Batch / Offline RL Policy Learning

Emma Brunskill

March 2 2023

CS234

Thanks to Phil Thomas for some figures

Refresh Your Understanding

Importance sampling (select all that are true)

- Requires the behavior policy to visit all the state--action pairs that would be visited under the evaluation policy in order to get an unbiased estimator
- Is likely to be high variance
- Not Sure

Behavior cloning from demonstrations:

- Reduces batch/offline learning to supervised learning
- May learn a low performing policy if the demonstrations come from a non-expert
- May learn a low performing policy if the demonstrations from an expert
- Could be used to warm start an online reinforcement learning algorithm
- Requires a human to label what they would do at the states visited by the policy learned
- Not Sure

Refresh Your Understanding

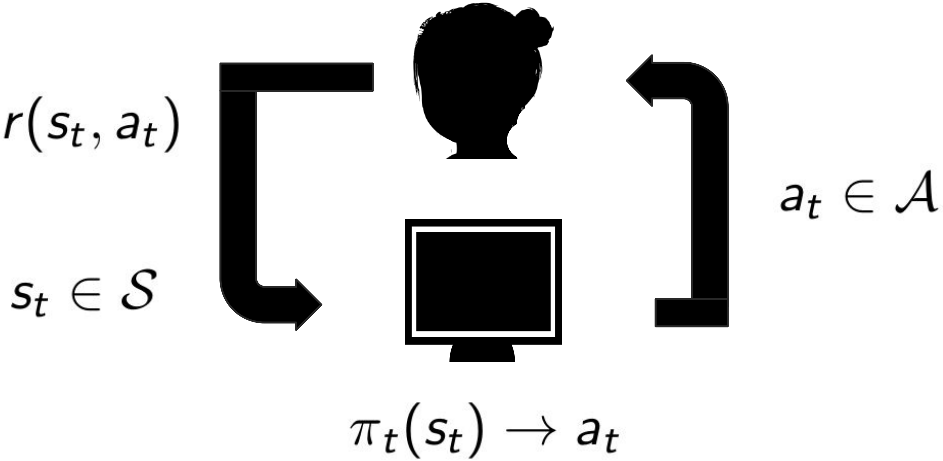
Importance sampling (select all that are true)

- Requires the behavior policy to visit all the state--action pairs that would be visited under the evaluation policy in order to get an unbiased estimator (true)
- Is likely to be high variance (true)
- Not Sure

Behavior cloning from demonstrations:

- Reduces batch/offline learning to supervised learning
- May learn a low performing policy if the demonstrations come from a non-expert
- May learn a low performing policy if the demonstrations from an expert
- Could be used to warm start an online reinforcement learning algorithm
- Requires a human to label what they would do at the states visited by the policy learned
- Not Sure

Today: Counterfactual / Batch RL



$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

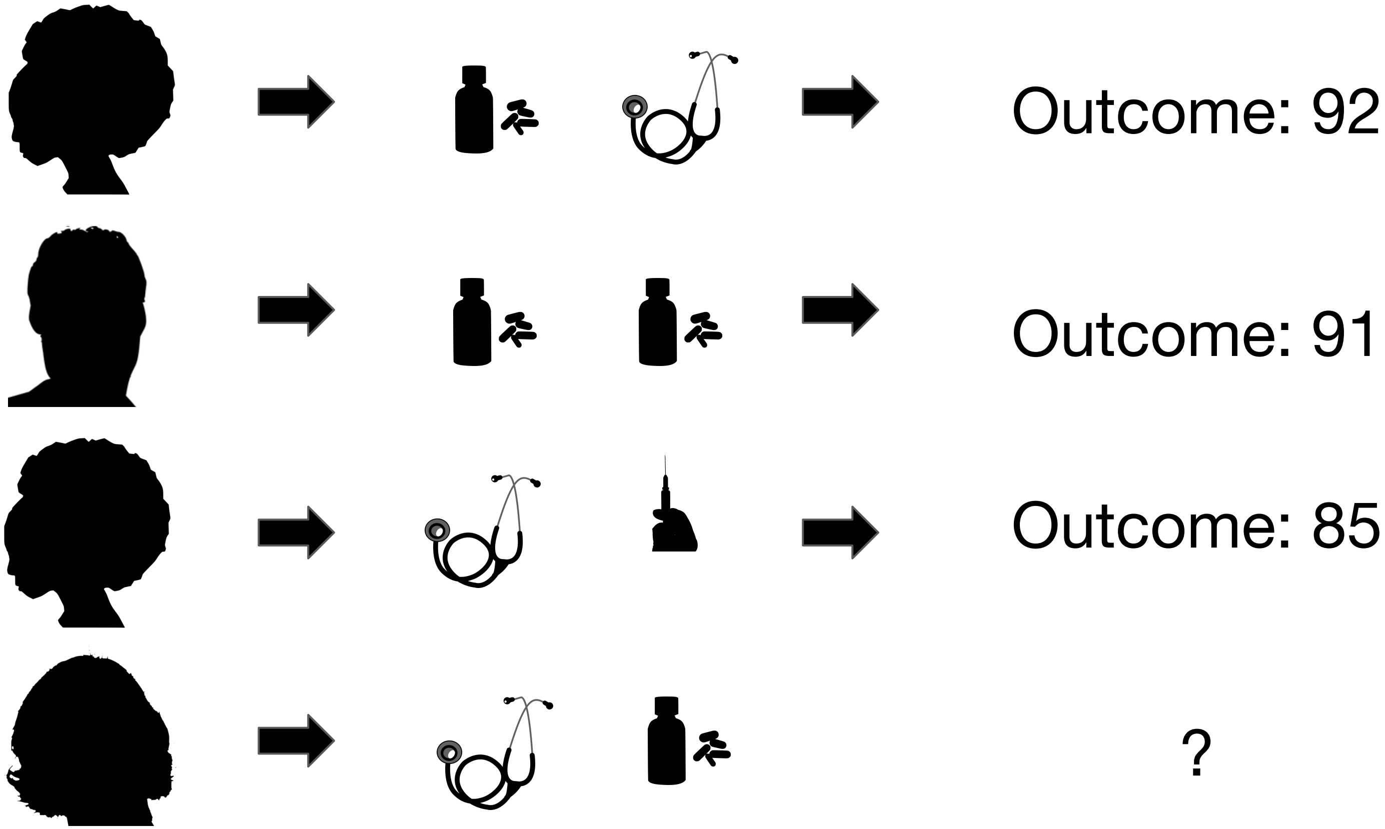
Where We Are In The Course

1. Learning from offline data
   a. Imitation learning
   b. Batch/offline policy evaluation
   c. Batch/offline policy learning
2. Next week
   a. Guest lecture
   b. Quiz

Today

1. Imitation vs batch/offline RL policy learning
2. Fitted Q Iteration / Offline Q Learning
3. Pessimism
4. Case Study

Is the Hope for Batch RL over Imitation Learning?

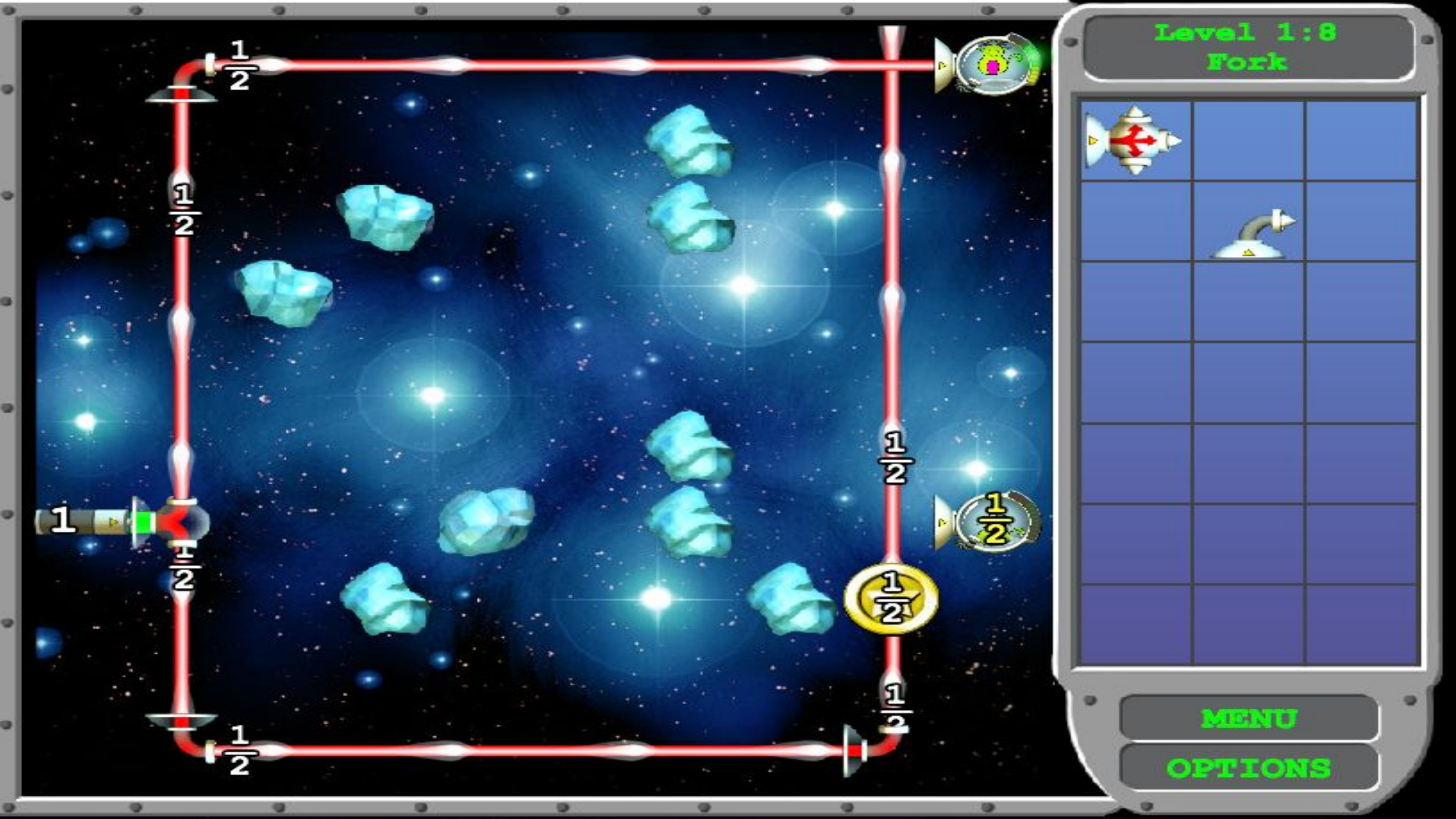


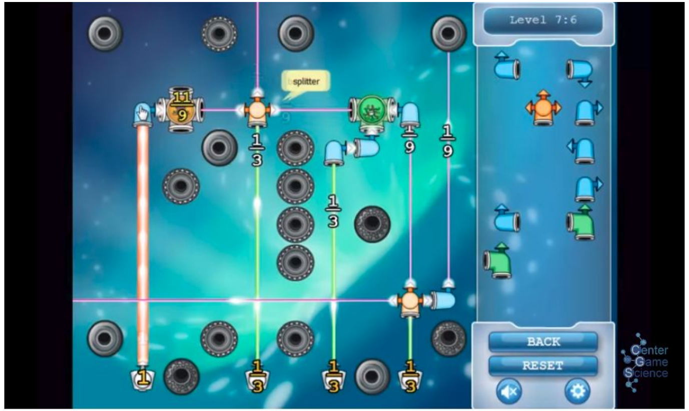
Outcome: 92

Outcome: 91

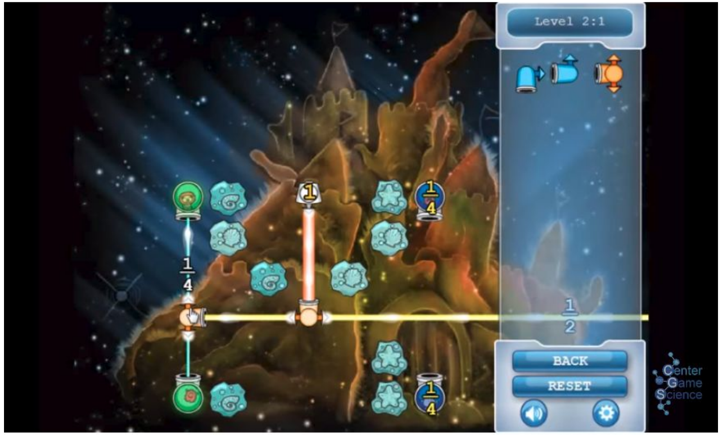
Outcome: 85

?

