*Note: we only went carefully through slides before slide 34. The remaining slides are kept for those interested but will not be material required for the quiz. See the last slide for a summary of what you should know
Select all that are true:

- Thompson sampling for MDPs the posterior over the dynamics can be updated after each transition.
- When using a Beta prior for a Bernoulli reward parameter for an \((s,a)\) pair, the posterior after \(N\) samples of that pair time steps can be the same as after \(N+2\) samples.
- The optimism bonuses discussed for MBIE-EB depend on the maximum reward but not on the maximum value function.
- In class we discussed adding a bonus term to the policy gradient update for a \((s,a,r,s')\) tuple using Q-learning with function approximation. Adding this bonus term will ensure all Q estimates used to make decisions online using DQN are optimistic with respect to \(Q^*\).
- Not sure.
Class Structure

• Last time: Fast Reinforcement Learning
• This time: Batch RL
• Next time: Guest Lecture
A Group

\[
\frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4} \\
\frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}
\]

1. Compare these fractions using the cross-multiplication strategy:
\[
\frac{4}{5} \quad \frac{9}{10}
\]

2. Finally, reduce the sum to lowest terms:
\[
\frac{4 \times 10}{40} = \frac{9 \times 5}{45}
\]

Avg Score: 95
A Scientific Experiment

Lecture 14: Batch RL

Winter 2020 Slides drawn from Philip Thomas with modifications

A Group

\[ \frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4} \]

2. Finally, reduce the sum to lowest terms:

\[ \frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20} \]

Avg Score: 95

B Group

1. Compare these fractions using the cross-multiplication strategy.

\[ \frac{4}{5} \times \frac{9}{10} = \frac{36}{50} \]

2. Finally, reduce the sum to lowest terms:

\[ \frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20} \]

Avg Score: 92
What Should We Do For a New Student?

A Group

1. Compare these fractions using the cross-multiplication strategy.
   \[
   \frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}
   \]

2. Finally, reduce the sum to lowest terms:
   \[
   \frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}
   \]

B Group

1. Compare these fractions using the cross-multiplication strategy.
   \[
   \frac{4}{3} \quad \frac{9}{10}
   \]

2. Finally, reduce the sum to lowest terms:
   \[
   \frac{4}{10} + \frac{3}{5} = \frac{19}{20} = \frac{19}{20}
   \]

Avg Score: 95

Avg Score: 92
Involves Counterfactual Reasoning

A Group

Avg Score: 95

B Group

Avg Score: 92

B Group

???
Involves Generalization

A Group

B Group

Avg Score: 95

Avg Score: 92

B Group

???
Batch Reinforcement Learning

A Group

1. Compare these fractions using the cross-multiplication strategy.

2. Finally, reduce the sum to lowest terms:

   \[
   \frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}
   \]

   \[
   \frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}
   \]

Avg Score: 95

B Group

1. Compare these fractions using the cross-multiplication strategy.

2. Finally, reduce the sum to lowest terms:

   \[
   \frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}
   \]

   \[
   \frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}
   \]

Avg Score: 92

B Group

1. Compare these fractions using the cross-multiplication strategy.

2. Finally, reduce the sum to lowest terms:

   \[
   \frac{3}{6} + \frac{2}{8} = \frac{18}{24} = \frac{3}{4}
   \]

   \[
   \frac{2}{10} + \frac{3}{4} = \frac{19}{20} = \frac{19}{20}
   \]

???
The Problem

- If you apply an existing method, do you have confidence that it will work?
A property of many real applications

- Deploying "bad" policies can be costly or dangerous
What property should a safe batch reinforcement learning algorithm have?

