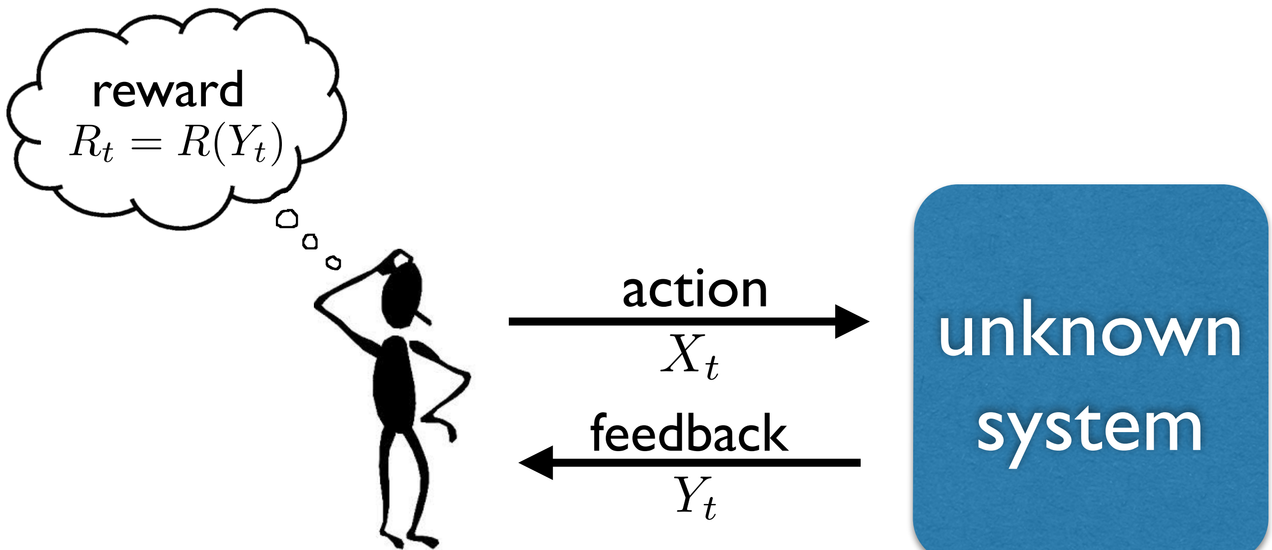
Review of Online Optimization

- Online optimization



- Special case: bandit optimization

- UCB
  - Construct set of statistically plausible models
  - Optimize optimistically

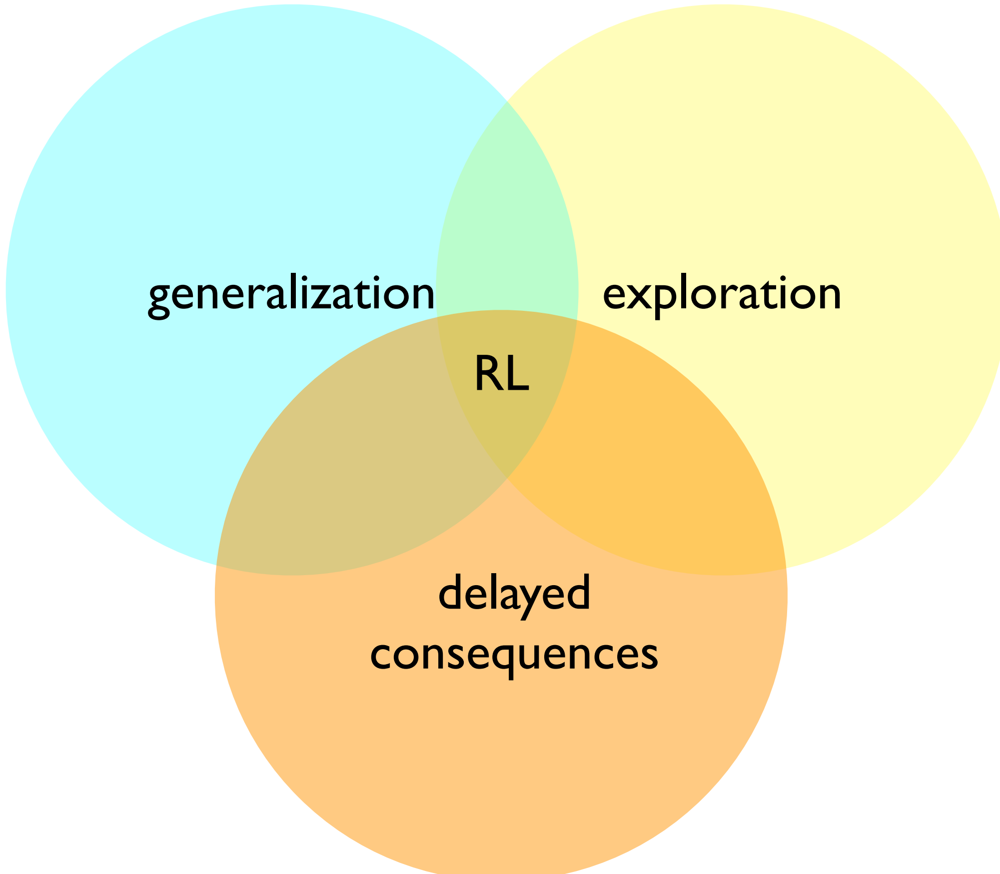
- Thompson Sampling
  - Sample from posterior model distribution
  - Optimize sampled model