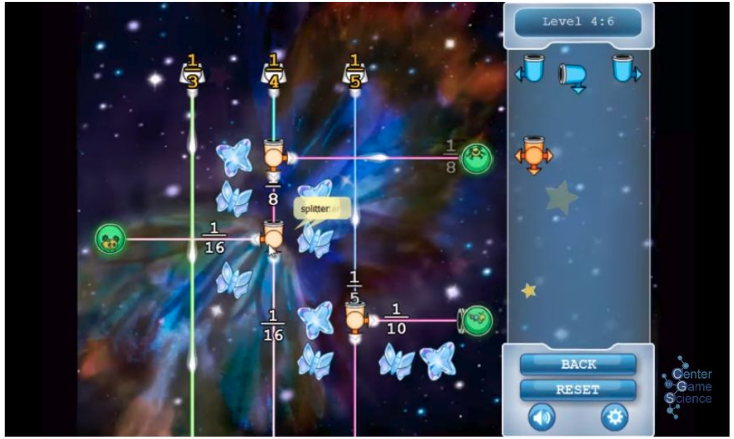

Level 1:8
Fork
1
1/2
1/2
1/2
1/2
1/2
1/2
1/2
1/2
1/2
MENU
OPTIONS

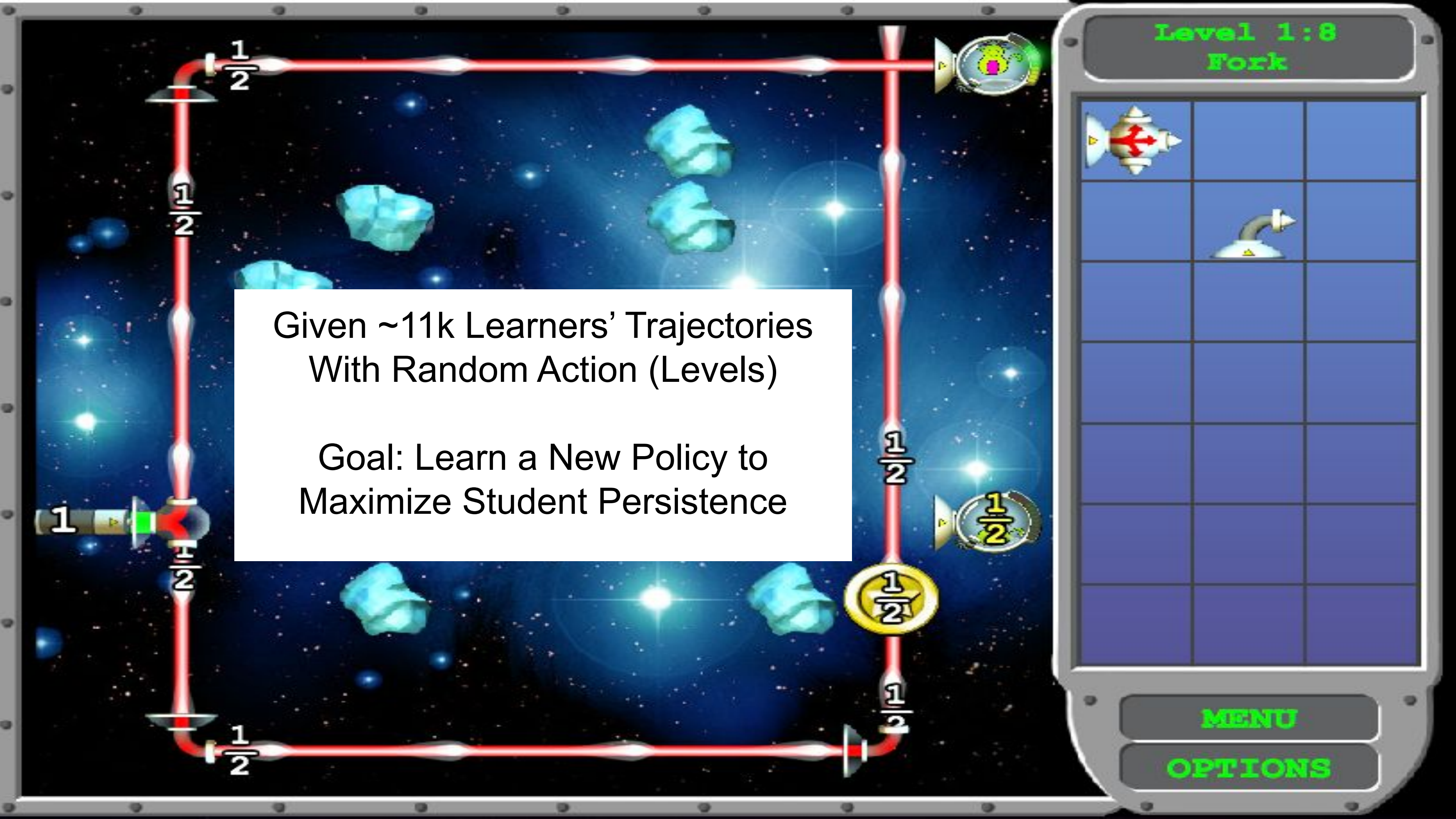


Took > 30s



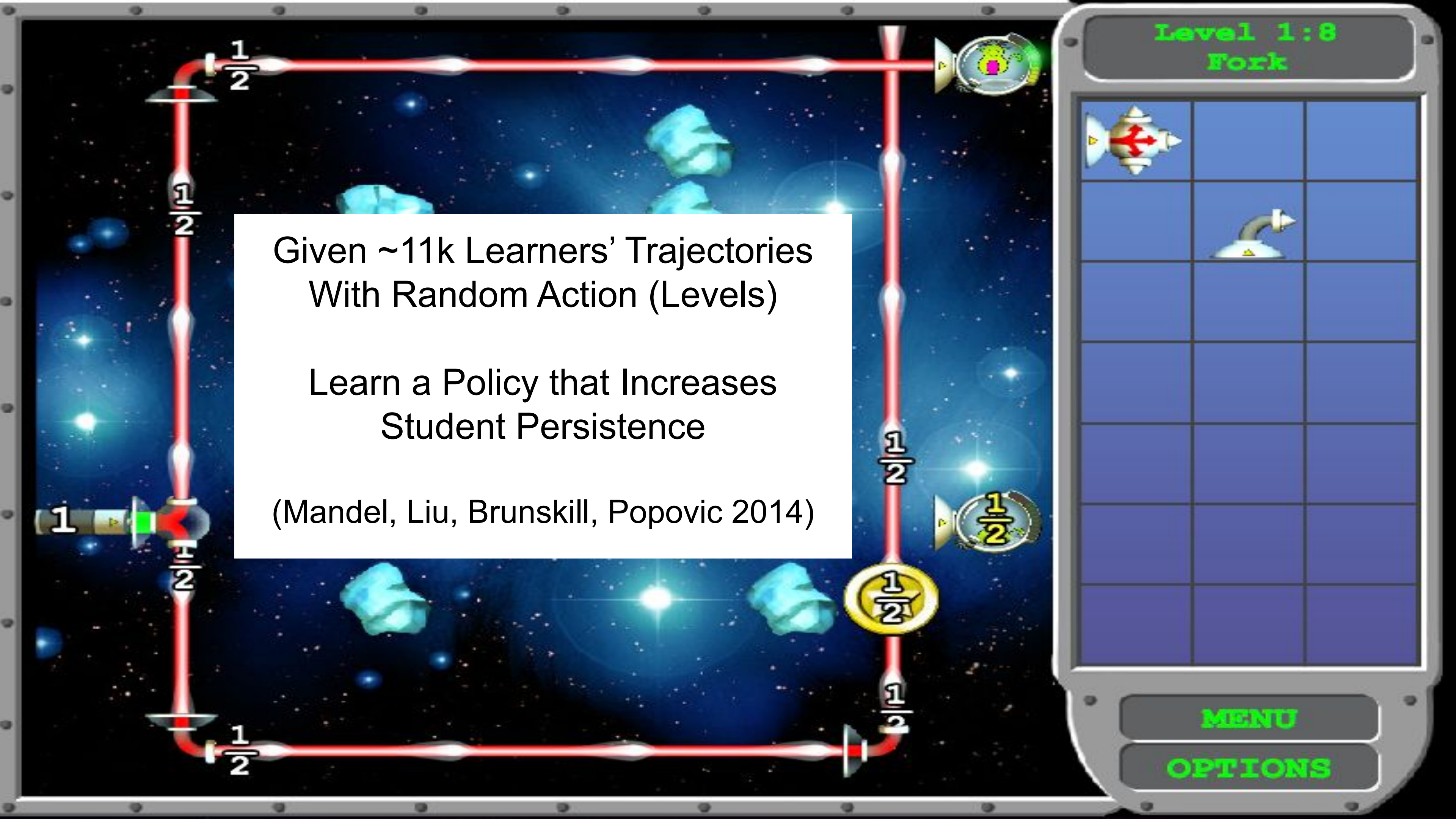
Took <= 30s





Given ~11k Learners’ Trajectories
With Random Action (Levels)

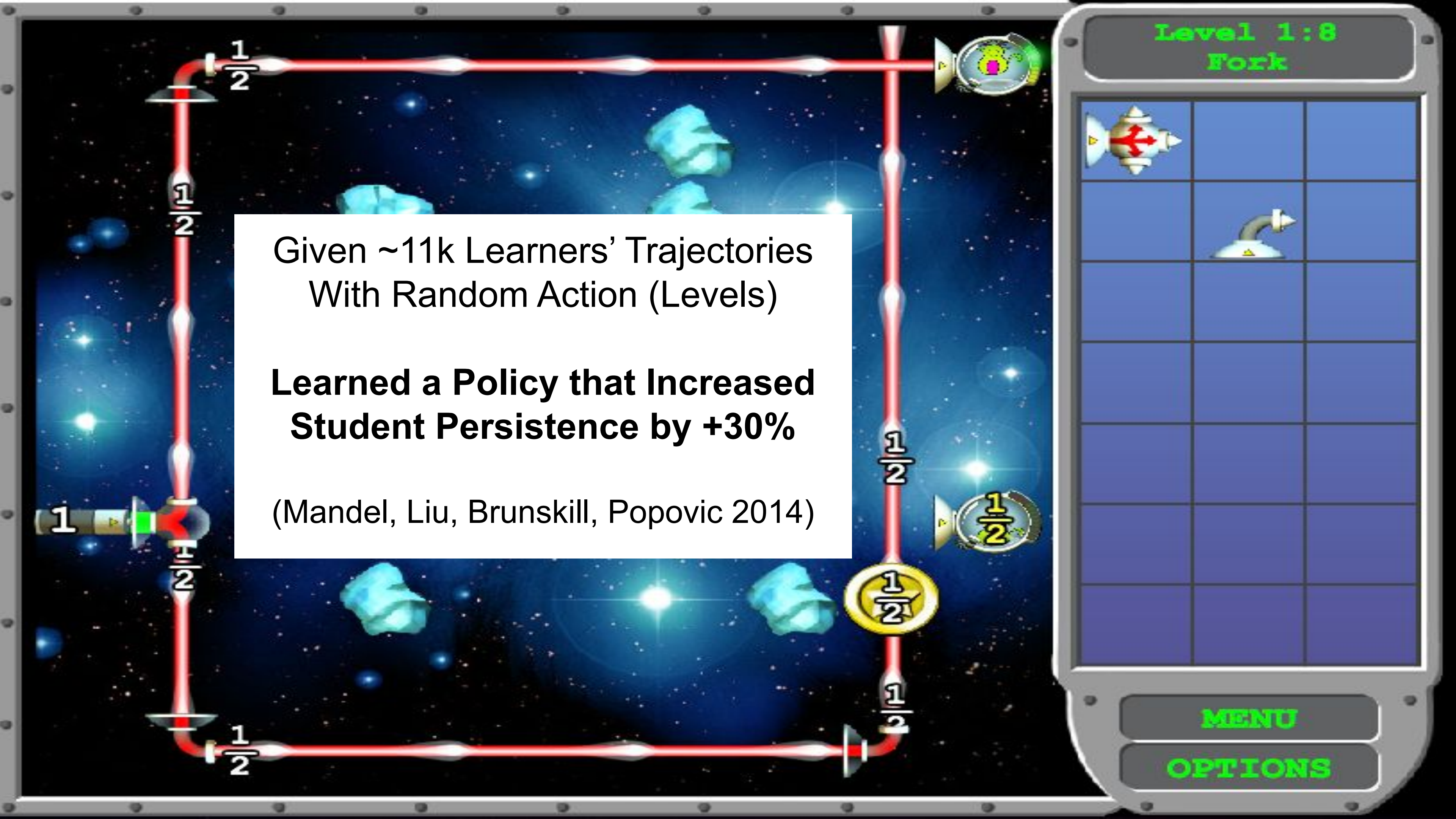
Goal: Learn a New Policy to
Maximize Student Persistence



Given ~11k Learners’ Trajectories With Random Action (Levels)

Learn a Policy that Increases Student Persistence

(Mandel, Liu, Brunskill, Popovic 2014)

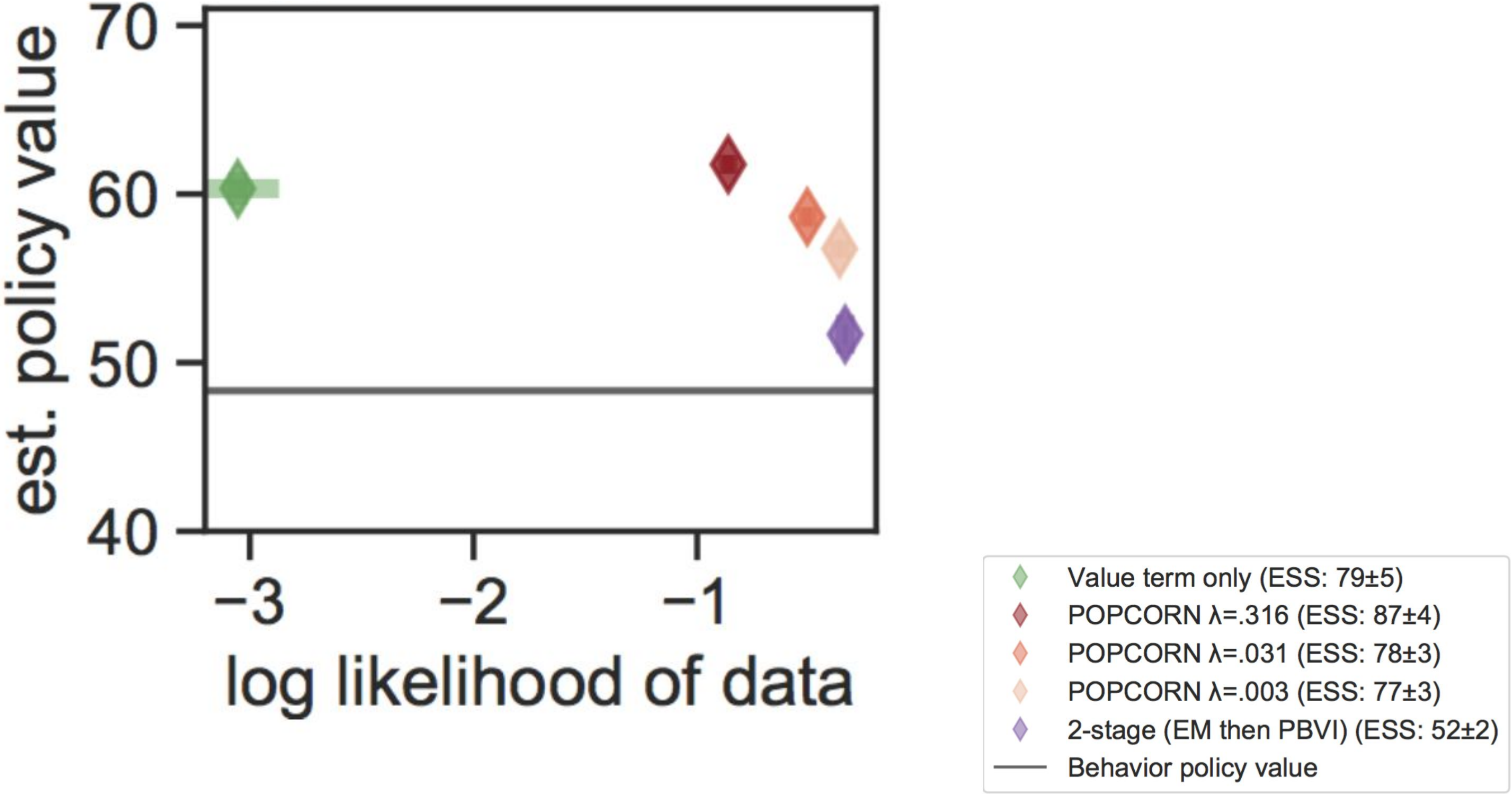


Given ~11k Learners' Trajectories With Random Action (Levels)

