P9120 - Statistical Learning and Data Mining

Lecture 9 - Reinforcement Learning I

Min Qian

Department of Biostatistics, Columbia University

October 31st, 2024

Outline

Causal Inference

- Notations: (X, A, R)
  - X: pre-treatment information,
  - $A \in \{0, 1\}$: treatment,
  - $R \in \mathbb{R}$: outcome.

- Potential outcomes:
  - $R(0)$: outcome that would have been observed under treatment 0.
  - $R(1)$: outcome that would have been observed under treatment 1.

- Causal Estimands:
  - Average Treatment Effect

$$\text{ATE} = E[R(1) - R(0)].$$

  - Conditional Average Treatment Effect

$$\text{CATE}(x) = E[R(1) - R(0)|X = x].$$

Causal Inference Assumptions

- Data:

X_i	10	4	13	5	...	7	9	4	6
A_i	1	1	1	1	...	0	0	0	0
R_i	-1	1	6	2	...	-4	4	1	2
$R_i(0)$	?	?	?	?	...	-4	4	1	2
$R_i(1)$	-1	1	6	2	...	?	?	?	?

- How can we use data to estimate ATE and CATE?

- Three key assumptions:
  - SUTVA (stable unit treatment value assumption): $R_i = R_i(A_i)$;
  - Positivity: $0 < P(A_i = 1 | X_i = x) < 1$;
  - No unmeasured confounders: $A_i \perp\!\!\!\perp R_i(0), R_i(1) | X$.

The Effect of Confounders: Simpson's Paradox

Kidney Stone Treatment Example (Charig et al. 1986)

- Treatment $A = 1$ for an open surgical procedure, and $A = 0$ for a small puncture procedure.
- Outcome $R = 1$ for success and $R = 0$ for failure.

	whole sample $n = 700$		small stone group $n = 357$		large stone group $n = 343$	
	$R = 1$	$R = 0$	$R = 1$	$R = 0$	$R = 1$	$R = 0$
$A = 1$	273	77	81	6	192	71
(success rate)	78%		93%		73%	
$A = 0$	289	61	234	36	55	25
(success rate)	83%		87%		69%	

- In the whole sample, treatment 0 has higher success rate than 1.
- In each subgroup, treatment 1 has higher success rate than 0.

Making Causal Inference from Data

- Three key assumptions are satisfied in randomized trials.
- Under the assumptions

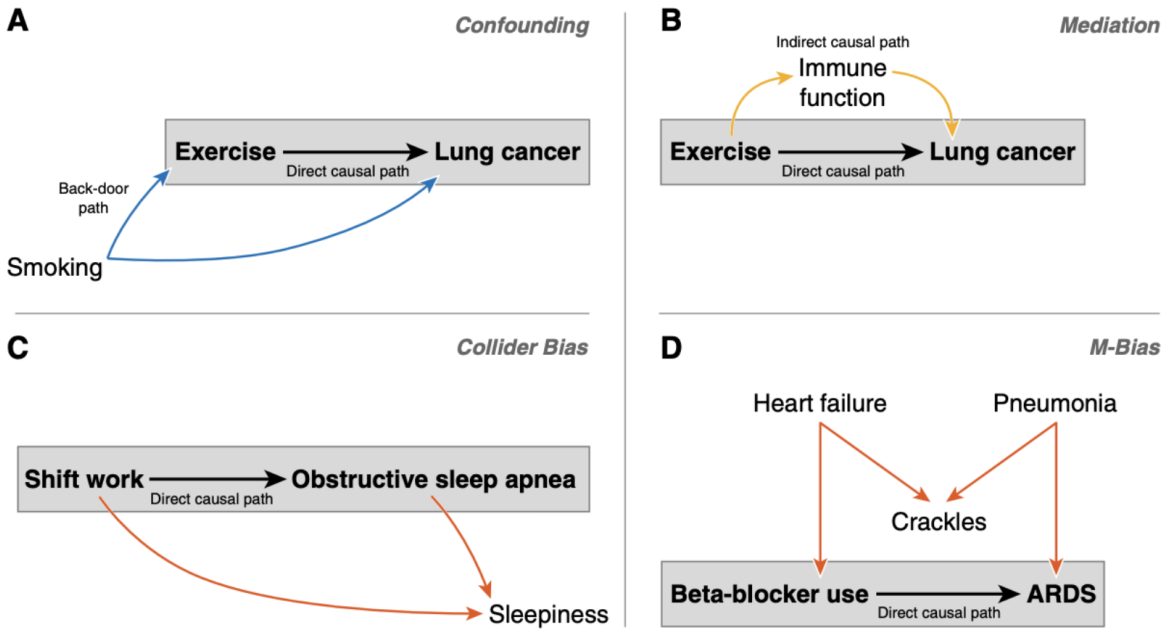
$$\begin{aligned} E[R(a)|X=x] =& E[R(a)|X=x, A=a] \\ =& E[R|A=a, X=x] \triangleq Q^*(x,a) \text{ for } a=0,1. \end{aligned}$$