- Given past experience from current policy/policies, produce a new policy
  - “Guarantee that with probability at least $1 - \delta$, will not change your policy to one that is worse than the current policy.”
- You get to choose $\delta$
- Guarantee not contingent on the tuning of any hyperparameters
1. Notation

2. Create a safe batch reinforcement learning algorithm
   - Off-policy policy evaluation (OPE)
   - Safe policy improvement (SPI)
Notation

- Policy $\pi$: $\pi(a \mid s) = P(a_t = a \mid s_t = s)$
- Trajectory: $T = (s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_L, a_L, r_L)$
- Historical data: $D = \{T_1, T_2, \cdots, T_n\}$
- Historical data from behavior policy, $\pi_b$
- Objective:
  \[ V^\pi = \mathbb{E}[\sum_{t=1}^{L} \gamma^t R_t \mid \pi] \]
Safe batch reinforcement learning algorithm

- Reinforcement learning algorithm, $\mathcal{A}$
- Historical data, $D$, which is a random variable
- Policy produced by the algorithm, $\mathcal{A}(D)$, which is a random variable
- A safe batch reinforcement learning algorithm, $\mathcal{A}$, satisfies:

$$\Pr(V^{\mathcal{A}(D)} \geq V^{\pi_b}) \geq 1 - \delta$$

or, in general

$$\Pr(V^{\mathcal{A}(D)} \geq V_{\text{min}}) \geq 1 - \delta$$
## Table of Contents

1. **Notation**

2. **Create a safe batch reinforcement learning algorithm**
   - Off-policy policy evaluation (OPE)
   - Safe policy improvement (SPI)
Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, $\pi_e$, Convert historical data, $D$, into $n$ independent and unbiased estimates of $V^{\pi_e}$

- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the $n$ independent and unbiased estimates of $V^{\pi_e}$ into a $1 - \delta$ confidence lower bound on $V^{\pi_e}$

- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforcement learning algorithm,

- Methods today focused on work by Philip Thomas UAI and ICML 2015 papers.
Off-policy policy evaluation (OPE)

\[
\text{Historical Data, } D \\
\text{Proposed Policy, } \pi_e \\
\text{Estimate of } \sqrt{\pi_e}
\]
High-confidence off-policy policy evaluation (HCOPE)

Historical Data, $D$
Proposed Policy, $\pi_e$
Probability, $1 - \delta$

$\rightarrow 1 - \delta$ confidence lower bound on $\sqrt[\pi_e]$
Safe policy improvement (SPI)

Historical Data, \( D \)  
Probability, \( 1 - \delta \)  
\[ \rightarrow \]  
New policy \( \pi \), or  
No Solution Found
Create a safe batch reinforcement learning algorithm

- **Off-policy policy evaluation (OPE)**
  - For any evaluation policy, $\pi_e$, Convert historical data, $D$, into $n$ independent and unbiased estimates of $V^{\pi_e}$

- **High-confidence off-policy policy evaluation (HCOPE)**
  - Use a concentration inequality to convert the $n$ independent and unbiased estimates of $V^{\pi_e}$ into a $1 - \delta$ confidence lower bound on $V^{\pi_e}$

- **Safe policy improvement (SPI)**
  - Use HCOPE method to create a safe batch reinforcement learning algorithm,
Monte Carlo (MC) Off Policy Evaluation

- Aim: estimate value of policy $\pi_1$, $V^{\pi_1}(s)$, given episodes generated under behavior policy $\pi_2$
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from $\pi_2$
- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ in MDP $M$ under policy $\pi$
- $V^{\pi}(s) = \mathbb{E}_\pi[G_t|s_t = s]$
- Have data from a different policy, behavior policy $\pi_2$
- If $\pi_2$ is stochastic, can often use it to estimate the value of an alternate policy (formal conditions to follow)
- Again, no requirement that have a model nor that state is Markov
Monte Carlo (MC) Off Policy Evaluation: Distribution Mismatch

- Distribution of episodes & resulting returns differs between policies
Importance Sampling