Learned a Policy that Increased Student Persistence by +30%

(Mandel, Liu, Brunskill, Popovic 2014)

Encouraging Recent Work on Observational Health Data (MIMIC) Hypotension



Futoma, Hughes, Doshi-Velez AISTATS 2020

Today

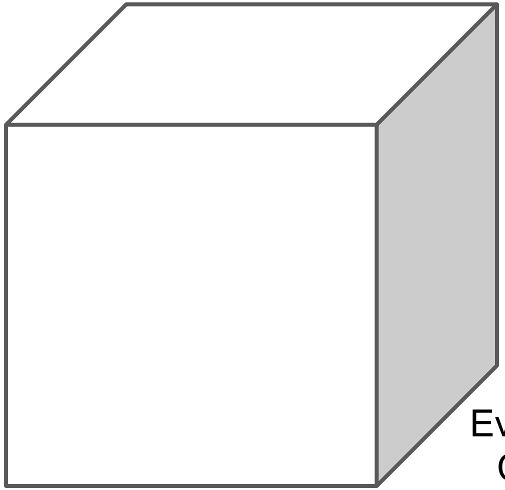
1. Imitation vs batch/offline RL policy learning
2. Fitted Q Iteration / Offline Q Learning
3. Pessimism
4. Case study

Offline / Batch Reinforcement Learning

Tasks

$$\int_{s\in S_0} \hat{V}^{\pi}(s,\mathcal{D})ds$$

$$\arg\max_{\pi\in\mathcal{H}_i} \int_{s\in S_0} \hat{V}^{\pi}(s,\mathcal{D})ds$$



Evaluation Criteria

- Empirical accuracy
- Consistency
- Robustness
- Asymptotic efficiency
- Finite sample bounds
- Computational cost

Assumptions

- Markov?
- Overlap?
- Sequential ignorability?

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

π: Policy mapping $s \to a$

S_0: Set of initial states

$\hat{V}^{\pi}(s,\mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Batch Policy Optimization: Find a Good Policy That Will Perform Well in the Future

$$\underbrace{\arg\max_{\pi\in\mathcal{H}_i}\ \max_{\mathcal{H}_i\in\{\mathcal{H}_1,\mathcal{H}_2,\dots\}}}_{\text{Policy Optimization}}\quad \underbrace{\int_{s\in S_0}\hat{V}^{\pi}(s,\mathcal{D})ds}_{\text{Policy Evaluation}}$$

$$\mathcal{H} = \mathcal{M}, \mathcal{V},\ \mathit{\Pi}\ ?$$

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

π: Policy mapping $s \to a$

S_0: Set of initial states

$\hat{V}^{\pi}(s,\mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

- Today will not be a comprehensive overview, but instead highlight some of the challenges involved & some approaches with desirable statistical properties convergence, sample efficiency & bounds

Policy Optimization: Find Good Policy to Deploy

$$\arg\max_{\pi\in\mathcal{H}_i} \quad \max_{\mathcal{H}_i\in\{\mathcal{H}_1,\mathcal{H}_2,\dots\}} \quad \int_{s\in S_0} \hat{V}^{\pi}(s,\mathcal{D})ds$$

$$\mathcal{H} = \mathcal{M}, \mathcal{V}, \ \Pi \ ?$$

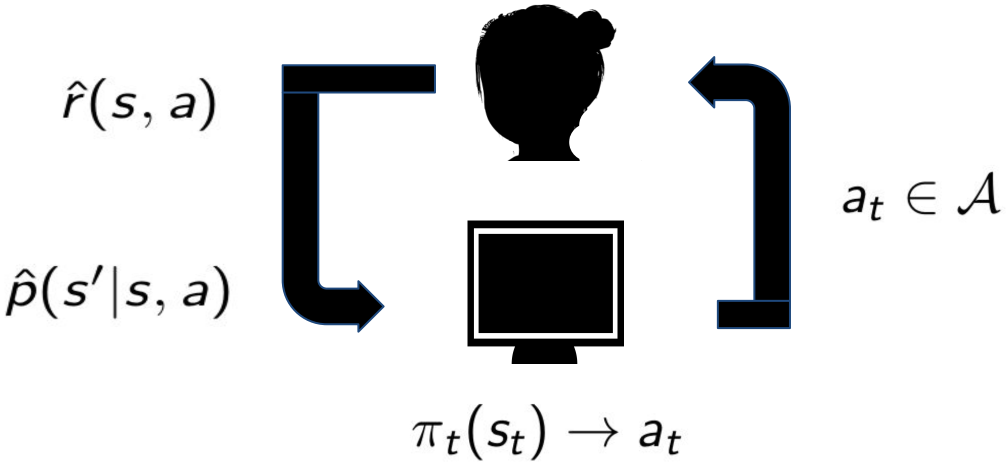
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

π: Policy mapping $s \rightarrow a$

S_0: Set of initial states

$\hat{V}^{\pi}(s,\mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Learn Dynamics and Reward Models from Data, Plan



$$\hat{V}^*(s) = \max_a \hat{r}(s, a) + \gamma \sum_{s'} \hat{p}(s'|s, a) \hat{V}^*(s')$$

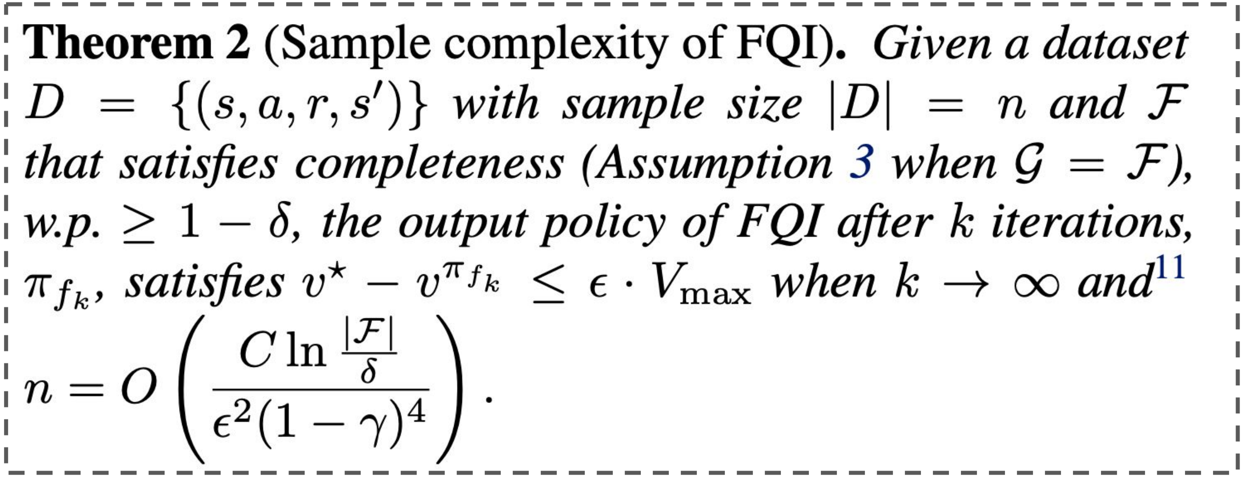
Model Free Value Function Approximation: Fitted Q Iteration

$$\mathcal{D} = (s_i, a_i, r_i, s_{i+1}) \; \forall i$$

$$(\mathcal{T} f)(s, a) := R(s, a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[V_f(s')]$$

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Value Function Estimation, Fitted Q Iteration



Theorem 2 (Sample complexity of FQI)**.** *Given a dataset $D = \{(s, a, r, s')\}$ with sample size $|D| = n$ and $\mathcal{F}$ that satisfies completeness (Assumption 3 when $\mathcal{G} = \mathcal{F}$), w.p. $\geq 1 - \delta$, the output policy of FQI after k iterations, π_{f_k}, satisfies $v^\star - v^{\pi_{f_k}} \leq \epsilon \cdot V_{\max}$ when $k \to \infty$ and*[11]

$$n = O\left(\frac{C \ln \frac{|\mathcal{F}|}{\delta}}{\epsilon^2 (1-\gamma)^4}\right).$$

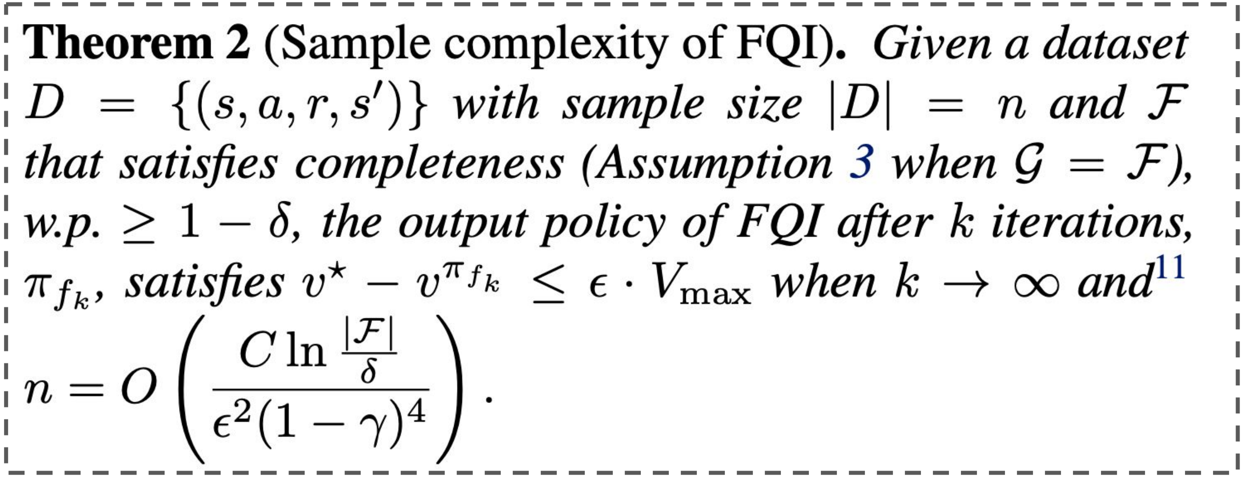
$$\forall f \in \mathcal{F}, \mathcal{T} f \in \mathcal{G}.$$

$$Q^\star \in \mathcal{F}$$

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A},\ \frac{\nu(s, a)}{\mu(s, a)} \leq C.$$

Chen & Jiang ICML 2019

Value Function Estimation, Fitted Q Iteration



Theorem 2 (Sample complexity of FQI). *Given a dataset $D = \{(s, a, r, s')\}$ with sample size $|D| = n$ and $\mathcal{F}$ that satisfies completeness (Assumption 3 when $\mathcal{G} = \mathcal{F}$), w.p. $\geq 1 - \delta$, the output policy of FQI after k iterations, π_{f_k}, satisfies $v^\star - v^{\pi_{f_k}} \leq \epsilon \cdot V_{\max}$ when $k \to \infty$ and*[11]

$$n = O\left(\frac{C \ln \frac{|\mathcal{F}|}{\delta}}{\epsilon^2(1-\gamma)^4}\right).$$

$$\forall f \in \mathcal{F}, \mathcal{T} f \in \mathcal{G}.$$

Completeness

$$Q^\star \in \mathcal{F}$$

Realizability

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \frac{\nu(s,a)}{\mu(s,a)} \leq C.$$

Overlap assumption: Concentratability coefficient

Chen & Jiang ICML 2019

Today

1. Imitation vs batch/offline RL policy learning
2. Fitted Q Iteration / Offline Q Learning
3. **Pessimism**
4. Case Study

Check Your Intuition

- Optimism under uncertainty can enable sublinear regret in online multi-armed bandits
- Pessimism under uncertainty can lead to linear regret in online multi-armed bandits
- With high probability the optimistic upper bound on the selected arm in UCB algorithms is an upper bound on the performance of any arm
- In offline / batch RL selecting the optimistic best arm is likely to be best
- In offline / batch RL selecting the arm with the highest mean is likely to be best
- Not sure