- TS ≅ randomized approximation of UCB

Time-Varying Action Sets

- Actions sets $\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \cdots$
  - Each observed immediately before decision
- UCB and TS
  - Can be applied with time-varying action sets
  - Simply maximize over current action set
  - Regret analysis extends
- Contextual models
  - Think of meta-action as context-action pair
- Adaptive adversaries
  - Think of meta-action as adversary-self action pair
- Cautious learning
  - Restrict to conservative actions

Facets of Statistical Learning



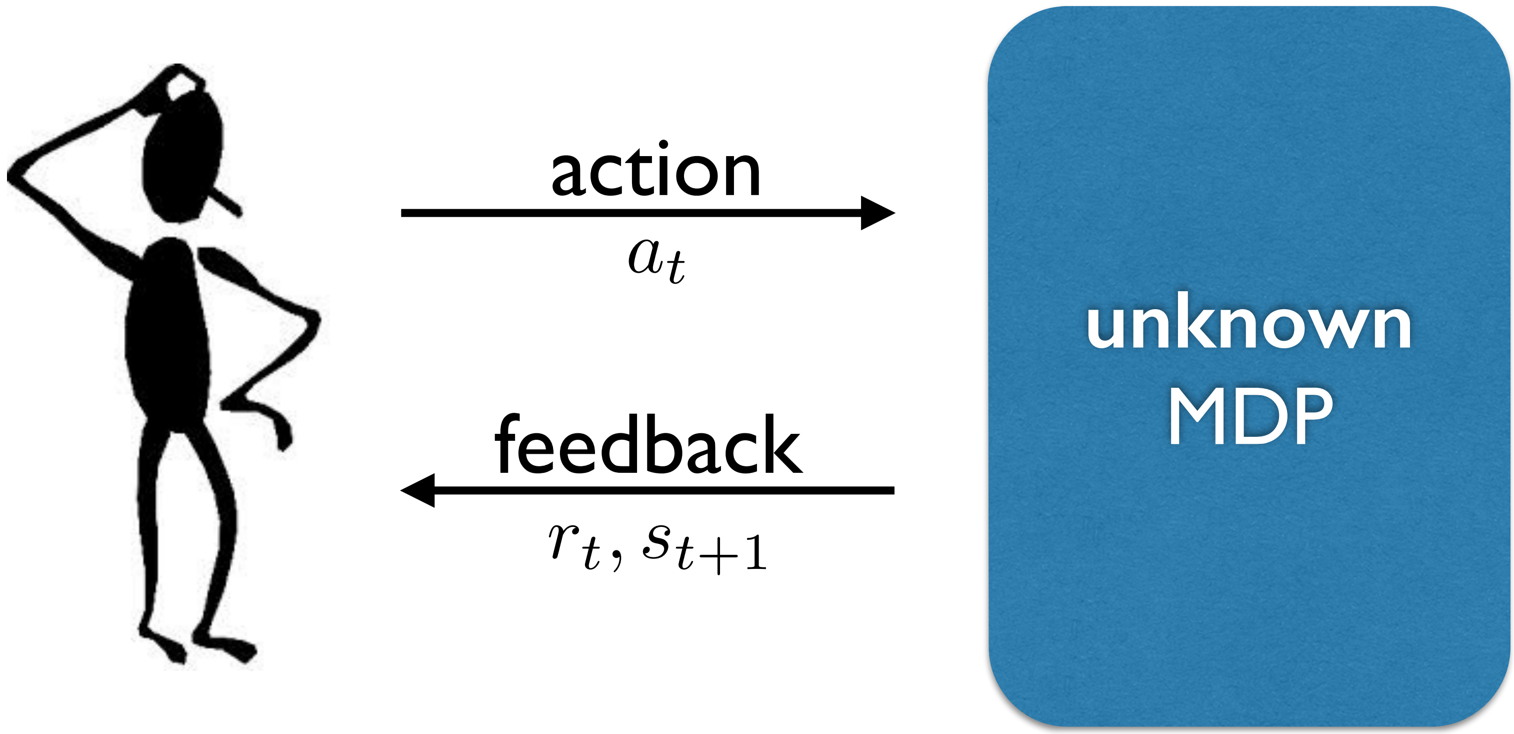
- Generalization
  - How to predict what we haven't seen from what we have seen?
  - Supervised and unsupervised learning
  - Regression and classification