- Goal: estimate the expected value of a function $f(x)$ under some probability distribution $p(x)$, $\mathbb{E}_{x \sim p}[f(x)]$
- Have data $x_1, x_2, \ldots, x_n$ sampled from distribution $q(s)$
- Under a few assumptions, we can use samples to obtain an unbiased estimate of $\mathbb{E}_{x \sim q}[f(x)] = \frac{1}{N} \sum_{i=1}^{N} f(x_i)$ where $x_i \sim q$
Importance Sampling

\[ E_{X \sim q} [f(X)] = \int x \cdot q(x) f(x) \, dx \]

\[ = \int x \cdot \frac{q(x)}{q(x)} p(x) f(x) \, dx \]

\[ = \int x \cdot q(x) \left[ \frac{p(x)}{q(x)} \right] f(x) \, dx \]

\[ = E_{X \sim q} \left[ \frac{p(x)}{q(x)} f(x) \right] \]

\[ \approx \frac{1}{N} \sum_{i=1}^{N} \frac{p(x_i)}{q(x_i)} f(x_i) \]

\[ q(x) > 0 \quad \text{forall} \quad x \quad \text{such that} \quad p(x) > 0 \]
We can use importance sampling to do batch bandit policy evaluation. Consider we have a dataset for pulls from 3 arms. Consider that arm 1 is a Bernoulli where with probability \(0.98\) we get 0 and with probability \(0.02\) we get 100. Arm 2 is a Bernoulli where with probability \(0.55\) the reward is 2 else the reward is 0. Arm 3 has a probability of yielding a reward of 1 with probability \(0.5\) else it gets 0. Select all that are true.

- Data is sampled from \(\pi_1\) where with probability 0.8 it pulls arm 3 else it pulls arm 2. The policy we wish to evaluate, \(\pi_2\), pulls arm 2 with probability 0.5 else it pulls arm 1. \(\pi_2\) has higher true reward than \(\pi_1\).

- We cannot use \(\pi_1\) to get an unbiased estimate of the average reward \(\pi_2\) using importance sampling.

- We can use \(\pi_1\) to get a lower bound on the average reward of \(\pi_2\) using importance sampling.

- If rewards can be positive or negative, we can still get a lower bound on \(\pi_2\) using data from \(\pi_1\) using importance sampling.

- Now assume \(\pi_1\) selects arm 1 with probability 0.2 and arm 2 with probability 0.8. We can use importance sampling to get an unbiased estimate of \(\pi_2\) using data from \(\pi_1\).

- Still with the same \(\pi_1\), it is likely with \(N=20\) pulls that the estimate using IS for \(\pi_2\) will be higher than the empirical value of \(\pi_1\).

- Not sure
• Let $h_j$ be episode $j$ (history) of states, actions and rewards

$$h_j = (s_{j,1}, a_{j,1}, r_{j,1}, s_{j,2}, a_{j,2}, r_{j,2}, \ldots, s_{j,L_j}(\text{terminal}))$$
Importance Sampling (IS) for Policy Evaluation

- Let $h_j$ be episode $j$ (history) of states, actions and rewards

$$h_j = (s_{j,1}, a_{j,1}, r_{j,1}, s_{j,2}, a_{j,2}, r_{j,2}, \ldots, s_{j,L_j(\text{terminal})})$$

$$p(h_j|\pi, s = s_{j,1}) = p(a_{j,1}|s_{j,1})p(r_{j,1}|s_{j,1}, a_{j,1})p(s_{j,2}|s_{j,1}, a_{j,1})$$

$$p(a_{j,2}|s_{j,2})p(r_{j,2}|s_{j,2}, a_{j,2})p(s_{j,3}|s_{j,2}, a_{j,2}) \ldots$$

$$= \prod_{t=1}^{L_j-1} p(a_{j,t}|s_{j,t})p(r_{j,t}|s_{j,t}, a_{j,t})p(s_{j,t+1}|s_{j,t}, a_{j,t})$$

$$= \prod_{t=1}^{L_j-1} \pi(a_{j,t}|s_{j,t})p(r_{j,t}|s_{j,t}, a_{j,t})p(s_{j,t+1}|s_{j,t}, a_{j,t})$$
Importance Sampling (IS) for Policy Evaluation