Check Your Intuition Solutions

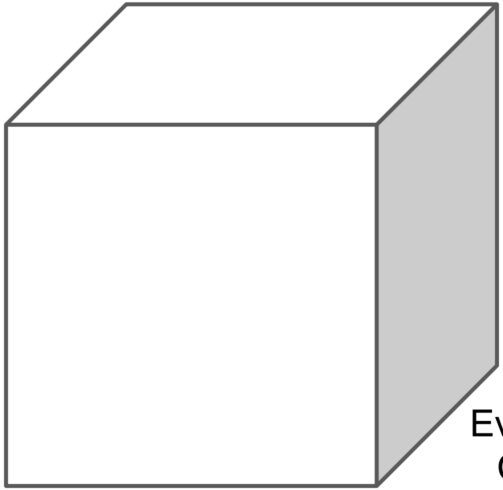
- Optimism under uncertainty can enable sublinear regret in online multi-armed bandits
- Pessimism under uncertainty can lead to linear regret in online multi-armed bandits
- With high probability the optimistic upper bound on the selected arm in UCB algorithms is an upper bound on the performance of any arm
- In offline / batch RL selecting the optimistic best arm is likely to be best
- In offline / batch RL selecting the arm with the highest mean is likely to be best
- Not sure

Offline / Batch Reinforcement Learning

Tasks

$$\int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds$$

$$\arg\max_{\pi \in \mathcal{H}_i} \int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds$$



Evaluation Criteria

- Empirical accuracy
- Consistency
- Robustness
- Asymptotic efficiency
- Finite sample bounds
- Computational cost
- **Constraints?**

Assumptions

- Markov?
- **Overlap?**
- **Sequential ignorability?**

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \to a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Standard Assumptions for Off Policy / Counterfactual Estimation & Optimization

- Overlap
  - Have to take all actions that target policy would take
  - In infinite data / finite data
- No confounding

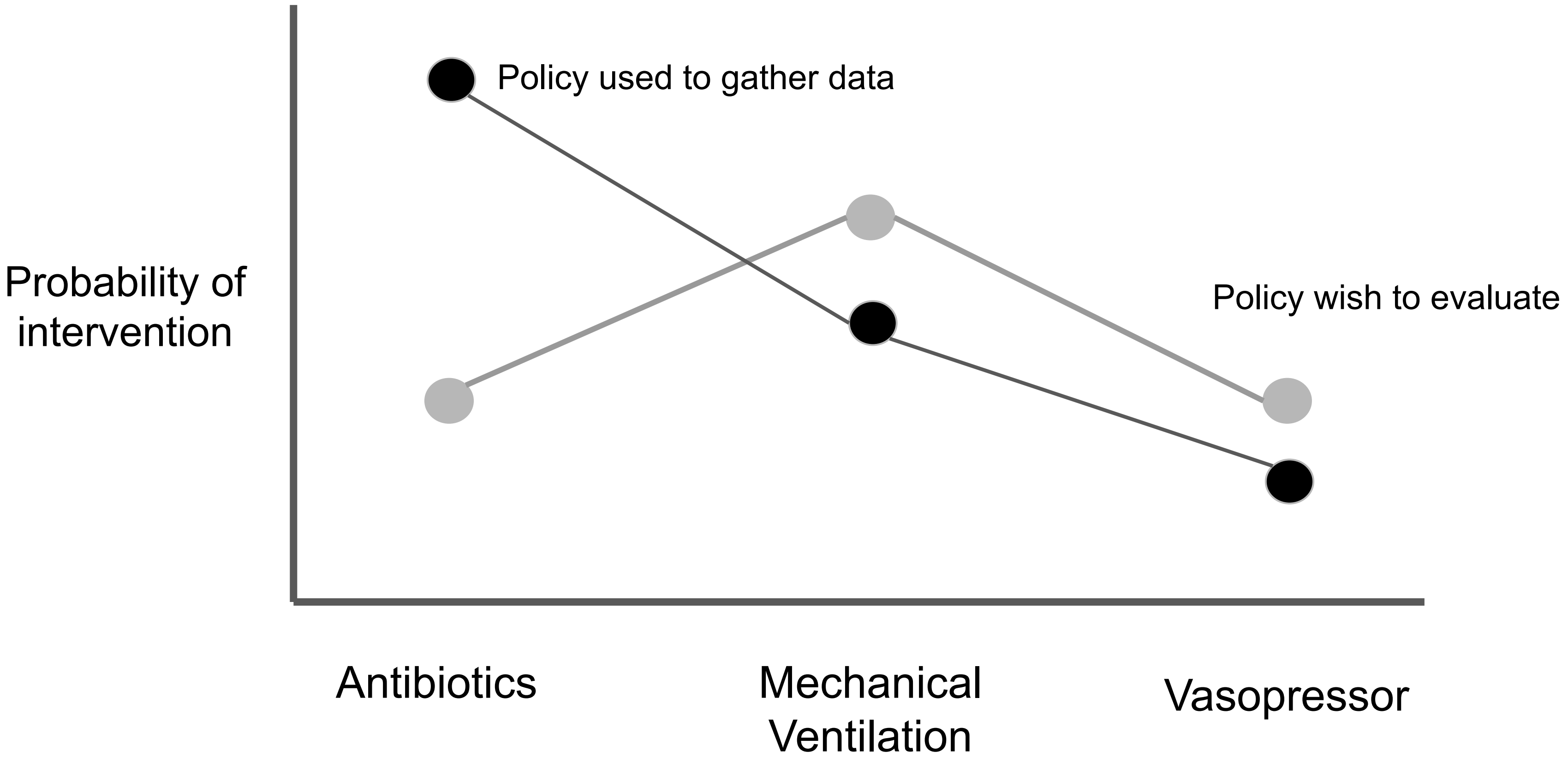
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

π: Policy mapping $s \to a$

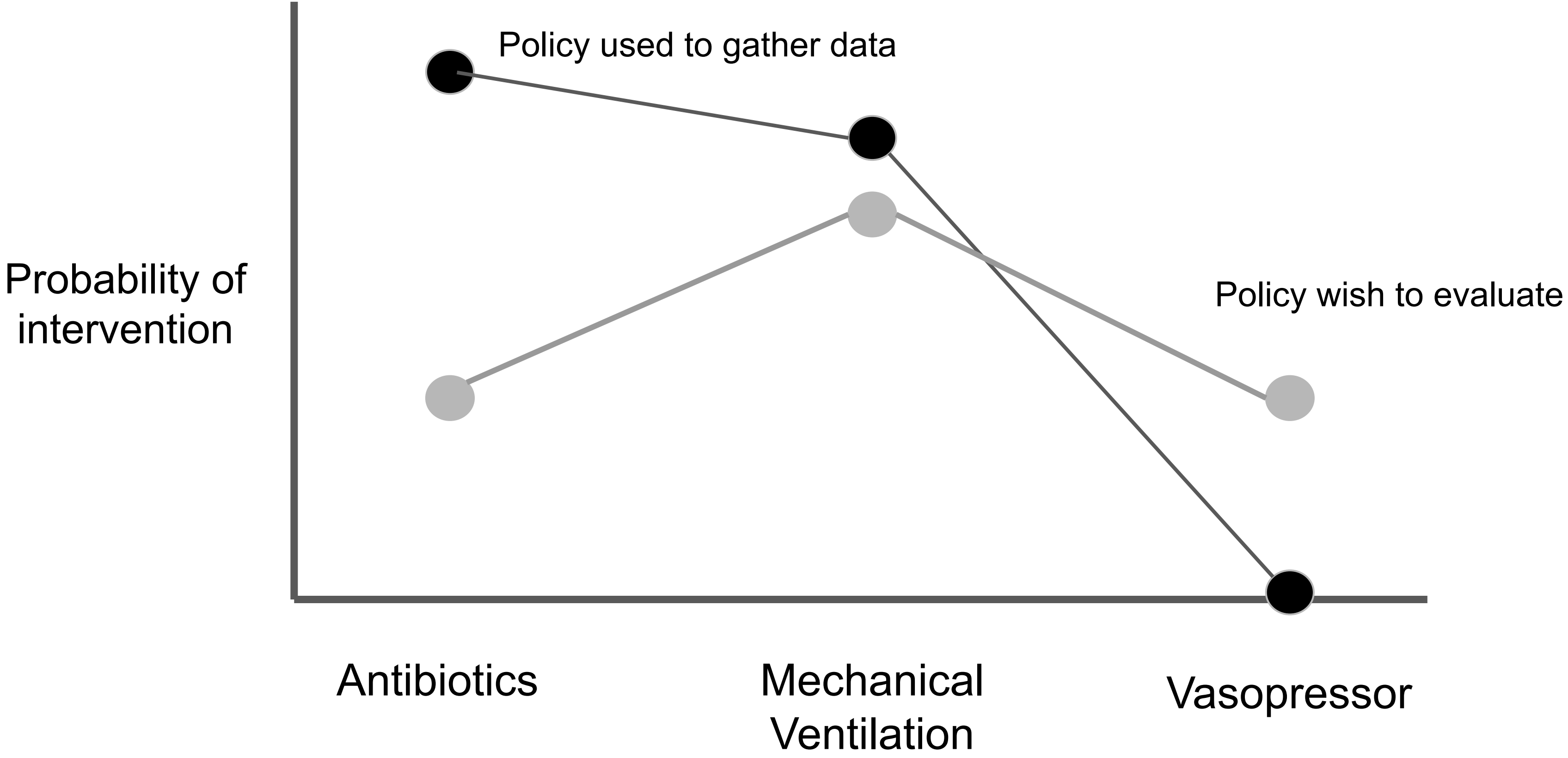
S_0: Set of initial states

$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Overlap Requirement: Data Must Support Policy Wish to Evaluate



No Overlap for Vasopressor⇒ Can’t Do Off Policy Estimation for Desired Policy



Limitations of Prior Work

- Typically assume overlap
    - Off policy estimation: for policy of interest
    - Off policy optimization: for all policies including optimal one (see concentrability assumption in batch RL)
- Unlikely to be true in many settings
- Many real datasets don’t include complete random exploration

Limitations of Prior Work

- Typically assume overlap
  - Off policy estimation: for policy of interest
  - Off policy optimization: for all policies including optimal one (see concentrability assumption in batch RL)
- Unlikely to be true in many settings
- Many real datasets don't include complete random exploration
- Assuming overlap when it's not there can be a problem:
  - We can end up with a policy with estimated high performance, but actually does poorly when deployed

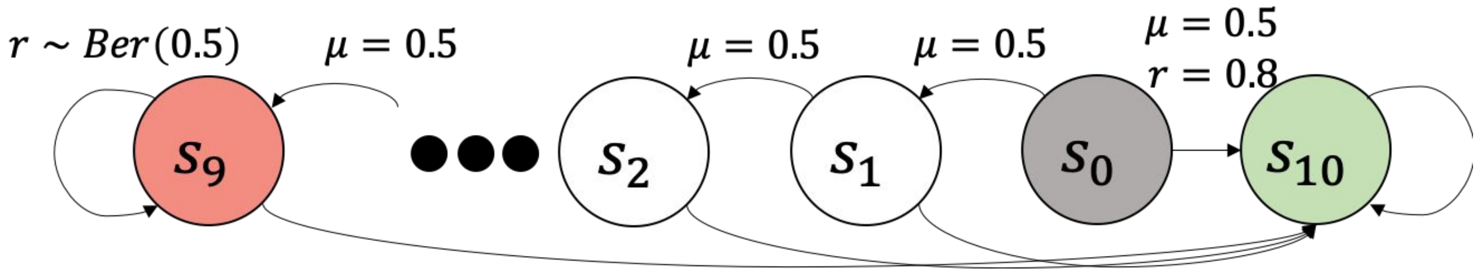
Doing the Best with What We've Got: Off Policy Optimization Without Full Data Coverage

- Idea: restrict off policy optimization to those with overlap in data
- Computationally tractable algorithm
- Simple idea: assume **pessimistic outcomes** for areas of state--action space with insufficient overlap/support

Common challenge that's attracted substantial interest in last few years but...

