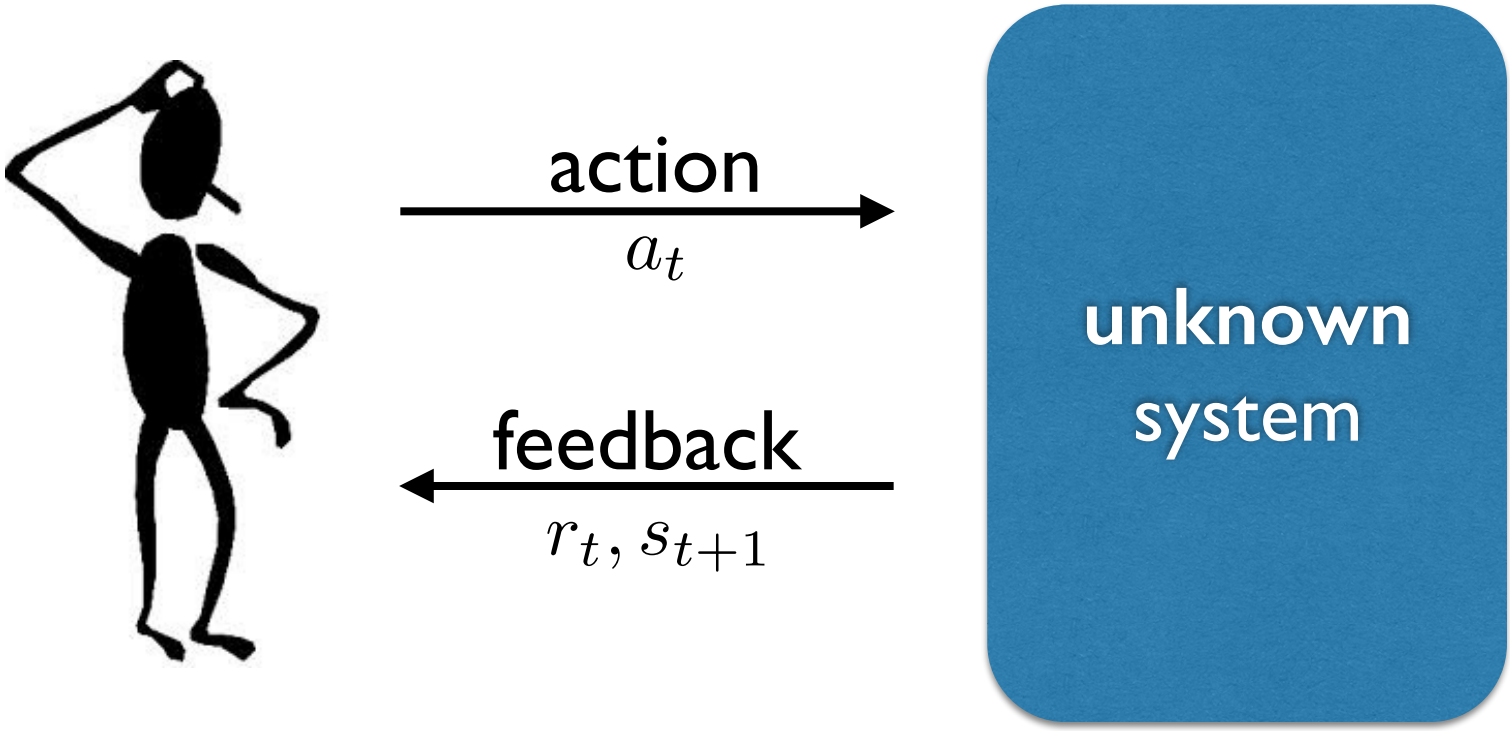
Reinforcement Learning

- The reinforcement learning problem



- Key issues
  - Exploration
  - Generalization
  - Delayed consequences

- Examples
  - What should the robot do next?
  - Which search results to display?
  - What medical treatments to prescribe?
  - What online course content to provide?
  - What orders to place to acquire a portfolio position?