  - $Q^*(x,a)$ can be estimated from the observed data.
  - $\text{CATE}(x) = Q^*(x,1) - Q^*(x,0)$,
    $\text{ATE} = \int [Q^*(x,1) - Q^*(x,0)]p(x)dx$.

Research Questions in Classic Causal Inference

Focus on making causal inference from observational studies.

- How to estimate ATE / CATE?
  - Assumptions: validation, violation, etc.
  - Model: robustness to model misspecification, efficiency, etc.
- How to learn the causal structure?
- How to estimate causal mediation effects?
- ...



From Causal Inference to Decision Making

Decision to make: what is the optimal treatment for each patient?

- Assume large values of outcome R is preferred.
- Optimal treatment for $X = x$ can be obtained using decision rule:

$$\begin{aligned}\pi^*(x) &= \arg\max_{a\in\{0,1\}} E[R(a)|X=x] \\ &= \begin{cases} 1 & \text{if CATE}(x) = E[R(1)-R(0)|X=x] \geq 0; \\ 0 & \text{if CATE}(x) < 0. \end{cases}\end{aligned}$$

$\pi^*(x)$ is known as the optimal policy in RL literature.

Single Stage Policy Learning

- Focus on learning optimal policy π^* from data $(X_i, A_i, R_i), i = 1, \ldots, n$.

- Usually assume all three key assumptions hold (e.g. data from randomized trial), and the causal structure is known. In this case,

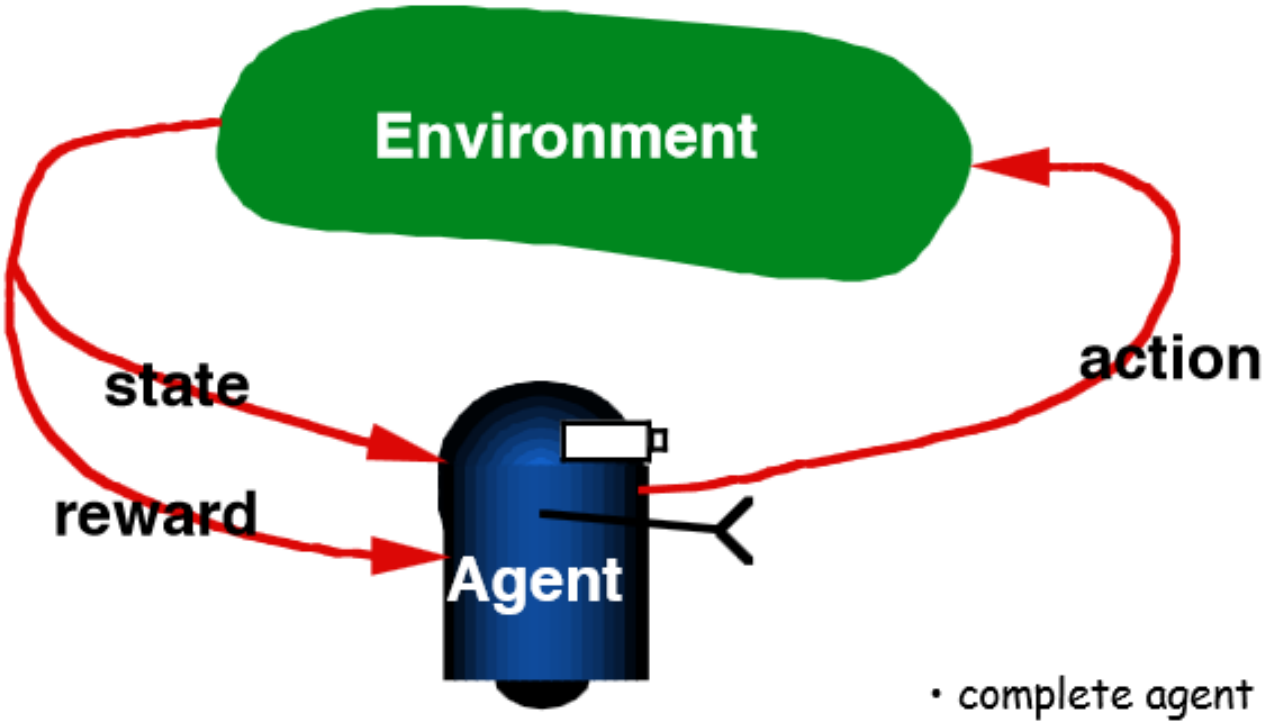
$$E[Y(a)|X = x] = E[R|A = a, X = x] \triangleq Q^*(x, a).$$

  - Estimate $Q^*(x, a)$ using supervised learning methods,
  - Using inverse probability weighting method,
  - ...

