Lecture 15: Batch RL

Emma Brunskill
CS234 Reinforcement Learning.

Winter 2019
Slides drawn from Philip Thomas with modifications
Class Structure

• Last time: Meta Reinforcement Learning
• This time: Batch RL
• Next time: Quiz
A Scientific Experiment

A Group

\[
\begin{align*}
\frac{3}{6} + \frac{2}{8} &= \frac{18}{24} + \frac{4}{8} = \frac{3}{4} \\
\frac{2}{10} + \frac{3}{4} &= \frac{19}{20} = \frac{19}{20}
\end{align*}
\]

Finally, reduce the sum to lowest terms:

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

\[
\frac{4}{5} \quad \frac{9}{10}
\]

Avg Score: 95
A Group

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

Finally, reduce the sum to lowest terms:

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

B Group

\[
\frac{4}{5} \div \frac{3}{10} = \frac{40}{30} = \frac{4}{3}
\]

Evaluate the cross-multiplication strategy:

\[
\frac{4}{3} \times \frac{3}{10} = \frac{12}{30} = \frac{2}{5}
\]

Finally, reduce the sum to lowest terms:

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

Avg Score: 95

Avg Score: 92

What to do for a new student?
What Should We Do For a New Student?

A Group

B Group

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

B Group

Avg Score: 95

Avg Score: 92

???
Batch Reinforcement Learning

**Off Policy Offline Batch RL**

A Group

B Group

B Group

Avg Score: 95

Avg Score: 92

???
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
Table of Contents

1 Notation

2 Create a safe batch reinforcement learning algorithm
   • Off-policy policy evaluation (OPE)
   • High-confidence off-policy policy evaluation (HCOPE)
   • Safe policy improvement (SPI)
Notation

- Policy $\pi$: $\pi(a) = P(a_t = a | s_t = s)$
- 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:

$$V^\pi = \mathbb{E}\left[ \sum_{t=1}^{L} \gamma^t R_t | \pi \right]$$
Safe batch reinforcement learning algorithm

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

$$\Pr(V_{A(D)} \geq V_{\pi_b}) \geq 1 - \delta$$

or, in general

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

1. Notation

2. Create a safe batch reinforcement learning algorithm
   - Off-policy policy evaluation (OPE)
   - High-confidence off-policy policy evaluation (HCOPE)
   - 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,
    $$\arg\max_{\pi} V^{\pi_e}$$
    with some confidence bounds
Off-policy policy evaluation (OPE)

Historical Data, $D$
Proposed Policy, $\pi_e$

$\rightarrow$ Estimate of $\sqrt[\pi_e]$
Importance Sampling

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

\[ \mathbb{E}[IS(D)] = V^{\pi_e} \]
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
think back to exploration

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

$1 - \delta$ confidence lower bound on $J(\pi_e)$
Hoeffding's inequality

- Let $X_1, \cdots, X_n$ 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 $X_i = \frac{1}{n} \sum_{i=1}^{n} (w_i \sum_{t=1}^{L} \gamma^t R_t^i)$ in our case.
Safe policy improvement (SPI)

Historical Data, $D$

Probability, $1 - \delta$

$\implies$

New policy $\pi$, or
No Solution Found
Safe policy improvement (SPI)

Historical Data

Training Set (20%)

Candidate Policy, $\pi$

Testing Set (80%)

Safety Test

Is $1 - \delta$ confidence lower bound on $J(\pi)$ larger than $J(\pi_{\text{cur}})$?
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 $\pi_2$

- $D = \{s_1, a_1, r_1, s_2, a_2, r_2, \ldots \}$ 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

\[ T - s \, a \, r \, s' \, ... \]

\[ \text{distrib} \not\to \quad \pi_0 \quad \not\to \quad \text{distrib} \not\to \quad \mu \]

\[ \rho(T) \not\subseteq \tau(c) \]
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 $\mathbb{E}_{x \sim q}[f(x)]$

$$\mathbb{E}_{x \sim q}[f(x)] = \int_x q(x)f(x) \, dx$$

When $x_i \sim p(x)$

$$= \int_x p(x) \frac{q(x)}{p(x)} f(x) \, dx$$

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

$$\approx \frac{1}{n} \sum_i \frac{q(x_i)}{p(x_i)} f(x_i)$$
Importance Sampling (IS) for Policy Evaluation

\[ \frac{p(h_j | \pi_e)}{p(h_j | \pi_b)} \]