- Let $h_j$ be episode $j$ (history) of states, actions and rewards, where the actions are sampled from $\pi_2$

$$h_j = (s_{j,1}, a_{j,1}, r_{j,1}, s_{j,2}, a_{j,2}, r_{j,2}, \ldots, s_{j,L_j(\text{terminal})})$$

$$V^{\pi_1}(s) \approx \sum_{j=1}^{n} \frac{p(h_j | \pi_1, s)}{p(h_j | \pi_2, s)} G(h_j)$$

$$= \prod_{i=1}^{L_j-1} \frac{p(a_{i+1} | \pi_1, s_i)}{p(a_{i+1} | \pi_2, s_i)} \frac{p(s_{i+1} | \pi_1, s_i)}{p(s_{i+1} | \pi_2, s_i)}$$
Importance Sampling for Policy Evaluation

- Aim: estimate $V^{\pi_1}(s)$ given episodes generated under policy $\pi_2$
  - $s_1, a_1, r_1, s_2, a_2, r_2, \ldots$ where the actions are sampled from $\pi_2$
- Have access to $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$ in MDP $M$ under policy $\pi_2$
- Want $V^{\pi_1}(s) = \mathbb{E}_{\pi_1}[G_t|s_t = s]$
- IS = Monte Carlo estimate given off policy data
- Model-free method
- Does not require Markov assumption
- Under some assumptions, unbiased & consistent estimator of $V^{\pi_1}$
- Can be used when agent is interacting with environment to estimate value of policies different than agent’s control policy
Leveraging Future Can’t Influence Past Rewards

- Importance sampling (IS):

\[
IS(D) = \frac{1}{n} \sum_{i=1}^{n} \left( \prod_{t=1}^{L} \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)} \right) \left( \sum_{t=1}^{L} \gamma^t R_t^{i} \right)
\]

- Per-decision importance sampling (PDIS)

\[
PSID(D) = \sum_{t=1}^{L} \gamma^t \frac{1}{n} \sum_{i=1}^{n} \left( \prod_{\tau=1}^{t} \frac{\pi_e(a_\tau | s_\tau)}{\pi_b(a_\tau | s_\tau)} \right) R_t^{i}
\]

Emma Brunskill (CS234 Reinforcement Learning)
Off-policy policy evaluation

*Note: we only went carefully through slides before this point. The remaining slides are kept for those interested but will not be material required for the quiz. See the last slide for a summary of what you should know

- Importance sampling (IS):

\[
IS(D) = \frac{1}{n} \sum_{i=1}^{n} w_i \left( \sum_{t=1}^{L} \gamma^t R^i_t \right)
\]

- Weighted importance sampling (WIS)

\[
WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left( \sum_{t=1}^{L} \gamma^t R^i_t \right)
\]
Off-policy policy evaluation

- Weighted importance sampling (WIS)

\[
WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left( \sum_{t=1}^{L} \gamma^t R_t^i \right)
\]

- Biased or unbiased?
Off-policy policy evaluation

- Weighted importance sampling (WIS)
  
  \[ WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left( \sum_{t=1}^{L} \gamma^t R_t^i \right) \]

- Biased. When \( n = 1 \), \( \mathbb{E}[WIS] = V(\pi_b) \)
- Strongly consistent estimator of \( V^{\pi_e} \)
  - i.e. \( \Pr(\lim_{n \to \infty} WIS(D) = V^{\pi_e}) = 1 \)
  - If
    - Finite horizon
    - One behavior policy, or bounded rewards
Control variates

- Given: \( X \)
- Estimate: \( \mu = \mathbb{E}[X] \)
- \( \hat{\mu} = X \)
- Unbiased: \( \mathbb{E}[\hat{\mu}] = \mathbb{E}[X] = \mu \)
- Variance: \( \text{Var}(\hat{\mu}) = \text{Var}(X) \)
Control variates