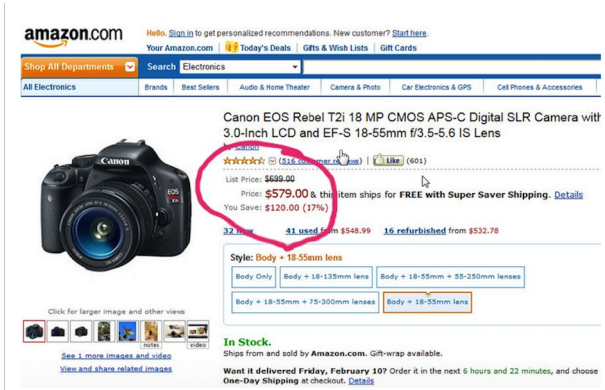
Reinforcement Learning: Sequential Decision Making

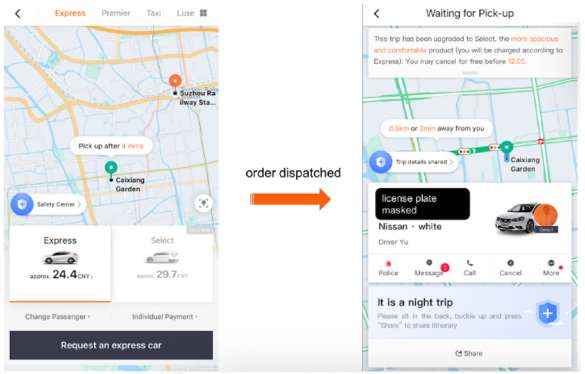
$$\{X_1, A_1, R_1, \ldots, X_t, A_t, R_t, \ldots\}$$

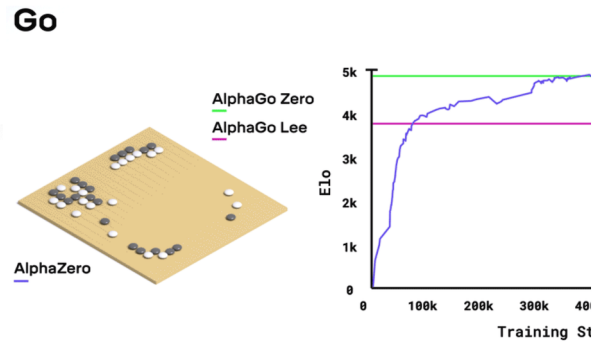
Learn policy $\{\pi_t : \mathcal{X}_t \to \mathcal{A}_t, t = 1, 2, \ldots\}$ to assign treatment based on history to optimize the expected cumulative outcomes/rewards $E\big[\sum_{t=1}^{\infty} \gamma^{t-1} R_t(\pi_t)\big]$ or $E\big[\sum_{t=1}^{T} R_t(\pi_t)\big]$