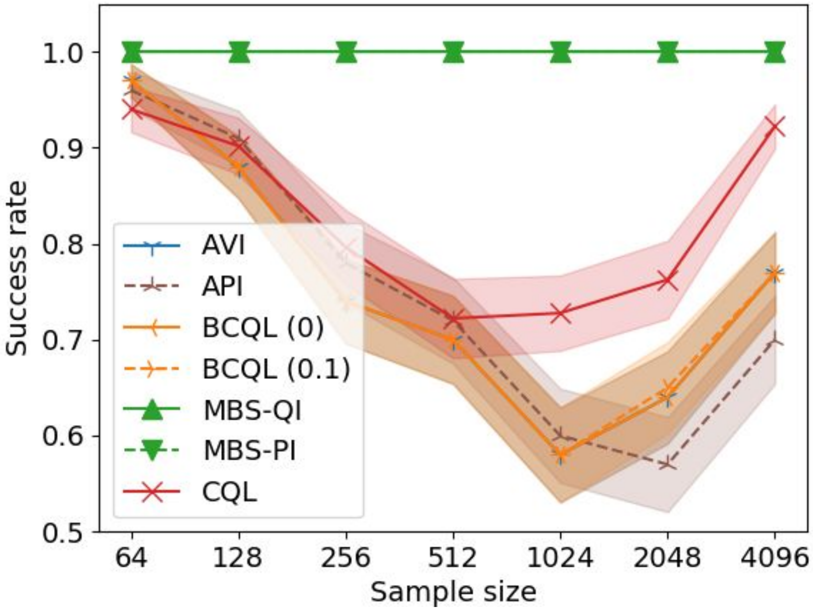
Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Illustrative Examples

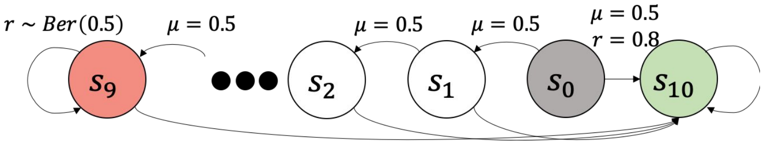


Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Recent Conservative Batch Reinforcement Learning Are Insufficient



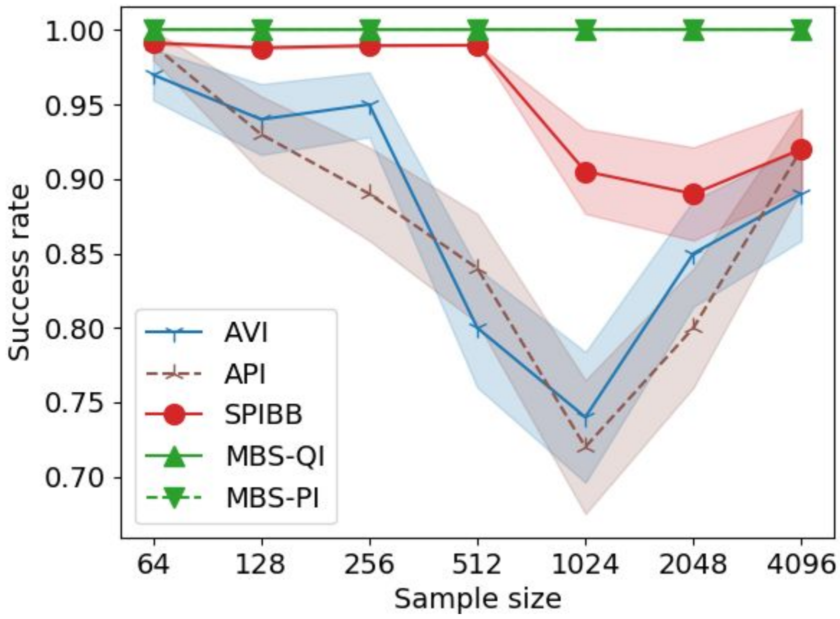
Success rate: #(getting the optimal policy)/#(trials)



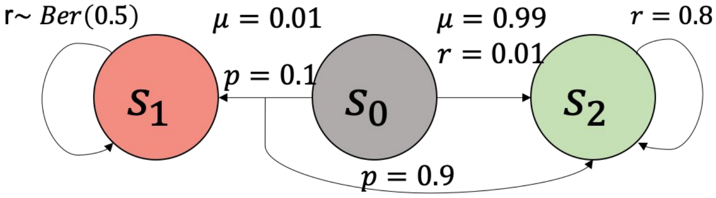
Reasons why baselines fail:

- Many baselines focus on penalty/constraints that are based on dist($\boldsymbol{\pi}$(a|s), $\boldsymbol{\pi}$_b(a|s)).
- In this example a sequence of large action conditional probabilities leads to a rare state.
- Due to finite samples, estimates of the reward of this rare state can be overestimated.

Recent Conservative Batch Reinforcement Learning Are Insufficient



Success rate: #(getting the optimal policy)/#(trials)



Reasons why baselines fail:

- SPIBB adds conservatism based on estimates of **π**_b & V of **π**_b.
- In this example, the actions which is rare under **π**_b also have a stochastic transition and reward, thus the **π**_b’s V is overestimated.

Idea: Use pessimistic value for state-action space with insufficient data

- Filtration function:

$$\zeta(s, a; \hat{\mu}, b) = 1(\hat{\mu}(s, a) > \mathrm{b})$$

Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Idea: Use pessimistic value for state-action space with insufficient data

- Filtration function:

$$\zeta(s, a; \hat{\mu}, b) = 1(\hat{\mu}(s, a) > \mathrm{b})$$

b can account for statistical uncertainty due to finite samples

Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Idea: Use pessimistic value for state-action space with insufficient data

- Filtration function:

$$\zeta(s, a; \hat{\mu}, b) = 1(\hat{\mu}(s, a) > \mathrm{b})$$

- Bellman operator and Bellman evaluation operator:

$$\mathcal{T} f(s, a) = r(s, a) + \gamma \mathbb{E}_{s'} \left[\max_{a'} \zeta(s', a') f(s', a') \right]$$

Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Idea: Use pessimistic value for state-action space with insufficient data

- Filtration function:

$$\zeta(s,a;\hat{\mu},b) = 1(\hat{\mu}(s,a) > \mathrm{b})$$

- Bellman operator and Bellman evaluation operator:

$$\mathcal{T}f(s,a) = r(s,a) + \gamma\mathbb{E}_{s'}\underbrace{\left[\max_{a'} \zeta(s',a')f(s',a')\right]}$$

⇒ = 0 for (s’,a’) with insufficient data.
We assume r(s,a) >= 0
Therefore pessimistic estimate for such tuples

Idea: Use pessimistic value for state-action space with insufficient data

- Filtration function:

$$\zeta(s, a; \hat{\mu}, b) = 1(\hat{\mu}(s, a) > \mathrm{b})$$

- Bellman operator and Bellman evaluation operator:

$$\mathcal{T} f(s, a) = r(s, a) + \gamma \mathbb{E}_{s'} \left[\max_{a'} \zeta(s', a') f(s', a') \right]$$

$$\mathcal{T}^{\pi} f(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P, a' \sim \pi} [\zeta(s', a') f(s', a')]$$

Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Marginalized Behavior Supported (MBI) Policy Optimization

- Filtration function:

$$\zeta(s, a; \hat{\mu}, b) = 1(\hat{\mu}(s, a) > \mathrm{b})$$

- Bellman operator and Bellman evaluation operator:

$$\mathcal{T} f(s, a) = r(s, a) + \gamma \mathbb{E}_{s'} \left[\max_{a'} \zeta(s', a') f(s', a') \right]$$

$$\mathcal{T}^{\pi} f(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P, a' \sim \pi} [\zeta(s', a') f(s', a')]$$

Majority of Past Model-Free Batch RL Theory for Function Approximation Setting

Assume for any *v*(s,a) distribution possible under some policy in this MDP

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \frac{\nu(s,a)}{\mu(s,a)} \leq C.$$

$$\Downarrow$$

$$V^* - V^{\pi_{\mathcal{A}}} \leq \epsilon$$

Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Best in Well Supported Policy Class*

Assume for any *v*(s,a) distribution possible under some policy in this MDP

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \frac{\nu(s,a)}{\mu(s,a)} \leq C.$$

$$\Downarrow$$

$$V^{*} - V^{\pi_{\mathcal{A}}} \leq \epsilon$$

$$\Rightarrow$$

Define

$$\Pi_{all} : \pi \ \ s.t.$$

$$\mathbb{E}_{s,a \sim \eta^{\pi}} \left[\mathbb{1} \left(\zeta(s,a) = 0 \right) \right] \leq \epsilon_{\zeta}$$

$$\Downarrow$$

$$\max_{\pi' \in \Pi_{all}} V^{\pi'} - V^{\pi_{\mathcal{A}}} \leq \epsilon$$

*Note: Policy set $\mathbf{\Pi}_{\text{all}}$ is not constructed, but implicitly our algorithm only considers elements in it

Assumption 1 (Bounded densities). *For any non-stationary policy* π *and* $h \geq 0$, $\eta_h^\pi(s,a) \leq U$.

Assumption 2 (Density estimation error). *With probability at least* $1-\delta$, $\|\widehat{\mu} - \mu\|_{TV} \leq \epsilon_\mu$.

Assumption 3 (Completeness under $\widetilde{\mathcal{T}}^\pi$). $\forall \pi \in \Pi$, $\max_{f \in \mathcal{F}} \min_{g \in \mathcal{F}} \|g - \widetilde{\mathcal{T}}^\pi f\|_{2,\mu}^2 \leq \epsilon_{\mathcal{F}}$.

Assumption 4 (Π Completeness). $\forall f \in \mathcal{F}$, $\min_{\pi \in \Pi} \|\mathbb{E}_\pi [\zeta \circ f(s,a)] - \max_a \zeta \circ f(s,a)\|_{1,\mu} \leq \epsilon_\Pi$.

$\eta_h^\pi(s) := \Pr[s_h = s|\pi]$,

$\eta_h^\pi(s,a) = \eta_h^\pi(s)\pi(a|s)$

$\zeta(s,a;\widehat{\mu},b) = \mathbb{1}\left(\widehat{\mu}(s,a) \geq b\right)$

Theoretical Result

We bound the error w.r.t. the best policy in the following policy set:

{all policies such that $\mathrm{Pr}(\zeta(s,a)=0|\pi) \leq \epsilon_{\zeta}$}

Error bounds [1]:

- PI:

$$O\left(\frac{\mathrm{V}_{\max}}{(1-\gamma)^3 b}\sqrt{\frac{\ln(|\mathcal{F}||\Pi|/\delta)}{n}}\right) + \frac{\mathrm{V}_{\max}\epsilon_{\zeta}}{1-\gamma}$$

- VI [2]:

$$O\left(\frac{\mathrm{V}_{\max}}{(1-\gamma)^2 b}\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{n}}\right) + \frac{\mathrm{V}_{\max}\epsilon_{\zeta}}{1-\gamma}$$

[1] We omit some constant terms that is same as standard ADP analysis with function approximation.

[2] For VI results there is another important constant term, see our paper for detailed result and discussion.

$\zeta(s,a;\widehat{\mu},b) = \mathbb{1}\left(\widehat{\mu}(s,a) \geq b\right)$

Theoretical Result

We bound the error w.r.t. the best policy in the following policy set:

{all policies such that $\Pr(\zeta(s,a) = 0|\pi) \leq \epsilon_{\zeta}$}

Note: Results are for function approximation, finite sample setting

Error bounds [1]:

- PI:

$$O\left(\frac{V_{\max}}{(1-\gamma)^3 b}\sqrt{\frac{\ln(|\mathcal{F}||\Pi|/\delta)}{n}}\right) + \frac{V_{\max}\epsilon_{\zeta}}{1-\gamma}$$

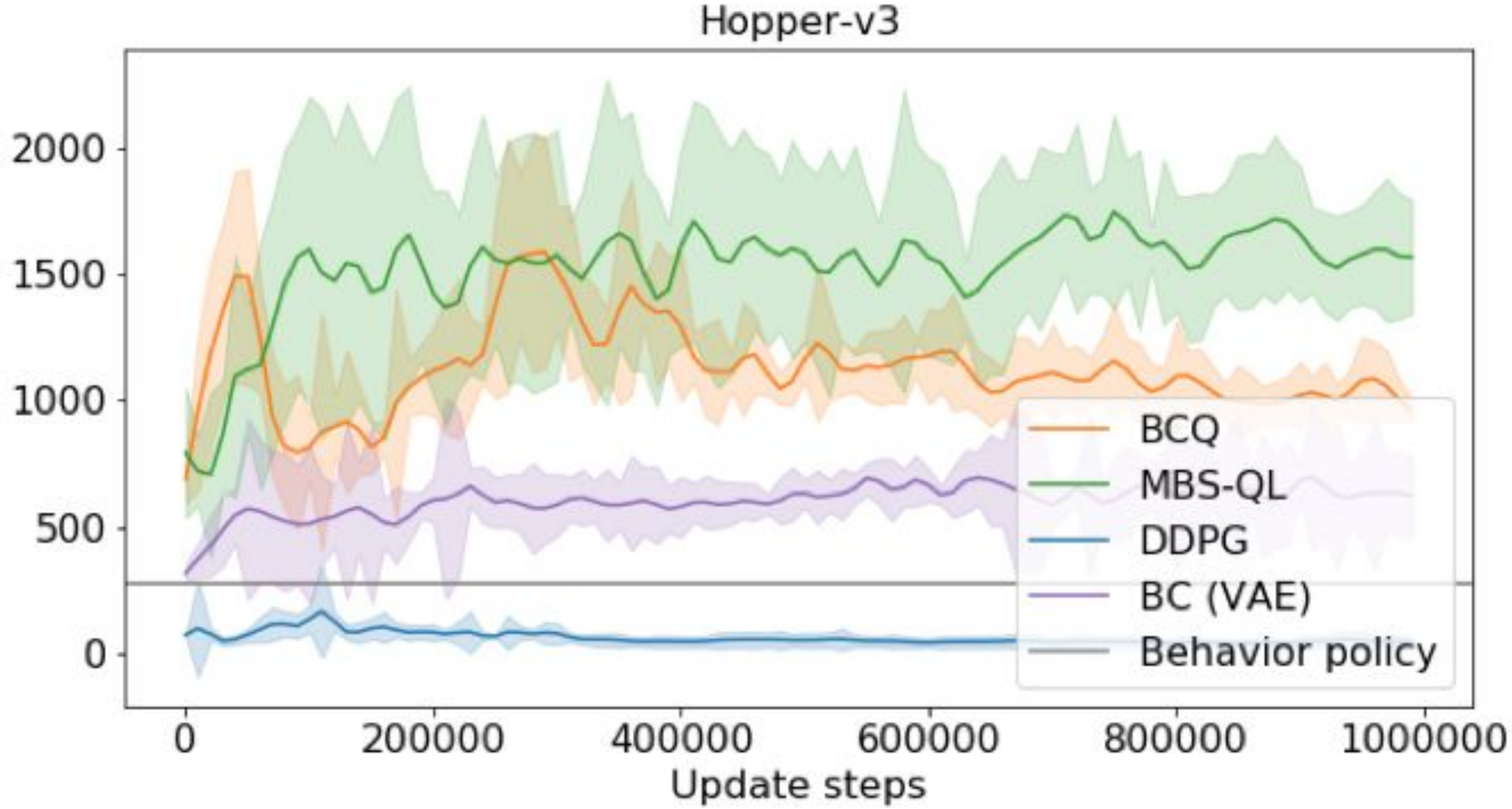
- VI [2]:

$$O\left(\frac{V_{\max}}{(1-\gamma)^2 b}\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{n}}\right) + \frac{V_{\max}\epsilon_{\zeta}}{1-\gamma}$$

[1] We omit some constant terms that is same as standard ADP analysis with function approximation.