• Let \( h_j \) be episode \( j \) (history) of states, actions and rewards

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

\[ p(h_j | \pi_e) = p(s_{j,1}) \prod_{j=1}^{L_j} p(a_j | s_j) p(s_{j+1} | s_j, a_j) p(r_j | a_j, s_j) \]

\[ \frac{p(h_j | \pi_e)}{p(h_j | \pi_b)} = \frac{p(s_{j,1})}{p(s_{j,1})} \prod_{j=1}^{L_j} \frac{p(a_j | s_j)^{\pi_e}}{p(a_j | s_j)^{\pi_b}} \frac{p(s_{j+1} | s_j, a_j)^{\pi_e}}{p(s_{j+1} | s_j, a_j)^{\pi_b}} \frac{p(r_j | a_j, s_j)^{\pi_e}}{p(r_j | a_j, s_j)^{\pi_b}} \]

\[ = \prod_{j=1}^{L_j} \frac{\pi_e(a_j | s_j)}{\pi_b(a_j | s_j)} \]

Markov condition: full history
Importance Sampling (IS) for Policy Evaluation

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

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

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

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

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

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

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

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

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

$$= \sum_{j=1}^{n} \left( \prod_{i=1}^{L} \frac{\pi_e(a_{j,i} | s_{j,i})}{\pi_b(a_{j,i} | s_{j,i})} \right) G(h_j)$$

$$\pi_b(a | s) = 0 \quad \text{but} \quad \pi_e(a | s) > 0$$

$$\pi_b(a | s) = 0 \quad \text{if} \quad \pi_e(a | s) = 0$$
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):

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

- Per-decision importance sampling (PDIS)

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

\[
G_t = r + r' + \cdots
\]

\[
\text{recall } \pi \text{ gradient}
\]

\[
\pi \nabla G_t = r + r' + \cdots
\]

\[
\text{only up to point got reward}
\]
Off-policy policy evaluation (revisited)

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

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

\[ \frac{\pi(\alpha)}{\pi_0(\alpha)} \leq \text{might be super small} \]
Off-policy policy evaluation (revisited)

- 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: $Var(\hat{\mu}) = Var(X)$
Control variates

- Given: $X, Y, \mathbb{E}[Y]$
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X - Y + \mathbb{E}[Y]$
- Unbiased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X - Y + \mathbb{E}[Y]] = \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[Y] = \mathbb{E}[X]$
- 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 $2\text{Cov}(X, Y) > \text{Var}(Y)$ if true: $\text{Var}(\hat{\mu}) < \text{Var}(X)$
- We call $Y$ a control variate
- We saw this idea before: baseline term in policy gradient estimation
Off-policy policy evaluation (revisited)

- 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

\[ E[Y] = 0 \text{ in this case} \]

PDIS\(_{CV}(D) = \text{PDIS}(D) - \text{CV}(D) \]

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

\[ \text{DR}(D) = \text{PDIS}_{CV}(D) \]

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

• 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

$$E[Y] = 0$$ in this case

$\text{PDIS}_{\text{CV}}(D) = \text{PDIS}(D) - \text{CV}(D)$

• 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 (revisited)

\[ DR(\pi_e \mid D) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{\infty} \gamma^t w_t^i (R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_t^i \hat{v}^{\pi_e}(S_t^i), \]

where \( w_t^i = \prod_{\tau=1}^{t} \frac{\pi_e(a_\tau \mid s_\tau)}{\pi_b(a_\tau \mid s_\tau)} \)
Empirical Results (Gridworld)

\[ V^\pi_e - V^\pi_t \]