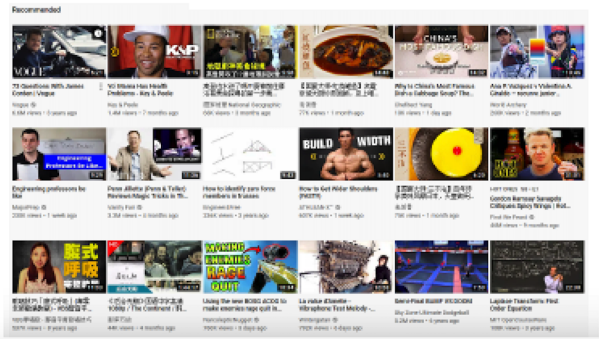


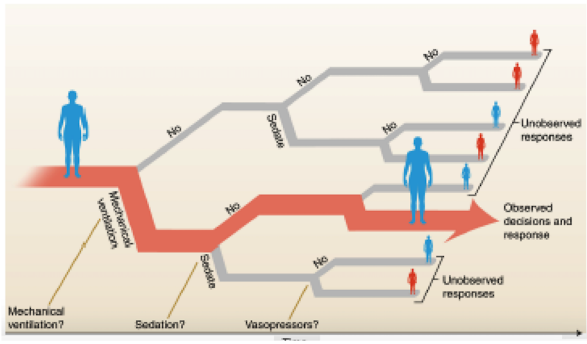
RL Applications





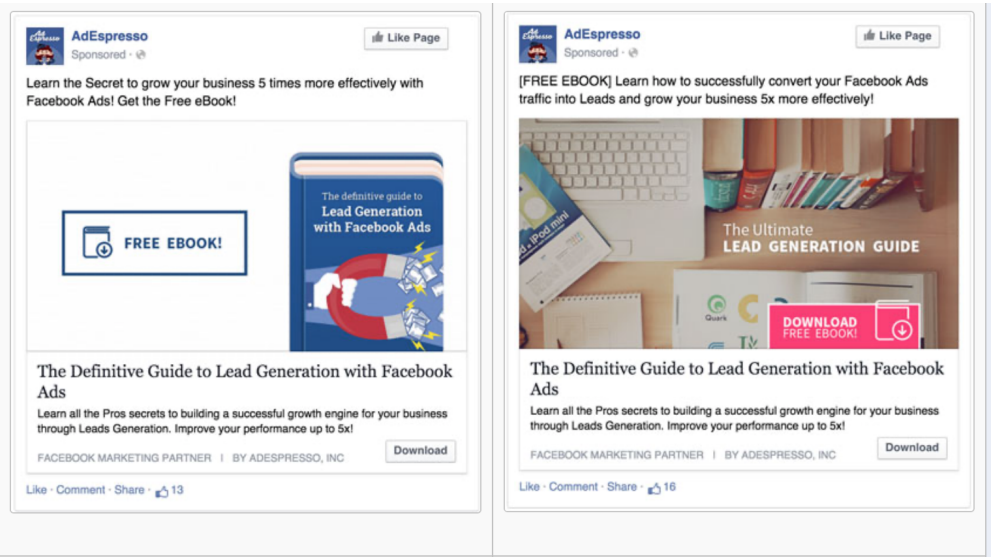






- Information and rewards
- learning and optimization
- trial and error

A/B Testing



10,000 Impression	10,000 Impression
237 Clicks *(CTR: 2.37%)*	187 Clicks *(CTR: 1.87%)*
28 Sales *(Conversion rate: 11.81%)*	16 Sales *(Conversion rate: 8.55%)*
Spent: $150	Spent: $150
Cost per Sale: $5.35	Cost per Sale: $9.37 (+75.14%)

- A/B testing is a simple RCT that involves two variants (A and B).
- The goal is to figure out which one (A or B) performs better.
- Can be generalized to more than two arms.

Multi-Armed Bandit

- K possible options/actions (e.g., K slot machines).
- The payout of a slot machine follows a fixed but unknown distribution.
- You could play, say, 1000 of times. Each time you need to decide which slot machine to play.
- How to play so as to maximize the total payout?

Multi-Armed Bandit Formulation

- You are repeatedly faced a choice of K actions $a \in \{1, 2, \ldots, K\}$.
- At each time point t, the reward following action a, denoted by $R_t(a)$, is randomly drawn from an unknown distribution $D(a)$ with expectation $Q^*(a)$, which only depends on a.
- Goal: maximize the expected total reward, $\sum_{t=1}^{T} E[R_t(a_t)]$, over some time period (say, $T = 1000$ steps).

How to choose action a_t at each step?

Greedy action

Warm up: Randomly choose actions $a \in \{1, 2, \ldots, K\}$ initially (say, first 100 times).

After that, if at the t-th play, action a has been chosen $N_t(a)$ times prior to t, yielding rewards $R_1, R_2, \ldots, R_{N_t(a)}$, then estimate $Q^*(a)$ by