[2] For VI results there is another important constant term, see our paper for detailed result and discussion.

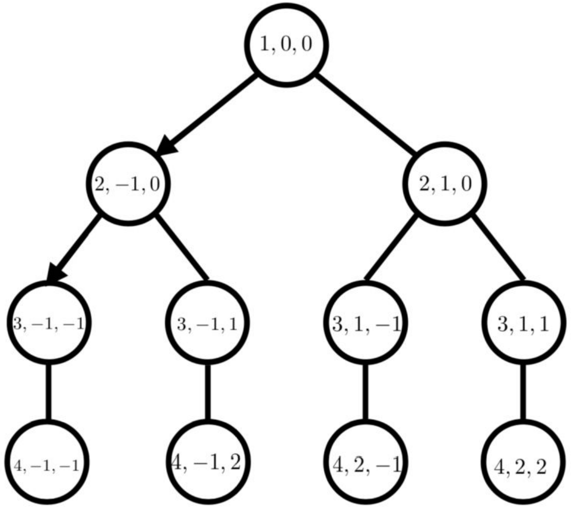
Can Do Get Substantially Better Solutions, With Same Data



Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

This Was Model Free. Might Models Be Even Better?

- Model based approaches can be provably more efficient than model free value function for *online* evaluation or control



Sun, Jiang, Krishnamurthy, Agarwal, Langford COLT 2019

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t \ ,$$

$$V^K(x) := \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} (x_t^\top Q x_t + u_t^\top R u_t - \lambda_K) \,\middle|\, x_0 = x\right]$$

Tu & Recht COLT 2019

Concurrent Work on Conservative Model-Based Offline Batch Reinforcement Learning

- Ex. Yu, Thomas, Yu, Ermon, Zou, Levine, Finn & Ma (NeurIPS 2020) and Kidambi, Rajeswaran, Netrapalli & Joachims (NeurIPS 2020)
- Learn a model and penalize model uncertainty during planning
- Empirically very promising on D4RL tasks
- Their work has more limited theoretical analysis

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Early Comparison with Concurrent Work

	MBS-BCQ	**MBS-BEAR**	BCQ	BEAR	MOPO	CQL
Hopper-medium	75.9	32.3	54.5	52.1	26.5	58.0

Early Comparison with Concurrent Work

	MBS-BCQ	**MBS-BEAR**	BCQ	BEAR	MOPO	CQL
Hopper-medium	75.9	32.3	54.5	52.1	26.5	58.0
HalfCheetah-medium	38.4	39.7	40.7	41.7	40.2	44.4
Walker2d-medium	64.4	75.4	53.1	59.1	14.0	79.2

Early Comparison with Concurrent Work

	MBS-BCQ	**MBS-BEAR**	BCQ	BEAR	MOPO	CQL
Hopper-medium	75.9	32.3	54.5	52.1	26.5	58.0
HalfCheetah-medium	38.4	39.7	40.7	41.7	40.2	44.4
Walker2d-medium	64.4	75.4	53.1	59.1	14.0	79.2

- Preliminary draft results: on some D4RL recent model-based pessimistic approaches or CQL do better
- In general suspect recent model-based approaches will dominate our MBS empirically but our theoretical results are stronger
- Interesting to see further theoretical work on model based approaches

Pessimistic Model-Free Batch/Offline Policy Learning

- Restrict off policy optimization to those with overlap in data
- Computationally tractable algorithm
- **Simple idea: assume pessimistic outcomes for areas of state--action space with insufficient overlap/support**
- Theoretical results bound distance to best supported policy
    - Considers finite sample & function approximation
- Model free value function method

⇒ Pessimism under uncertainty has received a lot of attention in last 1-2 years for offline RL

Today

1. Imitation vs batch/offline RL policy learning
2. Fitted Q Iteration / Offline Q Learning
3. Pessimism
4. **Case Study**

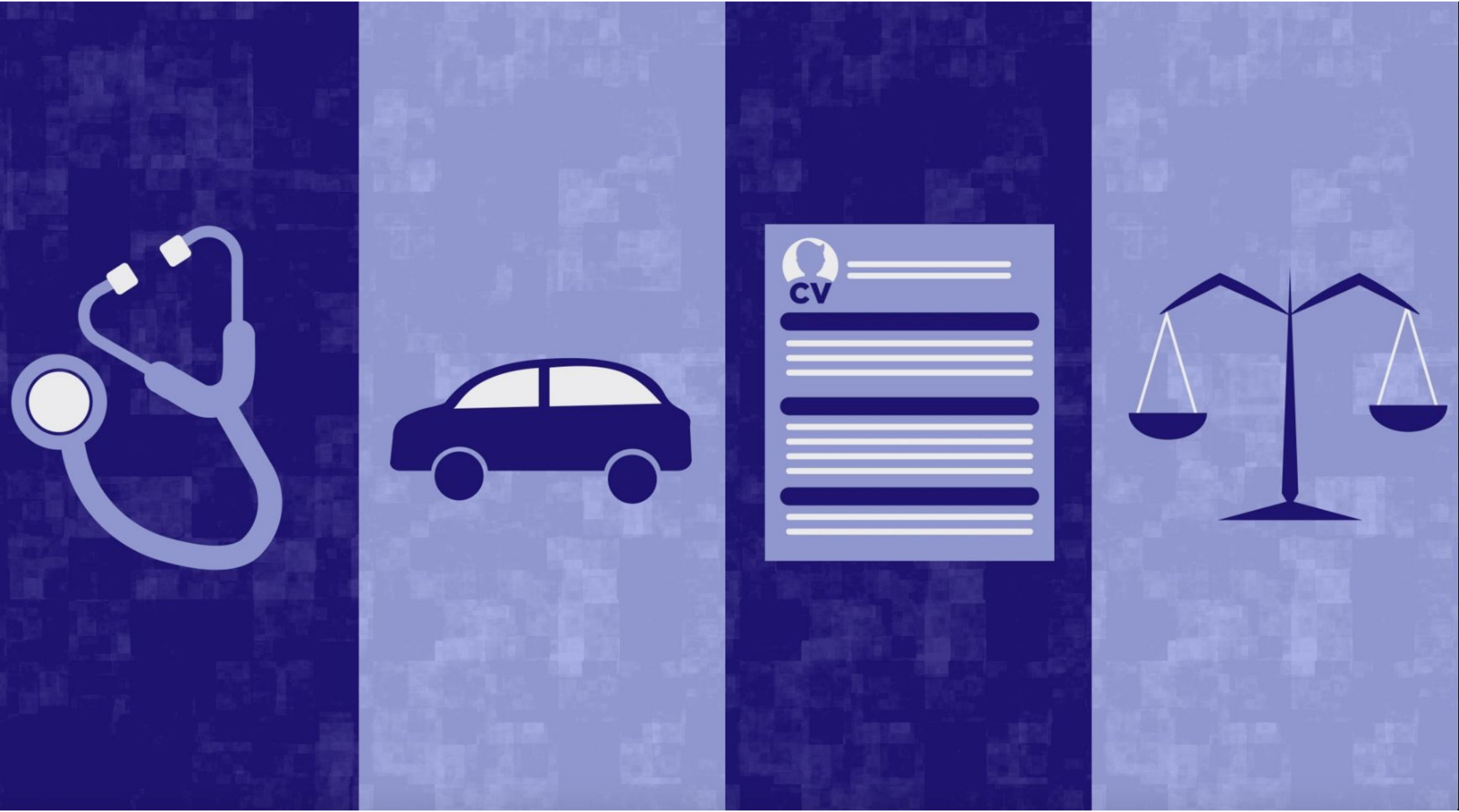
RESEARCH

COMPUTER SCIENCE

Preventing undesirable behavior of intelligent machines

Philip S. Thomas[1]*, Bruno Castro da Silva[2], Andrew G. Barto[1], Stephen Giguere[1], Yuriy Brun[1], Emma Brunskill[3]

Science November 2019

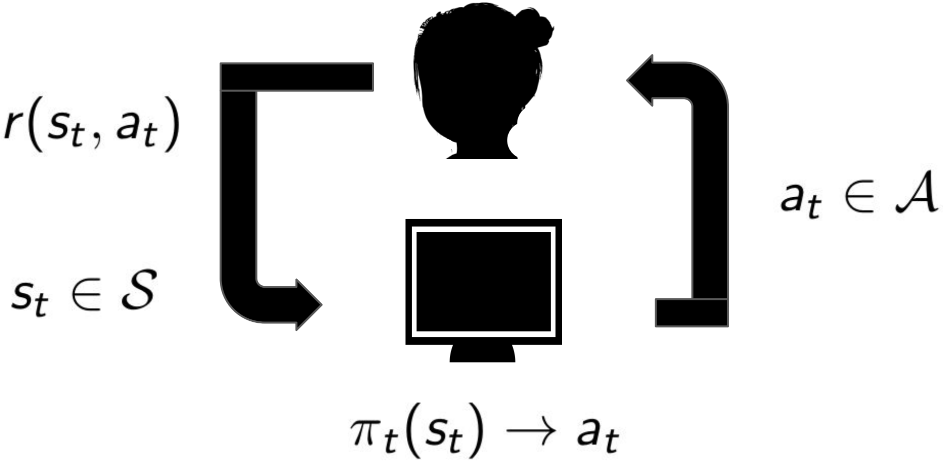

CV

Optimizing while Ensuring Solution Won’t, in the Future, Exhibit Undesirable Behavior

$$\underset{a \in \mathcal{A}}{\arg\max}\ f(a)$$

$$\text{s.t} \quad \text{s.t. } \forall i \in \{1, \ldots, n\}, \Pr\left(\underbrace{g_i(a(D)) \leq 0}_{\text{Constraints}}\right) \geq 1 - \delta_i$$

Counterfactual RL
with Constraints on Future Performance of Policy



$r(s_t, a_t)$

$s_t \in \mathcal{S}$

$a_t \in \mathcal{A}$

$\pi_t(s_t) \rightarrow a_t$

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

Related Work in Decision Making

$$\underset{a \in \mathcal{A}}{\arg\max}\; f(a)$$

$$\text{s.t. } \forall i \in \{1, \ldots, n\},\ \Pr\Big(g_i(a(D)) \leq 0\Big) \geq 1 - \delta_i$$

- Chance constraints, data driven robust optimization have similar aims
- Most of this work has focused on ensuring computational efficiency for *f* and/or constraints *g* with certain structure (e.g. convex)
- Also need to be able to capture broader set of aims & constraints

Batch RL with Safety Constraints

$$g(\theta) = \mathbf{E}[r'(H)|\underset{\substack{\uparrow\\ \text{Default policy}}}{\theta_0}] - \mathbf{E}[r'(H)|\underset{\substack{\uparrow\\ \text{Potential policy}}}{\theta}]$$

- r’(H) is a function of the trajectory H

Aside: Importance Sampling Estimators Unbiased for Policy Evaluation

$$V^{\pi}(s) \quad = \quad \sum_{i=1,\tau_i\sim\pi_b}^{N} R_{\tau_i} \prod_{t=1}^{H_i} \frac{p(a_{it}|\pi, s_{it})}{p(a_{it}|\pi_b, s_{it})}$$

- Using for policy optimization directly has challenges due to variance (e.g. Doroudi, Thomas, Brunskill UAI 2017)
- But can also be successful, especially with variants to reduce variance that yield lower bounds (e.g. Thomas et al. UAI/ICML 2015; Futoma et al. 2020)

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

1 Algorithm for Batch RL with Safety Constraints

- Take in desired behavior constraints g and confidence level & data

1 Algorithm for Batch RL with Safety Constraints

- Take in desired behavior constraints g and confidence level & data
- Given a finite set of decision policies, for each policy i
    - **Compute generalization bound for each constraint**
    - If passes all with desired confidence*, Safe(i) = true

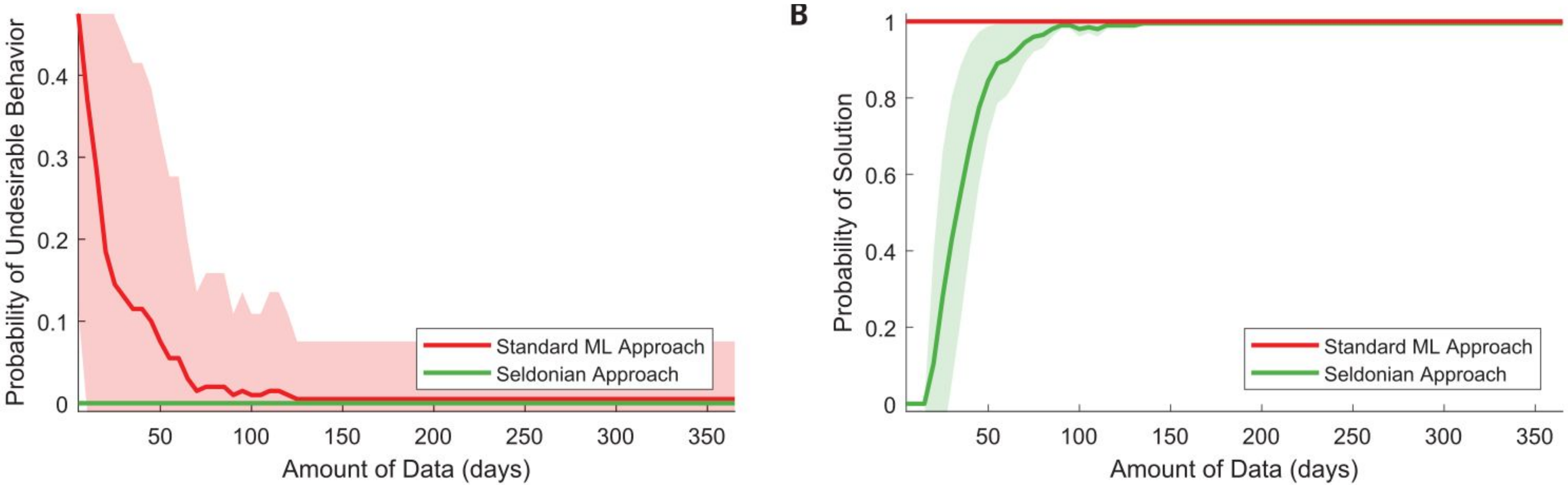
1 Algorithm for Batch RL with Safety Constraints