- Exploration
  - How to act when actions influence observations?
  - Online learning and MABs
  - UCB, Thompson sampling, etc.

- Delayed consequences
  - How to assign credit to past actions?
  - Reinforcement learning

- How to effectively combine these remains open
  - For many methods, regret can be exponential in horizon or # states

Reinforcement Learning



- Current state and action influence
  - Immediate reward
  - State transition

- Delayed consequences via state dynamics

- Do delayed consequences matter?
  - Robotics
  - Web site content optimization
  - Online education
  - Medical treatments

Markov Decision Processes

- Start with a simple case
  - finite horizon
  - finite state and action sets
  - time-inhomogeneous transition and reward distributions

- MDP: $(\mathcal{S}, \mathcal{A}, R, P, H, \rho)$

state space

action space

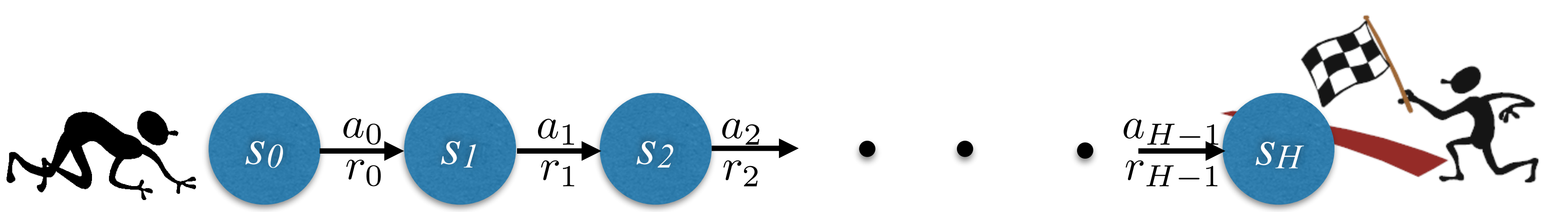
reward probs

transition probs

horizon

initial state distribution

- Sequence of H actions



- State and reward distributions

$$s_0 \sim \rho(\cdot) \qquad s_{t+1} \sim P_{t,a_t}(\cdot|s_t)$$

$$r_t \sim R_{t,a_t}(\cdot|s_t)$$

Policies and Value Functions

- (Deterministic) policy

$$\mu = (\mu_0, \ldots, \mu_{H-1}) \qquad a_t = \mu_t(s_t)$$

- Objective $\max_{\mu} \mathbb{E}\left[\sum_{t=0}^{H-1} r_t \,\middle|\, a_t = \mu_t(s_t)\right]$
- Value functions

$$V_t^{\mu}(s) = \mathbb{E}\left[\sum_{\tau=t}^{H-1} r_\tau \,\middle|\, a_\tau = \mu_\tau(s_\tau), s_t = s\right]$$

- Optimal value function

$$V_t^*(s) = \max_{\mu} V_t^{\mu}(s)$$

- A policy is optimal iff

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

- State-action value functions

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

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

- Stochastic policies

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

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