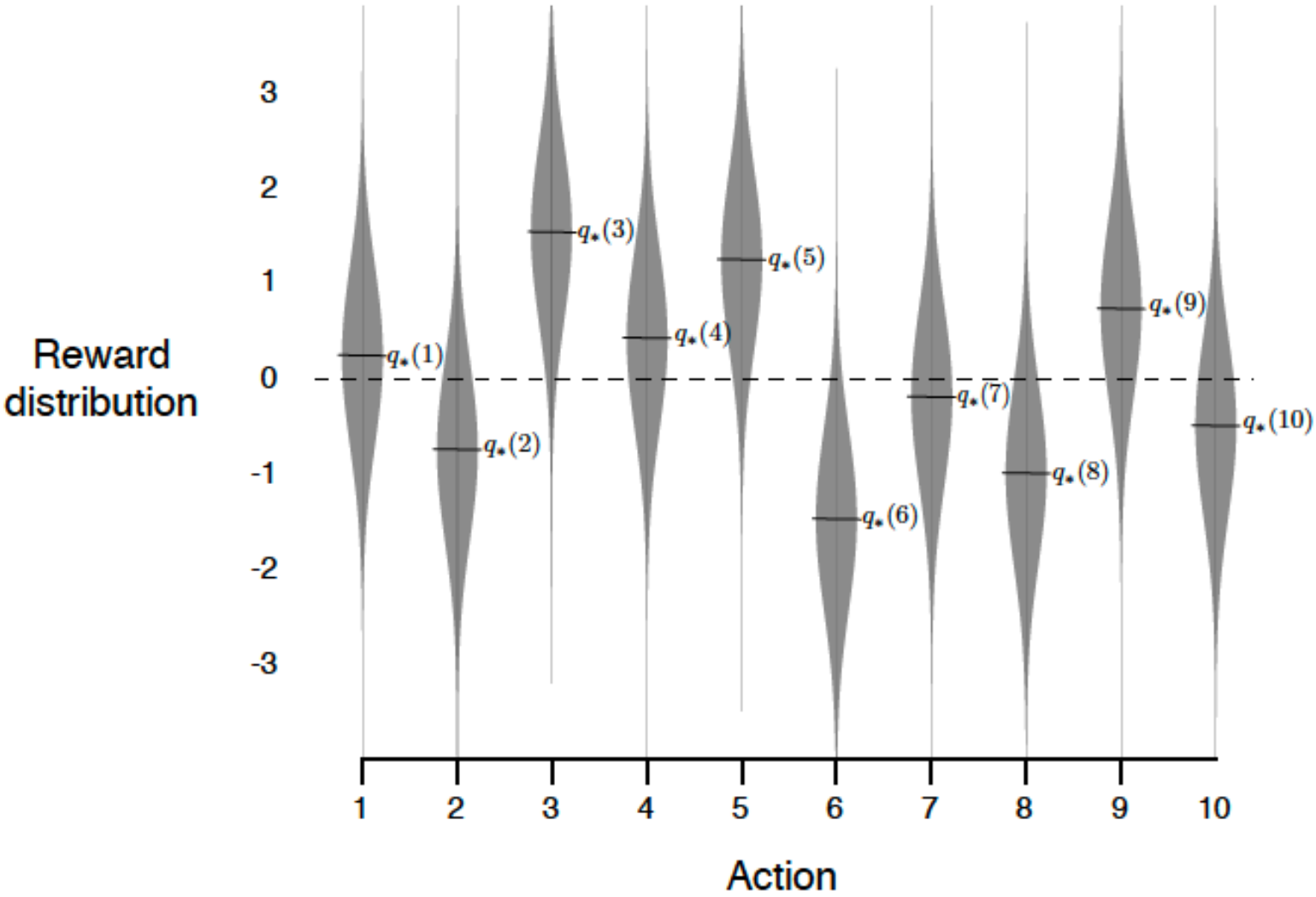
$$\widehat{Q}_t(a) = \frac{R_1 + R_2 + \ldots + R_{N_t(a)}}{N_t(a)}.$$

By LLN, $\widehat{Q}_t(a) \to Q^*(a)$ as $N_t(a) \to \infty$.

Choose the greedy action: $a \in \arg\max_a Q_t(a)$.

- This method always exploits current knowledge to maximize immediate reward.
- Maybe explore new actions in order to maker better action selections in the future.

A 10-armed Bandits Example



In all methods below, there is often a warm-up period.

ϵ-Greedy Method

Pick the greedy action w.p. $1 - \epsilon$, and pick a random action w.p. ϵ.