- RL vs. approximate dynamic programming

Course Information

- MS&E 338 Reinforcement Learning

- Topics covered
  - Bandit optimization with generalization
  - Optimistic exploration, UCB, Thompson sampling
  - Reinforcement learning in Markov decision processes
  - Real-time dynamic programming and Q-learning
  - Optimistic exploration in MDPs
  - Generalization in reinforcement learning
  - Synthesis of generalization and optimistic exploration
  - Exploration beyond optimism

- Prerequisites
  - Optimization (e.g., MS&E 310)
  - Stochastic processes (e.g., MS&E 321)

- Requirements
  - Attend course lectures and related talks
  - Present one paper that complements course content
  - Final project presented in ten-page paper

- Instructors
  - Benjamin Van Roy, bvr@stanford.edu, Packard 273
  - Ian Osband, iosband@stanford.edu, Packard 274

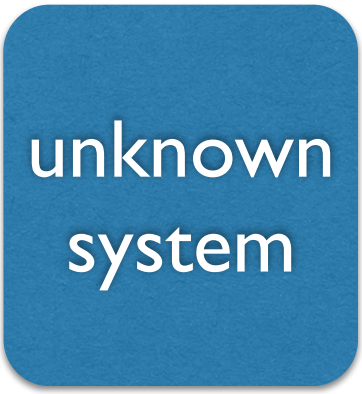
Course Schedule

Date	Time	Notes
January 5	14:15-15:45	
January 7	14:15-15:45	
January 12	14:15-15:45	Ian Osband
January 14	14:15-15:45	Ian Osband
January 21	14:15-15:45	
January 22		Thursday, time/room TBD
January 23	14:15-15:45	Friday, time/usual room
January 26	14:15-15:45	
January 27		Tuesday, time/room TBD
January 28	14:15-15:45	
February 2	14:15-15:45	Ian Osband
February 4	14:15-15:45	Ian Osband
February 9	14:15-15:45	Mohammad Ghavamzadeh
February 11	14:15-15:45	Mohammad Ghavamzadeh
February 18	14:15-15:45	
February 20	14:15-15:45	Friday, usual room
February 23	14:15-15:45	
February 25	14:15-15:45	
February 27	14:15-15:45	Friday, usual room
March 2	14:15-15:45	
March 4	14:15-15:45	

FIRST TALK TO ATTEND
Speaker: Dan Russo
January 7, 12:00-13:15
Knight Management Center P107

Single-Period Optimization

- RL without delayed consequences



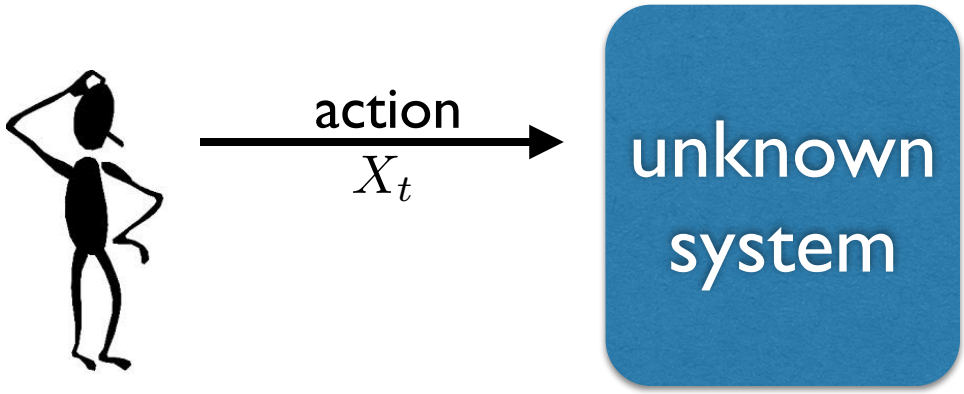
- Optimization problem

- Objective = expected reward

- Learn
  - about θ
  - from $\ldots, X_{t-3}, Y_{t-3}, X_{t-2}, Y_{t-2}, X_{t-1}, Y_{t-1}$

Single-Period Optimization

- RL without delayed consequences



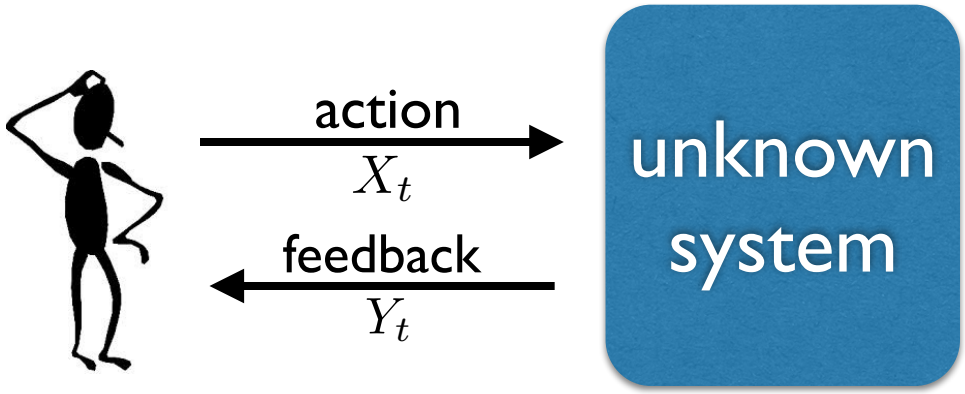
- Optimization problem

- Objective = expected reward

- Learn
  - about θ
  - from $\ldots, X_{t-3}, Y_{t-3}, X_{t-2}, Y_{t-2}, X_{t-1}, Y_{t-1}$