- Take in desired behavior constraints g and confidence level & data
- Given a finite set of decision policies, for each policy i
    - Compute generalization bound for each constraint
    - If passes all with desired confidence*, Safe(i) = true
- Estimate performance f of all policies that are safe
- Return best policy that is safe, or no solution if safe set is empty

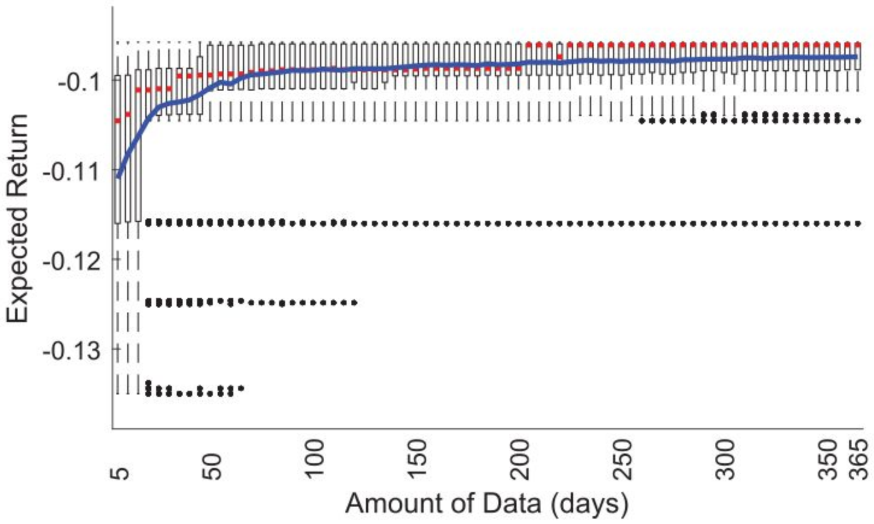
Diabetes Insulin Management

- Blood glucose control
- Action: insulin dosage
- Search over policies
- Constraint: hypoglycemia
- Very accurate simulator: approved by FDA to replace early stage animal trials

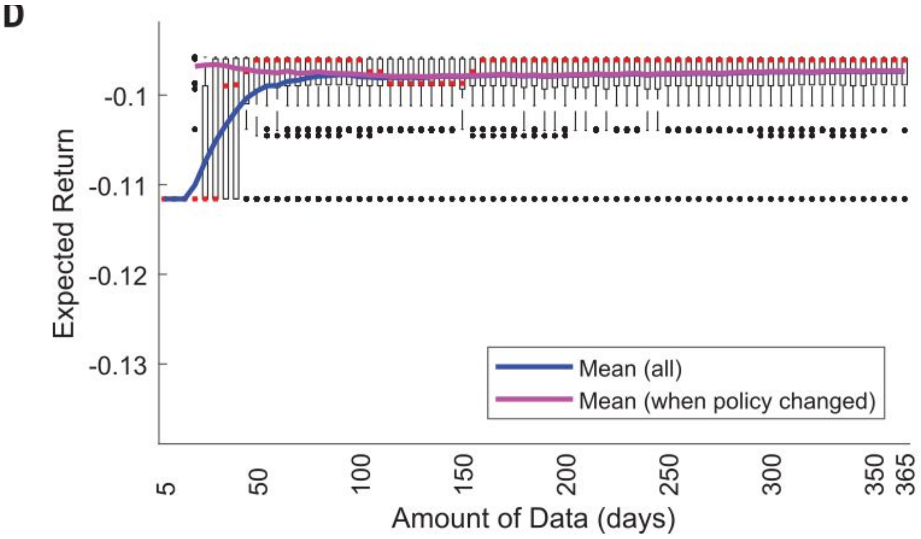
Personalized Insulin Dosage: Safe Batch Policy Improvement



Personalized Insulin Dosage: Quickly Can Have Confidence in Safe Better Policy



Standard RL



Our Safe Batch RL

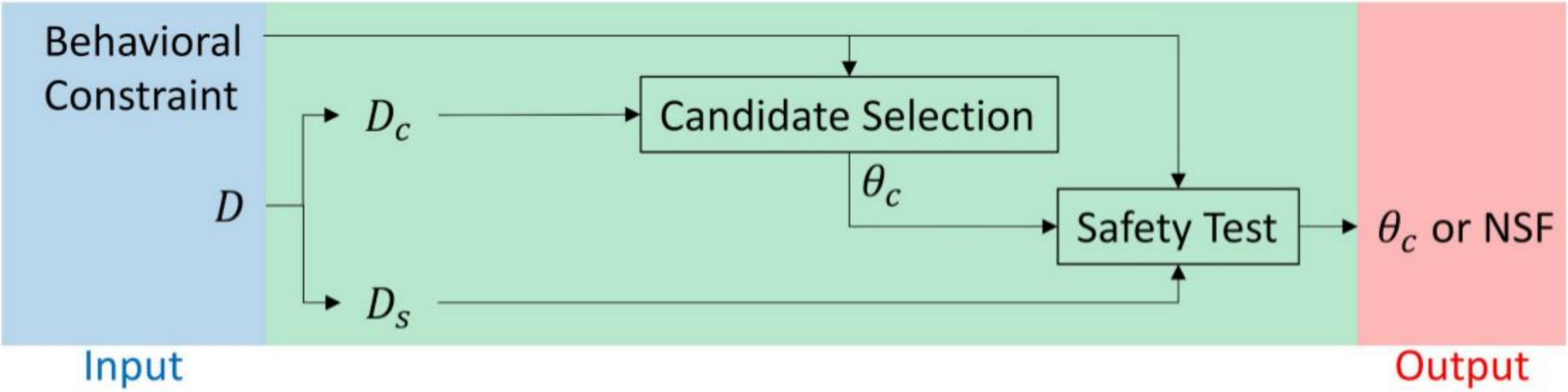
Often Want an Infinite Set of Decision Policies

- Take in desired behavior constraints g and confidence level & data
- Given a **finite set of decision policies**, for each policy i
    - Compute generalization bound for each constraint
    - If passes all with desired confidence*, Safe(i) = true
- Estimate performance f of all policies that are safe
- Return best policy that is safe, or no solution if safe set is empty

Offline Contextual Bandits with High Probability Fairness Guarantees

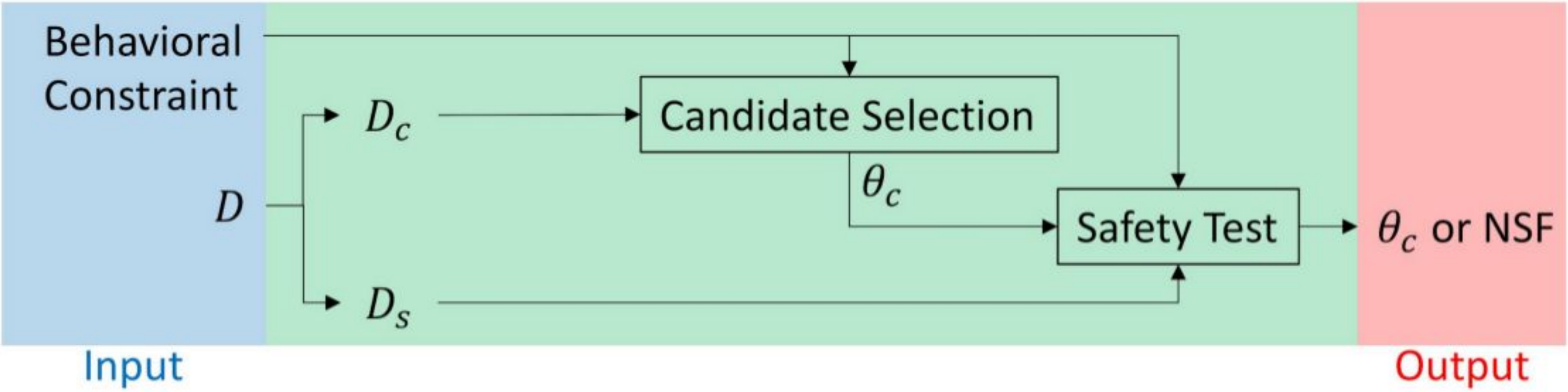
Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Offline Contextual Bandits with High Probability Fairness Guarantees



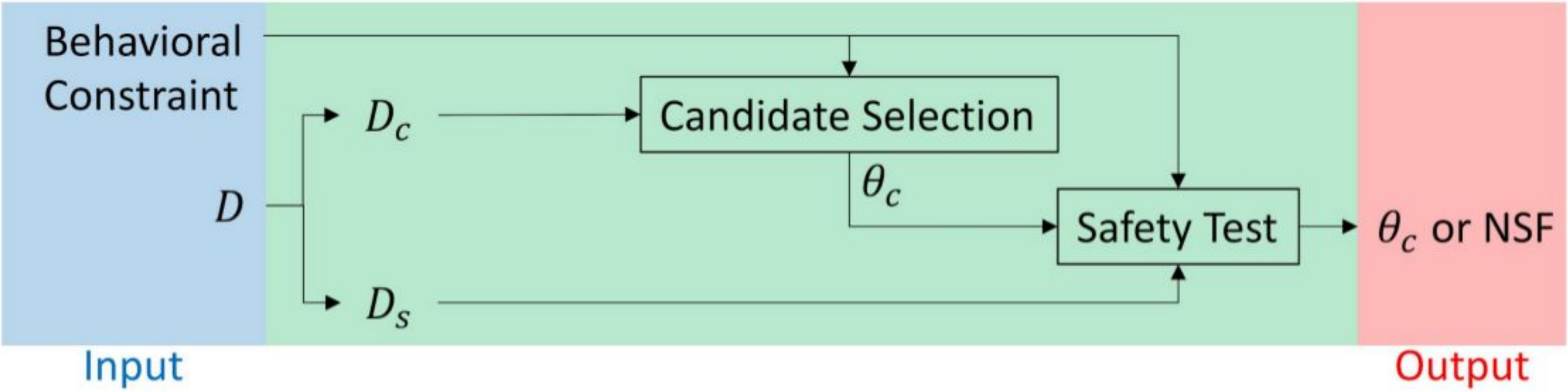
Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Infinite Policy Set, Union Bound Over Constraints for Each Fails



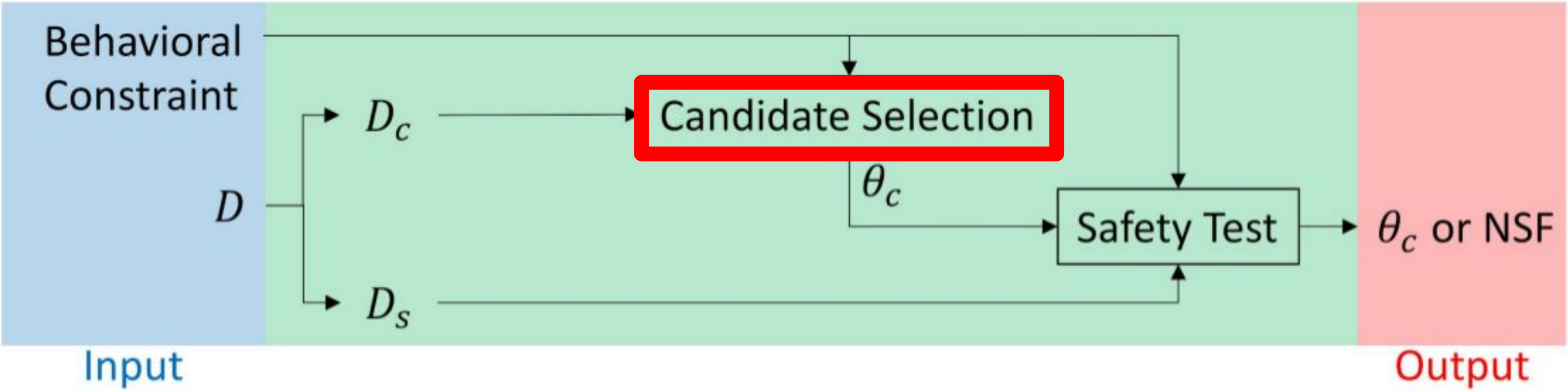
Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Idea 1: Split into 2 Datasets. Select Then Test



Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Idea 2: Do Policy Search So That Likely to Find a Good Policy that Will Satisfy Safety Test