- Given: $X, Y, \mathbb{E}[Y]$
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X - Y + \mathbb{E}[Y]$
- Unbiased:
  $\mathbb{E}[\hat{\mu}] =$
- Variance:

$$Var(\hat{\mu}) = Var(X - Y + \mathbb{E}[Y]) = Var(X - Y)$$
Control variates

- Given: $X, Y, \mathbb{E}[Y]$
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X - Y + \mathbb{E}[Y]$
- Unbiased:
  $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X - Y + \mathbb{E}[Y]] = \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[Y] = \mathbb{E}[X] = \mu$
- Variance:

  \[
  \text{Var}(\hat{\mu}) = \text{Var}(X - Y + \mathbb{E}[Y]) = \text{Var}(X - Y)
  \]

  \[
  = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)
  \]

- Lower variance if $2\text{Cov}(X, Y) > \text{Var}(Y)$
- We call $Y$ a control variate
- We saw this idea before: baseline term in policy gradient estimation
Off-policy policy evaluation

- Idea: add a control variate to importance sampling estimators
  - $X$ is the importance sampling estimator
  - $Y$ is a control variate build from an approximate model of the MDP

- Called the doubly robust estimator (Jiang and Li, 2015)
  - Robust to (1) poor approximate model, and (2) error in estimates of $\pi_b$
    - If the model is poor, the estimates are still unbiased
    - If the sampling policy is unknown, but the model is good, MSE will still be low

- Non-recursive and weighted forms, as well as control variate view provided by Thomas and Brunskill (ICML 2016)
Off-policy policy evaluation

\[ DR(\pi_e \mid D) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{\infty} \gamma^t w^i_t (R^i_t - \hat{q}^{\pi_e}(S^i_t, A^i_t)) + \gamma^t \rho^i_t - 1 \hat{v}^{\pi_e}(S^i_t), \]

where \( w^i_t = \prod_{\tau_1}^{t} \frac{\pi_e(a_{\tau} \mid s_{\tau})}{\pi_b(a_{\tau} \mid s_{\tau})} \)
Empirical Results (Gridworld)

![Graph showing the mean squared error as a function of the number of episodes.](image)

- **Approximate model** (Dudik, 2011)
- **Indirect method** (Sutton and Barto, 1998)
Empirical Results (Gridworld)

Below is a graph showing the mean squared error vs. the number of episodes. The graph compares different methods labeled as IS, PDIS, and AM. The x-axis represents the number of episodes, and the y-axis represents the mean squared error on a logarithmic scale.

The graph illustrates that as the number of episodes increases, the mean squared error decreases for all methods, indicating improved performance or accuracy.
Empirical Results (Gridworld)

![Graph showing mean squared error vs. number of episodes](image)

- **IS**
- **PDIS**
- **DR**
- **AM**
Empirical Results (Gridworld)
Empirical Results (Gridworld)

![Graph showing empirical results for different algorithms in Gridworld. The x-axis represents the number of episodes, and the y-axis represents the mean squared error. The graph includes lines for IS, PDIS, WIS, CWPDIS, DR, AM, and WDR, each indicated by a different color and style.]
Off-policy policy evaluation: Blending

- Importance sampling is unbiased but high variance
- Model based estimate is biased but low variance
- Doubly robust is one way to combine the two
- Can also trade between importance sampling and model based estimate within a trajectory
- MAGIC estimator (Thomas and Brunskill ICML 2016)
- Can be particularly useful when part of the world is non-Markovian in the given model, and other parts of the world are Markov
Can Need an Order of Magnitude Less Data To Get Good Estimates

![Graph showing Mean Squared Error vs Number of Episodes]

- IS
- DR
- AM
- WDR
- MAGIC

Number of Episodes, n

Mean Squared Error

10
1
0.1
0.01
0.001
1
10
100
1,000
10,000
• What if \( \text{supp}(\pi_e \subset \text{supp}(\pi_b)) \)
• There is a state-action pair, \((s, a)\), such that \(\pi_e(a \mid s) = 0\), but \(\pi_b(a \mid s) \neq 0\).
• If we see a history where \((s, a)\) occurs, what weight should we give it?