Single-Period Optimization

- RL without delayed consequences



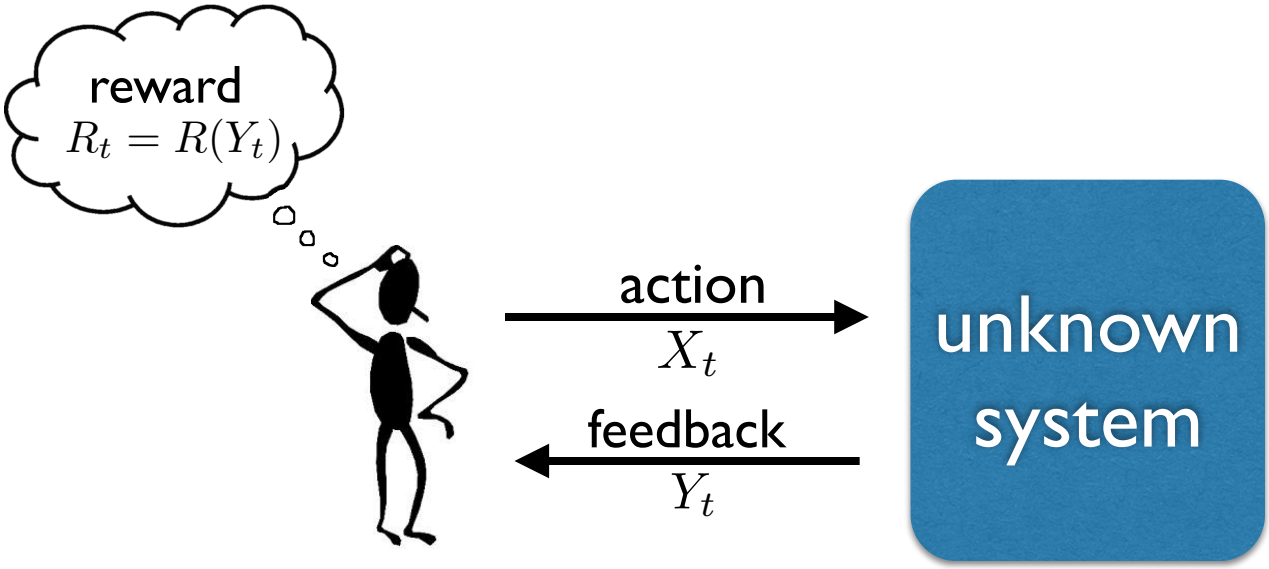
- Optimization problem

- Objective = expected reward

- Learn
  - about θ
  - from $\ldots, X_{t-3}, Y_{t-3}, X_{t-2}, Y_{t-2}, X_{t-1}, Y_{t-1}$

Single-Period Optimization

- RL without delayed consequences



- Optimization problem

$$\max_{x \in \mathbb{X}} f_\theta(x) \qquad \theta \in \Theta$$

- Objective = expected reward

$$f_\theta(X_t) = E\left[R_t | X_t, \theta\right]$$

- Learn
  - about θ
  - from $\ldots, X_{t-3}, Y_{t-3}, X_{t-2}, Y_{t-2}, X_{t-1}, Y_{t-1}$

Bayesian Formulation

- Prior probability PDF $p_0(\theta)$
- Learning $p_t(\theta) = \mathbb{P}\left(\theta \mid X_1^t, Y_1^t\right)$
- Bayes rule

$$p_{t+1}(\theta) = \frac{\mathbb{P}\left(Y_{t+1} \mid \theta, X_{t+1}\right) p_t(\theta)}{\mathbb{P}\left(Y_{t+1} \mid p_t, X_{t+1}\right)}$$