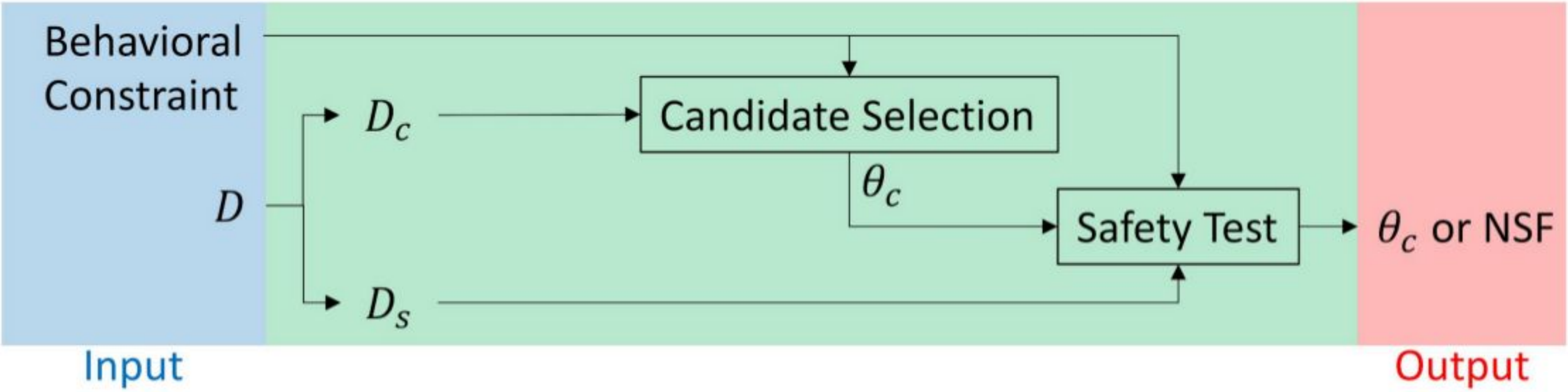


Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Algorithm 3 ComputeUpperBounds($\theta, D, \Delta, \hat{Z}, \mathcal{E}, \texttt{inflateBounds}$)

```
$\texttt{out} = [\,]$
for $i = 1, ..., k$ do
    $\hat{Z}_i = \{\hat{z}^i_j\}_{j=1}^{d_i} \subseteq \hat{Z}$
    $L_i, U_i = \texttt{Bound}(E_i, \theta, D, \delta_i / d_i, \hat{Z}_i, \texttt{inflateBounds})$
    $\texttt{out.append}(U_i)$
return $\texttt{out}$
```

Batch Contextual Bandit that Avoids Undesirable Behavior



Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Batch Fair Contextual Bandit Properties

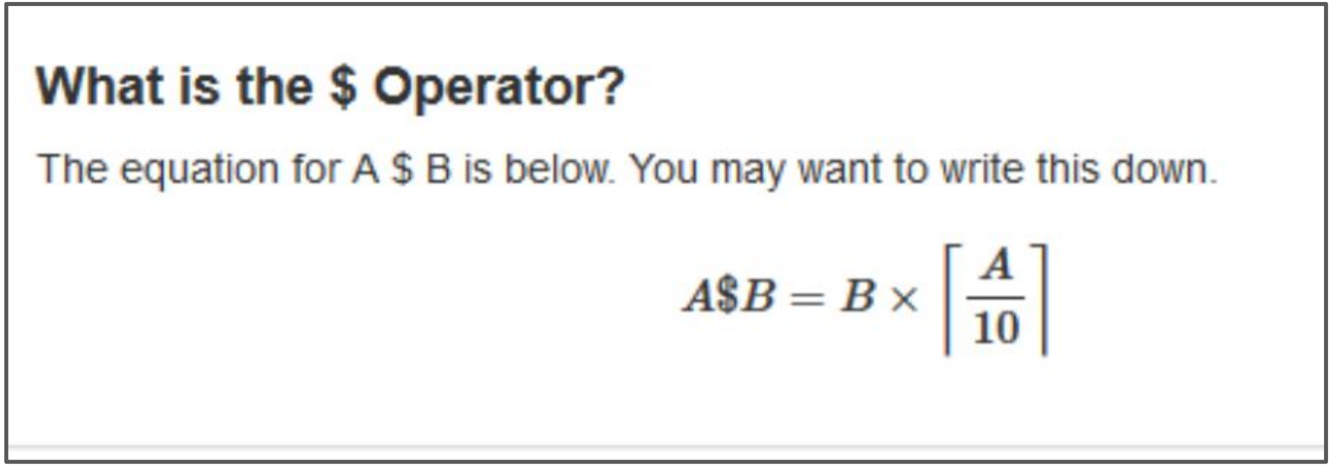
- Usable with a variety of fairness constraints: disparate impact, statistical parity, …
- As amount of data increases, probability that return a fair solution if one exists goes to 1

$$\lim_{|D|\to\infty} \Pr\Big(a(D) \neq \texttt{NSF}, g\big(a(D)\big) \leq 0\Big) = 1$$

Experiments

- Loan approval [Statlog German Credit data]
- Criminal recidivism [Propublica recidivism data]
- Simple sample tutoring experiment

Tutoring, with One Policy with Biased Decisions



What is the $ Operator?

The equation for A $ B is below. You may want to write this down.

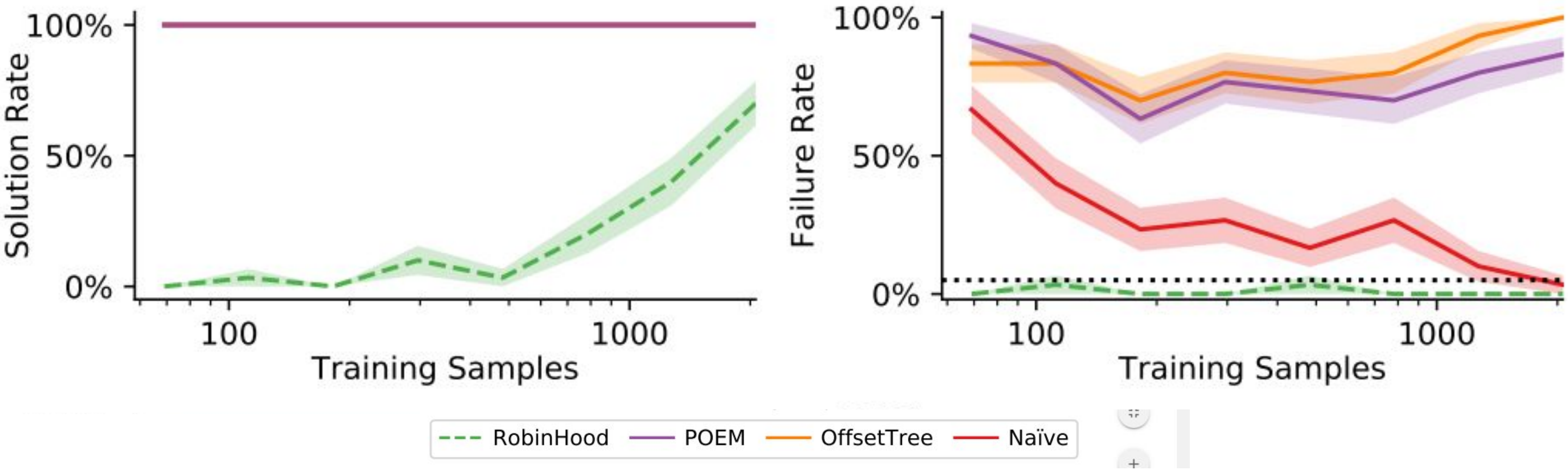
$$A\$B = B \times \left\lceil \frac{A}{10} \right\rceil$$

Fairness Constraints

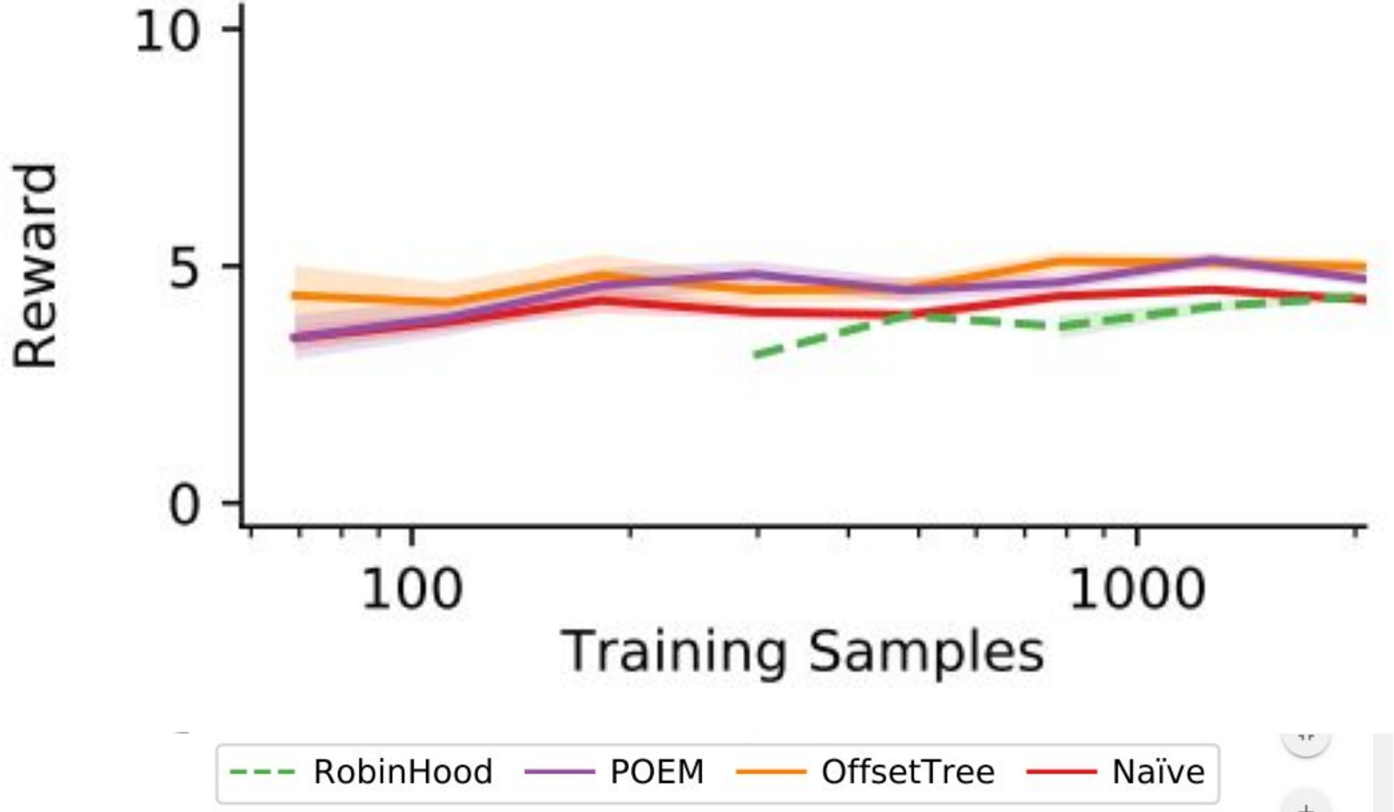
$$g_f(\theta) := \frac{1}{|F|}\sum_{\iota=1}^{|D|} R_\iota \mathbb{I}(\mathtt{f}) - \mathbf{E}[R_\iota|\mathtt{f}] - \epsilon_f$$

$$g_m(\theta) := \frac{1}{|M|}\sum_{\iota=1}^{|D|} R_\iota \mathbb{I}(\mathtt{m}) - \mathbf{E}[R_\iota|\mathtt{m}] - \epsilon_m$$

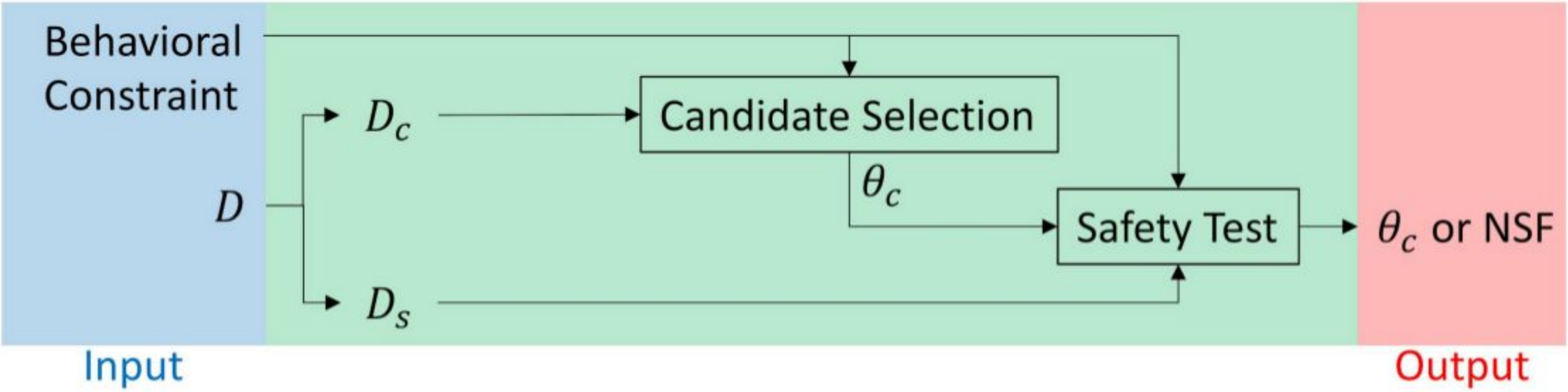
Data Needed and Constraint Satisfaction



Finds A High Performing Policy



Offline Contextual Bandits with High Probability Fairness Guarantees



Metevier, Giguere, Brockman, Kobren, Brun, Brunskill, Thomas NeurIPS 2019

Optimizing while Ensuring Solution Won’t, in the Future, Exhibit Undesirable Behavior

$$\arg\max_{a \in \mathcal{A}} f(a)$$

$$\text{s.t}\ \ \text{s.t.}\ \forall i \in \{1, \ldots, n\}, \Pr\Big(\underbrace{g_i(a(D)) \leq 0}_{\text{Constraints}}\Big) \geq 1 - \delta_i$$

⇒ Illustrated we can do this, for very general constraints, for several problems but many open questions around computational efficiency, other constraints …

What You Should Know

- Offline RL can do better than imitation learning / behavior cloning (Why?)
- Pessimism under uncertainty can be useful, particularly for high stakes applications
- Be able to give example application areas where offline RL might be useful

Where We Are In The Course

1. Learning from offline data
    a. Imitation learning
    b. Batch/offline policy evaluation
    c. Batch/offline policy learning
2. Next week
    a. Guest lecture: Maria Dimakopoulou
    b. Quiz