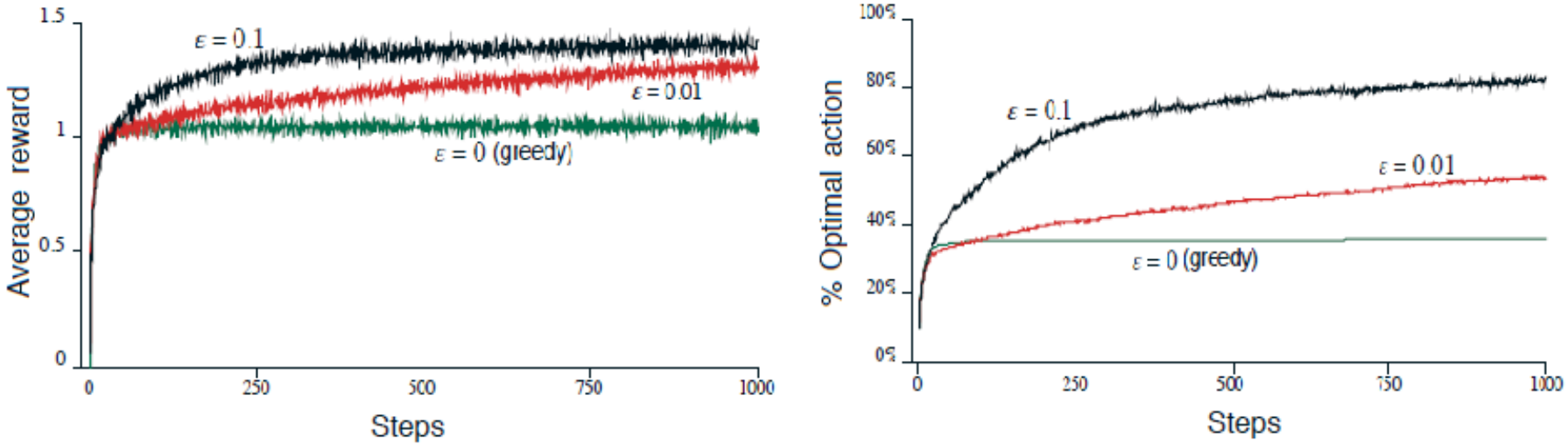
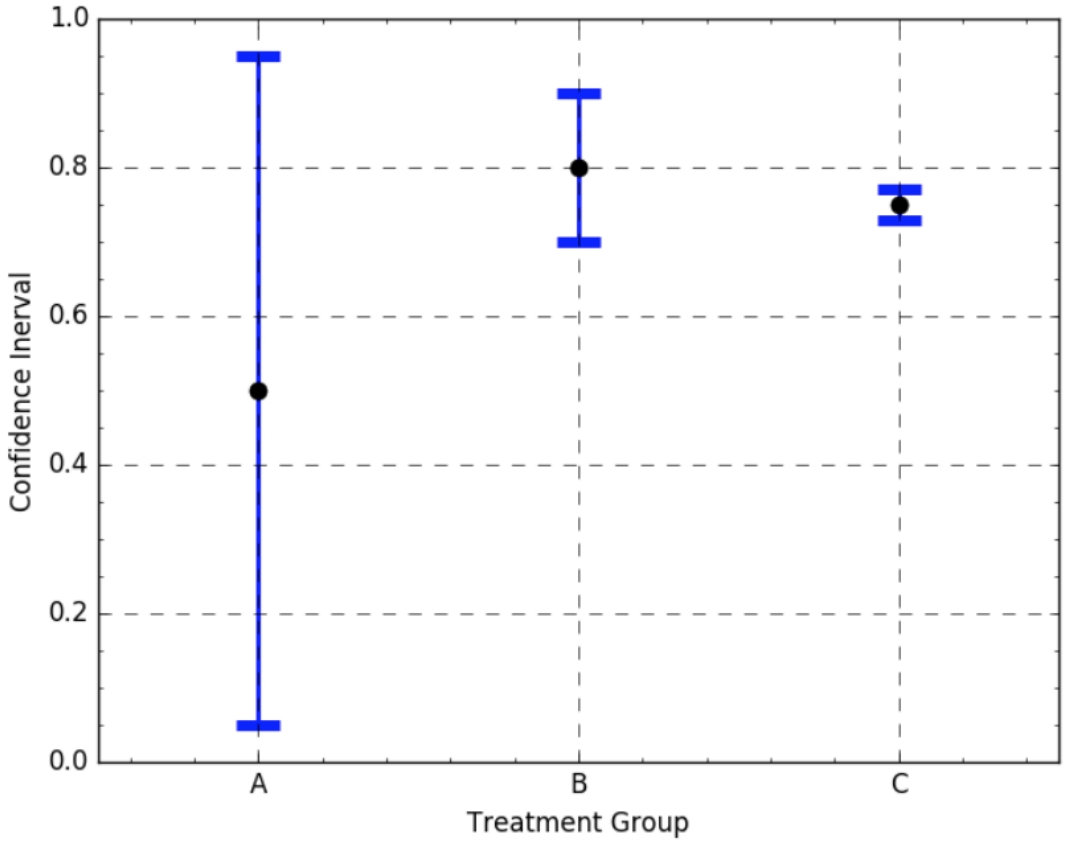


Figure: Average performance over 2000 runs.

- ϵ too small, learn too slow; ϵ too large, explore too much.
- Often start with large ϵ and then decrease over time.