\[
IS(D) = \frac{1}{n} \sum^n_{i=1} \left( \prod^{L}_{t=1} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \right) \left( \sum^L_{t=1} \gamma^t R_t^i \right)
\]
Off-policy policy evaluation

- What if there are zero samples ($n = 0$)?
  - The importance sampling estimate is undefined
- What if no samples are in $\text{supp} (\pi_e)$ (or $\text{supp} (p)$ in general)?
  - Importance sampling says: the estimate is zero
  - Alternate approach: undefined
- Importance sampling estimator is unbiased if $n > 0$
- Alternate approach will be unbiased given that at least one sample is in the support of $p$
- Alternate approach detailed in Importance Sampling with Unequal Support (Thomas and Brunskill, AAAI 2017)
Off-policy policy evaluation

Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, $\pi_e$, Convert historical data, $D$, into $n$ independent and unbiased estimates of $V^{\pi_e}$

- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the $n$ independent and unbiased estimates of $V^{\pi_e}$ into a $1 - \delta$ confidence lower bound on $V^{\pi_e}$

- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforcement learning algorithm,
High-confidence off-policy policy evaluation

- Consider using IS + Hoeffding’s inequality for HCOPE on mountain car

Figure 3: Mountain Car (Sarsa(λ))
Natural Temporal Difference Learning, Dabney and Thomas, 2014
Hoeffding’s inequality

• Let $X_1, \ldots, X_n$ be $n$ independent identically distributed random variables such that $X_i \in [0, b]$

• Then with probability at least $1 - \delta$:

$$\mathbb{E}[X_i] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}},$$

where $X_i = \frac{1}{n} \sum_{i=1}^{n} (w_i \sum_{t=1}^{L} \gamma^t R_{it})$ in our case.
High-confidence off-policy policy evaluation

- Using 100,000 trajectories
- Evaluation policy’s true performance is $0.19 \in [0, 1]$
- We get a 95% confidence lower bound of: $-5,8310,000$
What went wrong

\[ w_i = \prod_{t=1}^{L} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \]
High-confidence off-policy policy evaluation

- Removing the upper tail only decreases the expected value.
High-confidence off-policy policy evaluation

- Thomas et. al, High confidence off-policy evaluation, AAAI 2015

Theorem 1. Let $X_1, \ldots, X_n$ be $n$ independent real-valued random variables such that for each $i \in \{1, \ldots, n\}$, we have $\mathbb{P}[0 \leq X_i] = 1$, $\mathbb{E}[X_i] \leq \mu$, and some threshold value $c_i > 0$. Let $\delta > 0$ and $Y_i := \min\{X_i, c_i\}$. Then with probability at least $1 - \delta$, we have

$$
\mu \geq \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{n} Y_i - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \frac{7n \ln(2/\delta)}{3(n-1)} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left( \frac{Y_i}{c_i} - \frac{Y_j}{c_j} \right)^2}.
$$  \hspace{1cm} (3)

The terms $\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{n} \frac{1}{c_i}$ are empirical means, the term $\frac{7n \ln(2/\delta)}{3(n-1)}$ goes to zero as $1/n$ as $n \to \infty$, and the term $\sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left( \frac{Y_i}{c_i} - \frac{Y_j}{c_j} \right)^2}$ goes to zero as $1/\sqrt{n}$ as $n \to \infty$.
High-confidence off-policy policy evaluation

95% Confidence Lower Bound on Mean

\[ c \]

\[ 1, 10, 100, 1,000, 10,000, 100,000, 1,000,000 \]

\[ n=2, n=4, n=8, n=16, n=32, n=64, n=128, n=256, n=512, n=1024, n=2048, n=4096, n=8192, n=16384, n=32768 \]
High-confidence off-policy policy evaluation

