Class Structure

- Last time: Fast Reinforcement Learning / Exploration and Exploitation
- This time: Batch RL
- Next time: Monte Carlo Tree Search
1. What makes an RL algorithm safe?

2. Notation

3. Create a safe batch reinforcement learning algorithm
   - Off-policy policy evaluation (OPE)
   - High-confidence off-policy policy evaluation (HCOPE)
   - Safe policy improvement (SPI)
What does it mean to for a reinforcement learning algorithm to be safe?

Scholarly articles for safe reinforcement learning
Safe reinforcement learning - Thomas - Cited by 10
Lyapunov design for safe reinforcement learning - Perkins - Cited by 70
Reinforcement learning: A survey - Kaelbling - Cited by 5842

A Comprehensive Survey on Safe Reinforcement Learning
by J Garcia - 2015 - Cited by 27 - Related articles
A Comprehensive Survey on Safe Reinforcement Learning. Javier García figpolo@inf.uc3m.es.
Fernando Fernández ffernand@inf.uc3m.es. Universidad...
A Comprehensive Survey on Safe Reinforcement Learning

Javier García
Fernando Fernández
University Carlos III de Madrid,
Avenida de la Universidad 30,
28911 Leganes, Madrid, Spain

FJGPOLO@INF.UC3M.ES
FFERNAND@INF.UC3M.ES
Table 1: Overview of the approaches for Safe Reinforcement Learning considered in this survey.
Changing the objective

Policy 1
-50  +0  +20  +20  +20  +20  +20  +20  

Policy 2
+0  +0  +0  +0  +0  +0  +0  +20
Changing the objective

- Policy 1:
  - Reward = 0 with probability 0.999999
  - Reward = $10^9$ with probability 1-0.999999
  - Expected reward approximately 1000

- Policy 2:
  - Reward = 999 with probability 0.5
  - Reward = 1000 with probability 0.5
  - Expected reward 999.5
Another notion of safety

Safe and efficient off-policy reinforcement learning

Rémi Munos  
munos@gmail.com  
Google DeepMind

Anna Harutyunyan  
amna.harutyunyan@vub.ac.be  
Vrije Universiteit Brussel

Thomas Stepleton  
stepleton@gmail.com  
Google DeepMind

Marc G. Bellemare  
bellemare@gmail.com  
Google DeepMind
Another notion of safety (Munos et. al)

We start from the recent work of Harutyunyan et al. (2016), who show that naive off-policy policy evaluation, without correcting for the “off-policyness” of a trajectory, still converges to the desired $Q^\pi$ value function provided the behavior $\mu$ and target $\pi$ policies are not too far apart (the maximum allowed distance depends on the $\lambda$ parameter). Their $Q^\pi (\lambda)$ algorithm learns from trajectories generated by $\mu$ simply by summing discounted off-policy corrected rewards at each time step. Unfortunately, the assumption that $\mu$ and $\pi$ are close is restrictive, as well as difficult to uphold in the control case, where the target policy is greedy with respect to the current Q-function. In that sense this algorithm is not safe: it does not handle the case of arbitrary “off-policyness”.

Alternatively, the Tree-backup (TB($\lambda$)) algorithm (Precup et al., 2000) tolerates arbitrary target/behavior discrepancies by scaling information (here called traces) from future temporal differences by the product of target policy probabilities. TB($\lambda$) is not efficient in the “near on-policy” case (similar $\mu$ and $\pi$), though, as traces may be cut prematurely, blocking learning from full returns.
Another notion of safety

Reachability-Based Safe Learning with Gaussian Processes

Anayo K. Akametalu*  
Shahab Kaynama

Jaime F. Fisac*  
Melanie N. Zeilinger

Jeremy H. Gillula  
Claire J. Tomlin
SAFE REINFORCEMENT LEARNING

A Dissertation Presented

by

PHILIP S. THOMAS
• If you apply an existing method, do you have confidence that it will work?
Reinforcement learning success
A property of many real applications

- Deploying "bad" policies can be costly or dangerous
Deploying bad policies can be costly
Deploying bad policies can be 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. What makes an RL algorithm safe?

2. Notation

3. 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** $\pi$: $\pi(a) = P(a_t = a \mid s_t = s)$
- **History**: $H = (s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_L, a_L, r_L)$
- **Historical data**: $D = \{H_1, H_2, \cdots, H_n\}$
- **Historical data from behavior policy**, $\pi_b$
- **Objective**:

  $$V^\pi = \mathbb{E}\left[\sum_{t=1}^{L} \gamma^t R_t \mid \pi\right]$$
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

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

1. What makes an RL algorithm safe?

2. Notation

3. 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, \( a \)
Off-policy policy evaluation (OPE)

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

$\rightarrow$ Estimate of $J(\pi_e)$
Importance Sampling (Reminder)

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

\[ \mathbb{E}[IS(D)] = V^{\pi_e} \]
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, $\alpha$
High-confidence off-policy policy evaluation (HCOPE)

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 $X_i = \frac{1}{n} \sum_{i=1}^{n} (w_i \sum_{t=1}^{L} \gamma^t R^i_t)$ in our case.
Safe policy improvement (SPI)

Historical Data, $D$
Probability, $1 - \delta$

\[ \{ \]

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

Historical Data

Training Set (20%)

Testing Set (80%)