Upper-Confidence-Bound Method (UCB)

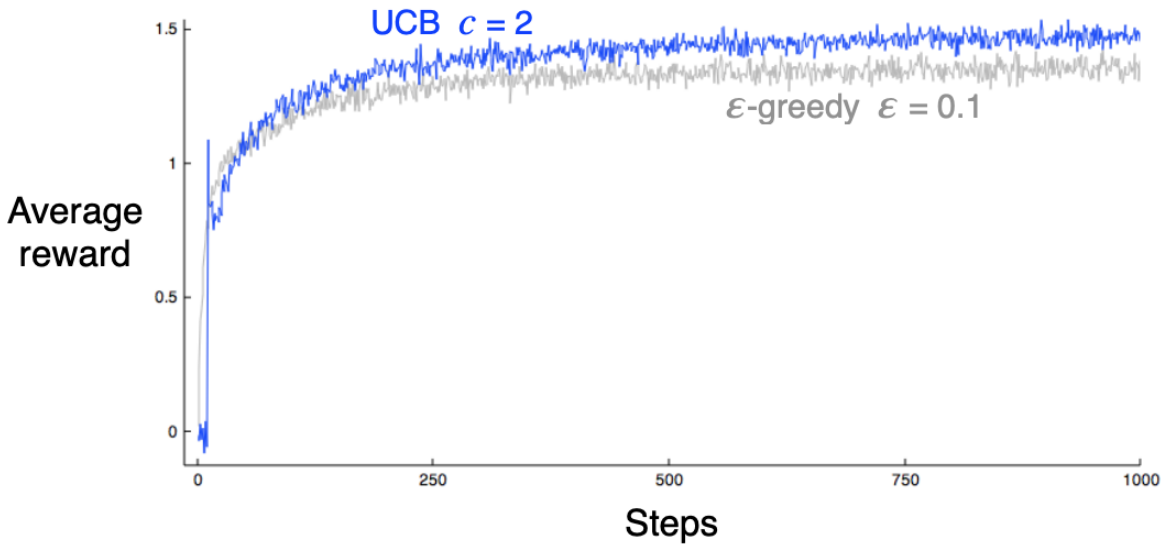
Idea: select actions according to their potential for actually being optimal - based on both $\widehat{Q}_t(a)$ and the uncertainty in $\widehat{Q}_t(a)$.



w.p. $\geq 1 - t^{-2c^2}$, $Q^*(a) \leq \widehat{Q}_t(a) + c[\log t / N_t(a)]^{1/2}$.

UCB vs ϵ-Greedy

At each time point t, select $a_t = \arg\max_a \{\widehat{Q}_t(a) + c[\log t / N_t(a)]^{1/2}\}$.



- UCB generally performs better that ϵ-greedy action selection after the first several plays.
- ϵ-greedy has no preference for those that are nearly greedy or particularly uncertain.
- UCB is more difficult to generalize to general RL settings than ϵ-greedy.

Boltzmann Exploration (Softmax) Method

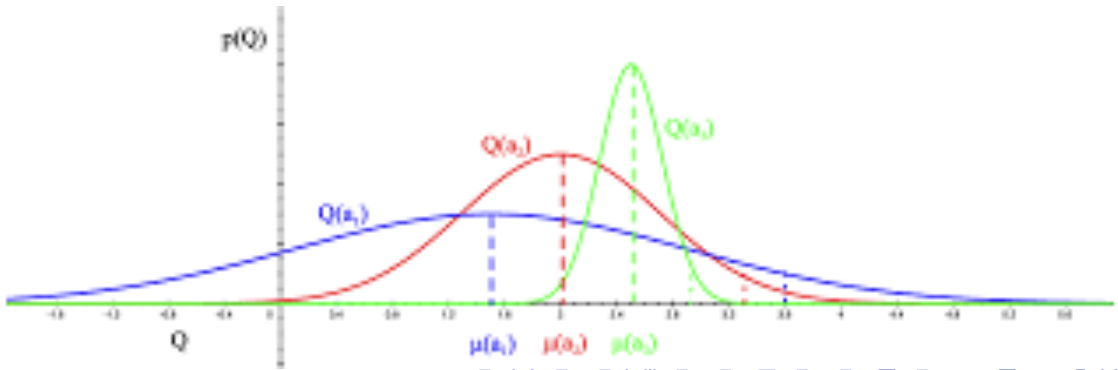
- Choose an action with probability that is proportional to its average reward at each time point .
- At time point t, choose action a with probability

$$\pi_{t+1}(a) = \frac{\exp(\widehat{Q}_t(a)/\tau)}{\sum_{a'=1}^{K} \exp(\widehat{Q}_t(a')/\tau)}.$$

- τ is a scaling parameter. As $\tau \to \infty$, all actions are chosen uniformly.