- Use 20% of the data to optimize $c$ (cutoff)
- Use 80% to compute lower bound with optimized $c$
- Mountain car results:

<table>
<thead>
<tr>
<th></th>
<th>CUT</th>
<th>Chernoff-Hoeffding</th>
<th>Maurer</th>
<th>Anderson</th>
<th>Bubeck et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Confidence lower bound on the mean</td>
<td>0.145</td>
<td>$-5,831,000$</td>
<td>$-129,703$</td>
<td>0.055</td>
<td>$-.046$</td>
</tr>
</tbody>
</table>
High-confidence off-policy policy evaluation

Digital marketing:

![Graph showing expected returns and confidence levels](image)
Cognitive dissonance:

\[ \mathbb{E}[X_i] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}} \]
High-confidence off-policy policy evaluation

- Student’s t-test
  - Assumes that $IS(D)$ is normally distributed
  - By the central limit theorem, it (is as $n \to \infty$)

$$
\Pr \left( \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] \geq \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}}{\sqrt{n}} t_{1-\delta, n-1} \\
\geq 1 - \delta
$$

- Efron’s Bootstrap methods (e.g., BCa)
  - Also, without importance sampling: Hanna, Stone, and Niekum, AAMAS 2017
High-confidence off-policy policy evaluation

Create a safe batch reinforcement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, \( \pi_e \), Convert historical data, \( D \), into \( n \) independent and unbiased estimates of \( V^{\pi_e} \)

- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the \( n \) independent and unbiased estimates of \( V^{\pi_e} \) into a \( 1 - \delta \) confidence lower bound on \( V^{\pi_e} \)

- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforcement learning algorithm
Safe policy improvement

Thomas et al, ICML 2015

Historical Data

Training Set (20%)

Testing Set (80%)

Candidate Policy, $\pi$

Safety Test

Is $1 - \delta$ confidence lower bound on $J(\pi)$ larger than $J(\pi_{\text{cur}})$?
Empirical Results: Digital Marketing

Agent → Environment
Action, \( a \) → State, \( s \) → Reward, \( r \)
Empirical Results: Digital Marketing

![Graph showing expected normalized return for different sample sizes and methods.]

- None, CUT
- None, BCa
- k-Fold, CUT
- k-Fold, Bca
Empirical Results: Digital Marketing

![Graph showing comparison between Initial Policy and New Policy in terms of Mean Return. The graph indicates a significant increase in Mean Return under the New Policy.]
Empirical Results: Digital Marketing
Other Relevant Work

- How to deal with long horizons? (Guo, Thomas, Brunskill NIPS 2017)
- How to deal with importance sampling being “unfair”? (Doroudi, Thomas and Brunskill, best paper UAI 2017)
- What to do when the behavior policy is not known? (Liu, Gottesman, Raghu, Komorowski, Faisal, Doshi-Velez, Brunskill NeurIPS 2018)
- What to do when the behavior policy is deterministic?
- What to do when care about safe exploration?
- What to do when care about performance on a single trajectory
- Many others also doing great work in this space, including the groups of Yisong Yue, Susan Murphy, Finale Doshi-Velez, Marco Pavone, Pieter Abbeel, Shie Mannor, Sergey Levine and Claire Tomlin, amongst others
• Very important topic: healthcare, education, marketing, ...
• Insights are relevant to on policy learning
• Big focus of my lab
• A number of others on campus also working in this area (e.g. Stefan Wager, Susan Athey...)  
• Very interesting area at the intersection of causality and control
• Our Science 2019 paper show how to do safe policy improvement for a high fidelity diabetes simulator and discusses the need for ensuring good behavior
What You Should Know: Off Policy Policy Evaluation and Selection

- Be able to define and apply importance sampling for off policy policy evaluation
- Define some limitations of IS (variance)
- List a couple alternatives (weighted IS, doubly robust)
- Define why we might want safe reinforcement learning
Class Structure

- Last time: Fast Reinforcement Learning
- This time: Batch RL
- Next time: Guest Lecture