Candidate Policy, $\pi$

Safety Test

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

WON’T WORK!
Off-policy policy evaluation (revisited)

• 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)

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

- Importance sampling (IS):

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

- 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 (revisited)

• Weighted importance sampling (WIS)

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

• NOT unbiased. When \( n = 1 \), \( \mathbb{E}[WIS] = J(\pi_b) \)

• 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
Off-policy policy evaluation (revisited)

- Weighted per-decision importance sampling
  - Also called consistent weighted per-decision importance sampling
  - A fun exercise!
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:
  \begin{align*}
  \mathbb{E}[\hat{\mu}] &= \mathbb{E}[X - Y + \mathbb{E}[Y]] = \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[Y] = \mathbb{E}[X] = \mu
  \end{align*}
- Variance:
  \begin{align*}
  \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)
  \end{align*}
- 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 (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
    - $\mathbb{E}[Y] = 0$ in this case
    - $PDIS_{CV}(D) = PDIS(D) - CV(D)$
  - 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
    - $DR(D) = PDIS_{CV}(D)$
  - Non-recursive and weighted forms, as well as control variate view provided by Thomas and Brunskill (2016)
Off-policy policy evaluation (revisited)

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

where \( w^i_t = \prod_{t=1}^{t} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \)

- Recall: we want the control variate \( Y \) to cancel with \( X \):

\[
R - q(S, A) + \gamma v(S')
\]
Empirical Results (Gridworld)

![Graph showing empirical results for Gridworld. The graph plots Mean Squared Error against the Number of Episodes, n. Two lines are shown: IS and AM. The IS line starts at a higher error and decreases more rapidly than the AM line. The graph also includes notes on Approximate model (Dudik, 2011) and Indirect method (Sutton and Barto, 1998).]
Empirical Results (Gridworld)

![Graph showing Empirical Results (Gridworld)]
Empirical Results (Gridworld)

![Graph showing empirical results for different methods in Gridworld. The x-axis represents the number of episodes, and the y-axis represents mean squared error. The graph compares IS, PDiS, DR, and AM methods, showing their performance over different numbers of episodes.]
Empirical Results (Gridworld)

The graph shows the mean squared error over the number of episodes. The x-axis represents the number of episodes, ranging from 2 to 2,000. The y-axis represents the mean squared error, ranging from 0.001 to 10,000. Several lines are plotted, each representing a different method or algorithm:

- IS
- PDIS
- WIS
- CWPDIS
- DR
- AM

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

The diagram shows the empirical results for different algorithms in a gridworld environment. The x-axis represents the number of episodes, n, ranging from 2 to 2,000. The y-axis represents the mean squared error. Various lines indicate different algorithms, with labels such as IS, PDIS, WIS, CWGDIS, DR, AM, and WDR. The lines demonstrate how each algorithm performs across different episode counts, with some algorithms showing faster convergence than others.
Important 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 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
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?

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

![Diagram showing evaluation and behavior policies](image-url)
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)
Off-policy policy evaluation (revisited)

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, $a$
High-confidence off-policy policy evaluation (revisited)

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

![Diagram of mountain car with inelastic wall and goal position.]

**Figure 3: Mountain Car (Sarsa(λ))**

Natural Temporal Difference Learning, Dabney and Thomas, 2014.
• 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

\[ wi = \prod_{t=1}^{L} \frac{\pi_e(a_t | s_t)}{\pi_b(a_t | s_t)} \]
• 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 $P[0 \leq X_i] = 1$, $E[X_i] \leq \mu$, 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

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

- Use 20% of the data to optimize $c$
- 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>$-0.046$</td>
</tr>
</tbody>
</table>
High-confidence off-policy policy evaluation (revisited)

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

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

Empirical Error Rate

Number of samples

0.05

0

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, $a$
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
Empirical Results: Digital Marketing

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

- n=10000
- n=30000
- n=60000
- n=100000

Expected Normalized Return
- None, CUT
- None, BCa
- k-Fold, CUT
- k-Fold, Bca
Empirical Results: Digital Marketing
Empirical Results: Digital Marketing
Example Results: Diabetes Treatment

Blood Glucose (sugar)

Eat Carbohydrates
Release Insulin

Graph showing blood glucose levels over time with arrows indicating changes due to carbohydrate intake and insulin release.
Example Results: Diabetes Treatment

Blood Glucose (sugar)

Eat Carbohydrates → Release Insulin

Hyperglycemia
Example Results: Diabetes Treatment

- **Blood Glucose (sugar)**
- **Eat Carbohydrates**
- **Release Insulin**

Graph showing:
- **Hyperglycemia**
- **Hypoglycemia**

Graph: Blood Glucose (mg/dL) vs. Time (minutes since midnight)
Example Results: Diabetes Treatment

\[
\text{injection} = \frac{\text{blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR}
\]
Example Results: Diabetes Treatment

Intelligent Diabetes Management

T1DMS
Type 1 Diabetes Metabolic Simulator
Example Results: Diabetes Treatment

![Graphs showing the change in probability policy and the probability policy worse over time.](image)

- Probability Policy Changed
- Probability Policy Worse
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?
- What to do when the behavior policy is deterministic?
- What to do when care about doing 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: Exploration and Exploitation
- This time: Batch RL
- Next time: Monte Carlo Tree Search