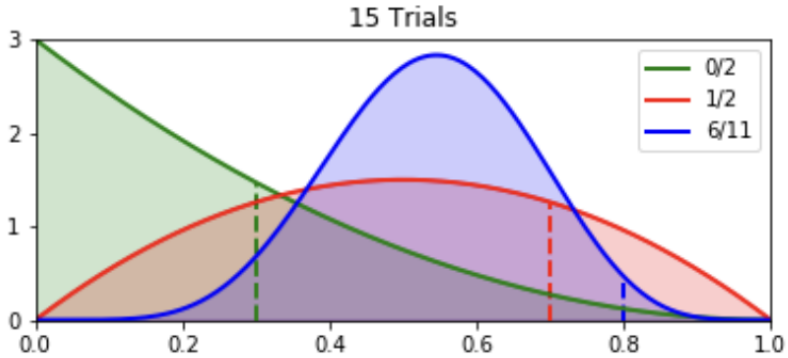
Thompson Sampling

1. Parameterize distribution of $R_t|a \sim P(r;\theta_a)$ for all $a = 1, \ldots, K$.
2. Set the prior of $\theta_a \sim P(\theta_a)$ for all a.
3. At each step, if an arm a is selected, then update the posterior distribution of parameter θ_a, $P(\theta_a|Data)$.
4. Choose the action $\widehat{A}_t = k$ with prob. $P(\arg\max_a E(R_t|a,\theta_a) = k)$, where P is taken w.r.t. the posterior distribution of θ_a's.
   - Equivalent to sample $\widehat{\theta}_a$ from $P(\theta_a|Data)$ and choose $\widehat{A}_t = \arg\max_a E(R_t|a;\widehat{\theta}_a)$.
   - Posterior mode enables exploitation
   - Posterior spread/variance enables exploration.



Thompson Sampling: Binary Rewards $Y_t \in \{0, 1\}$

- $P(R_t = 1|A = a) = \theta(a)$.
- Prior $\theta(a) \sim Beta(\alpha_0(a), \beta_0(a))$ for $a = 1, \ldots, K$.



- At each step t:
  - Sample $\widehat{\theta}_t(a)$ from $Beta(\alpha_{t-1}(a), \beta_{t-1}(a))$ for $a = 1, \ldots, K$.
  - Select $\widehat{A}_t = \arg\max_{a=1,\ldots,K} \widehat{\theta}_t(a)$, observe R_t.
  - Update distribution of $\theta(a)$ to $Beta(\alpha_t(a), \beta_t(a))$ by setting

$$(\alpha_t(a), \beta_t(a)) = \begin{cases} (\alpha_{t-1}(a), \beta_{t-1}(a)) & \text{if } a \neq \widehat{A}_t \\ (\alpha_{t-1}(a) + R_t, \beta_{t-1}(a) + 1 - R_t) & \text{if } a = \widehat{A}_t. \end{cases}$$

Multi-Armed Bandits: Exploration vs Exploitation Trade-off

Goal: Select $\{a_t :, t = 1, \ldots, T\}$ to maximize the expected total reward $\sum_{t=1}^{T} E[R_t(a_t)]$ or minimize Regret $= T \max_a Q^*(a) - \sum_{t=1}^{T} E[R_t(a_t)]$.

- Exploitation: Make the best decision for current subject given current information.
- Exploration: Gather more information (benefit future subjects).

When to explore and when to exploit?

Connection to Clinical Trials: Interventions A vs B

A/B testing

- Analog to conventional randomized clinical trial
- Sample size determined by formal power analysis
- Aim to benefit future patients.
- The reward does not need to be immediate.
- Can analyze multiple outcome metrics.

Multi-armed bandits

- Analog to response adaptive randomization.
- Sample size determined by stopping rule.
- Benefit patients in the study as well as future patients.
- Can add/remove arms in the middle of an experiment.

Reminders

- Hw #3 is due at 9pm on Saturday November 16th.
- Quiz for Lecture 9 is due at 9pm on Monday November 4th.
- In class group paper presentation:
  - Each presentation is about 40-45 minutes.
  - Every student in the group is expected to present.
  - Evaluation will be based on both individual performance (80%) and group performance (20%).
  - Next week (Nov 7th): V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation.
    Yangyang Chen, Serena Hu, Ze Li, Ziqiu Liu, Qu Sha, Eunice Wang