Error Scale

MSE

Approximate model
- Direct method (Dudik, 2011)
- Indirect method (Sutton and Barto, 1998)

Mean Squared Error

Number of Episodes, n

0.001 0.01 0.1 1 10 100 1000

IS
AM

\[ \text{Approximate model} \]

\[ \text{Direct method (Dudik, 2011)} \]

\[ \text{Indirect method (Sutton and Barto, 1998)} \]
Empirical Results (Gridworld)

The graph shows the mean squared error as a function of the number of episodes. The y-axis represents the mean squared error on a logarithmic scale, ranging from 0.001 to 10,000. The x-axis represents the number of episodes, ranging from 2 to 2,000.

Three lines are plotted:
- **IS** (solid line)
- **PDIS** (dashed line)
- **AM** (dotted line)

As the number of episodes increases, the mean squared error decreases for all three methods.
Empirical Results (Gridworld)

\[ \hat{V}^{\pi_c} - V^{\pi_c} \]

![Graph showing the comparison of IS, PDIS, DR, and AM methods over the number of episodes, with AM + IS highlighted.](image)
Empirical Results (Gridworld)

![Graph showing mean squared error versus number of episodes](Graph.png)
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. Lines for different algorithms are shown, indicating their performance over the number of episodes.]
Off-policy policy evaluation (revisited): Blending

\[
MSE = f(bias, var)
\]

- 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 against number of episodes.](image)
Off-policy policy evaluation (revisited)

- 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?
- $IS(D) = \frac{1}{n} \sum_{i=1}^{n} \left( \prod_{t=1}^{L} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \right) \left( \sum_{t=1}^{L} \gamma^t R_t^i \right)$

![Diagram showing probability of history and policies]
Off-policy policy evaluation (revisited)

- 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 \( \text{supp}(\rho) \) 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 \( \rho \)
- Alternate approach detailed in Importance Sampling with Unequal Support (Thomas and Brunskill, AAAI 2017)
Off-policy policy evaluation (revisited)

Create a safe batch reinforcement learning algorithm

\[ P \left( \forall A(d) - V^{\pi_e} > 0 \right) \geq 1 - \delta \]

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

- Let $X_1, \cdots, X_n$ 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 $X_i = \frac{1}{n} \sum_{i=1}^{n} (w_i \sum_{t=1}^{L} \gamma^t R_t)$ in our case.
High-confidence off-policy policy evaluation (revisited)

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

$V \text{ true is between } 0 \& 1$

$19 > -5 \text{ million}$
What went wrong

\[ W_i = \prod_{t=1}^{L} \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)} \]

Pretty small \( \frac{1}{10^2} \)
• Removing the upper tail only decreases the expected value.
High-confidence off-policy policy evaluation (revisited)

- 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

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

(3)
High-confidence off-policy policy evaluation (revisited)
High-confidence off-policy policy evaluation (revisited)

- Use 20% of the data to optimize $c$ (cutoff)
- Use 80% to compute lower bound with optimized $c$
- Mountain car results: $100 \%$ true

<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>−0.046</td>
</tr>
</tbody>
</table>

True $\frac{19}{100}$
High-confidence off-policy policy evaluation (revisited)

Digital marketing:

![Graph showing expected return vs confidence](image-url)
High-confidence off-policy policy evaluation (revisited)

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 (revisited)

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

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

$\geq 1 - \delta$

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

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 (revisited)

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

Expected Normalized Return

- 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. Initial Policy has a Mean Return of approximately 0.052, while the New Policy has a higher Mean Return of approximately 0.06.]
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
- For last two, see great work by Marco Pavone’s group, Pieter Abbeel’s group, Shie Mannor’s group and Claire Tomlin’s group, 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
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
- Define the scope of the guarantees implied by safe policy improvement as defined in this lecture
Class Structure

- Last time: Meta Reinforcement Learning
- **This time:** Batch RL
- Next time: Quiz
Off-policy policy evaluation (revisited)

- Weighted per-decision importance sampling
  - Also called consistent weighted per-decision importance sampling
  - A fun exercise!