- Learning algorithm $X_t = \mathcal{A}\left(p_0, X_1^{t-1}, Y_1^{t-1}, Z_1^t\right)$
  - Should suffice to have $X_t = \mathcal{A}(p_{t-1}, Z_t)$
- Finite-horizon expected reward

$$\max_{\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

- Exploration versus exploitation

Bayesian Formulation

- Prior probability PDF $p_0(\theta)$
- Learning $p_t(\theta) = \mathbb{P}\left(\theta \mid X_1^t, Y_1^t\right)$
- Bayes rule

$$p_{t+1}(\theta) = \frac{\mathbb{P}\left(Y_{t+1} \mid \theta, X_{t+1}\right) p_t(\theta)}{\mathbb{P}\left(Y_{t+1} \mid p_t, X_{t+1}\right)}$$

- Learning algorithm $X_t = \mathcal{A}\left(p_0, X_1^{t-1}, Y_1^{t-1}, Z_1^t\right)$

  - Should suffice to have $X_t = \mathcal{A}(p_{t-1}, Z_t)$
- Finite-horizon expected reward

$$\max_{\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$$

- Exploration versus exploitation

Example: Beta-Bernoulli

- Finite action space $X_t \in \{1, \ldots, N\}$

- Binary reward feedback $Y_t = R_t \in \{0, 1\}$

$$\mathbb{E}\left[R_t | \theta, X_t\right] = \theta_{X_t}$$

- Independent Beta prior PDF

$$p_0(\theta) = \prod_{n=1}^{N} \frac{1}{\mathcal{B}(\alpha_{n0}, \beta_{n0})} \theta_n^{\alpha_{n0}-1} (1-\theta_n)^{\beta_{n0}-1}$$

- Conjugate distribution

$$p_t(\theta) = \prod_{n=1}^{N} \frac{1}{\mathcal{B}(\alpha_{nt}, \beta_{nt})} \theta_n^{\alpha_{nt}-1} (1-\theta_n)^{\beta_{nt}-1}$$

- Bayes rule

$$\alpha_{X_{t+1},t+1} = \alpha_{X_{t+1},t} + R_{t+1}$$

$$\beta_{X_{t+1},t+1} = \beta_{X_{t+1},t} + (1 - R_{t+1})$$

$$\alpha_{n,t+1} = \alpha_{n,t} \qquad \forall n \neq X_{t+1}$$

$$\beta_{n,t+1} = \beta_{n,t} \qquad \forall n \neq X_{t+1}$$

Example: Independent Gaussian

- Finite action space $X_t \in \{1, \ldots, N\}$
- Noisy reward $Y_t = R_t \sim \mathcal{N}\left(\theta_{X_t}, \sigma_r^2\right)$
- Independent Gaussian prior PDF

$$p_0(\theta) = \prod_{n=1}^{N} \frac{1}{\sigma_{n0}\sqrt{2\pi}} e^{-\frac{(\theta_n - \mu_{n0})^2}{2\sigma_{n0}^2}}$$

- Conjugate distribution

$$p_t(\theta) = \prod_{n=1}^{N} \frac{1}{\sigma_{nt}\sqrt{2\pi}} e^{-\frac{(\theta_n - \mu_{nt})^2}{2\sigma_{nt}^2}}$$

- Bayes rule

$$\mu_{X_{t+1},t+1} = \frac{\mu_{X_{t+1},t}/\sigma^2_{X_{t+1},t} + R_{t+1}/\sigma_r^2}{1/\sigma^2_{X_{t+1},t} + 1/\sigma_r^2}$$

$$\sigma^2_{X_{t+1},t+1} = \frac{1}{1/\sigma^2_{X_{t+1},t} + 1/\sigma_r^2}$$

Example: Linear-Gaussian

- Action space $X_t \in \mathcal{X} \subset \Re^N$
- Noisy reward $Y_t = R_t \sim \mathcal{N}\left(\theta^\top X_t, \sigma_r^2\right)$
- Gaussian prior PDF $p_0(\theta) \sim \mathcal{N}(\mu_0, \Sigma_0)$
- Conjugate distribution $p_t(\theta) \sim \mathcal{N}(\mu_t, \Sigma_t)$
- Bayes rule

$$\mu_{t+1} = \mu_t + \frac{\Sigma_t X_{t+1}(R_{t+1} - \mu_t^\top X_{t+1})}{X_{t+1}^\top \Sigma_t X_{t+1} + \sigma_r^2}$$

$$\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t X_{t+1} X_{t+1}^\top \Sigma_t}{X_{t+1}^\top \Sigma_t X_{t+1} + \sigma_r^2}$$

- Special case: linear program
- Relation to supervised learning