Batch / Offline RL Policy Evaluation & Optimization

Emma Brunskill

CS234

Spring 2024

Refresh Your Understanding

Select all that are true

does learn does not

- RLHF and DPO both learn an explicit representation of a reward model from preference data false
- Both are constrained to be at most as good as the best examples in the pairwise preference data false
- DPO does not use a reference policy false
- Not Sure

Refresh Your Understanding Solutions

Select all that are true

- RLHF and DPO both learn an explicit representation of a reward model from preference data
- Both are constrained to be at most as good as the best examples in the pairwise preference data
- DPO does not use a reference policy
- Not Sure

Class Outline

- Last time: Learning from Past Human Preferences, RLHF and DPO
- **Today: Learning from Past Decisions and Actions, Offline RL**
- Next time: Fast / Data efficient RL (and bandits, relevant to HW3)

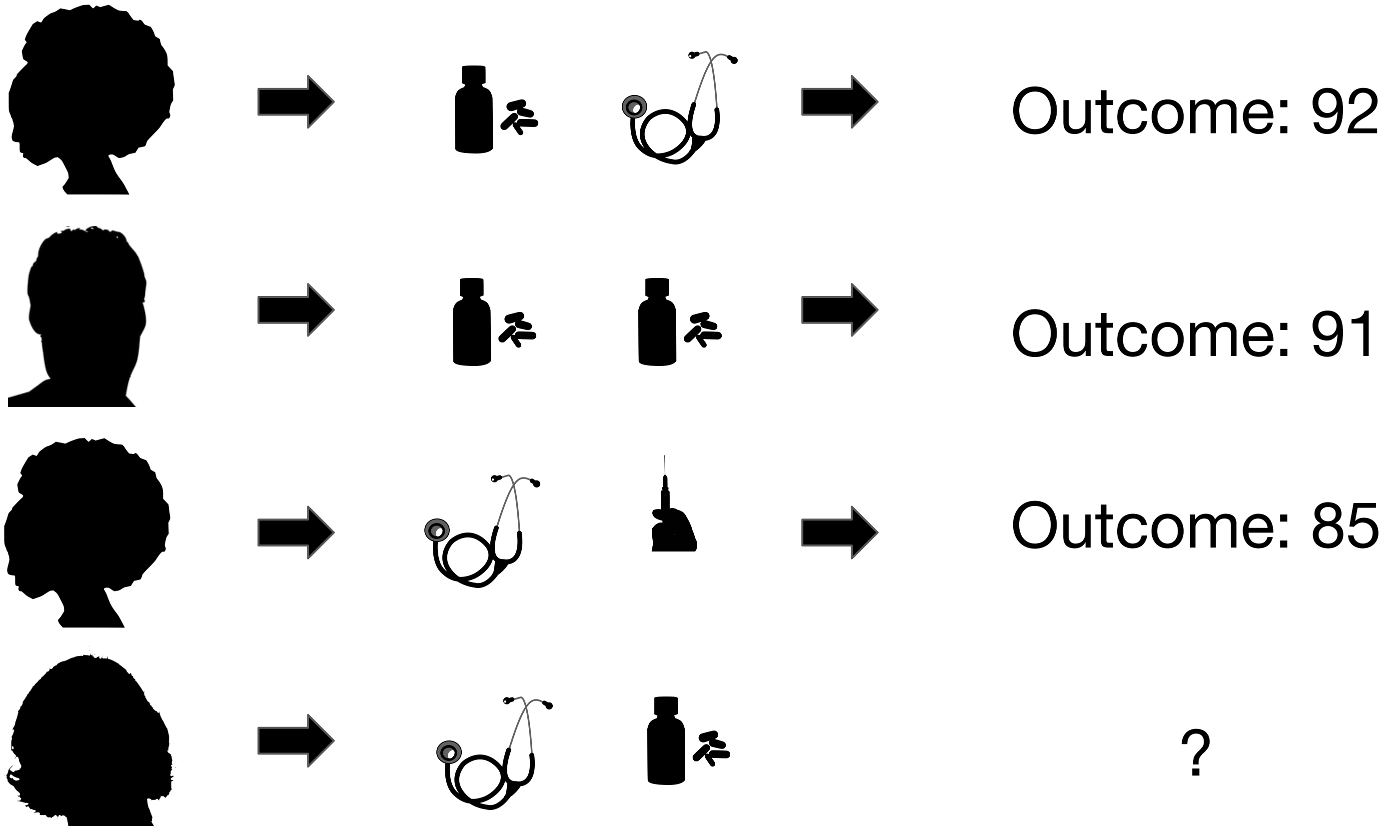
Learning from the Past

- Learning from Past Human Demonstrations: Imitation Learning
- Learning from Past Human Preferences: RLHF and DPO
- **Learning from Past Decisions and Actions: Offline RL**

Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
    a. Using models
    b. Using model free methods
    c. Use importance sampling
3. Offline policy learning / optimization

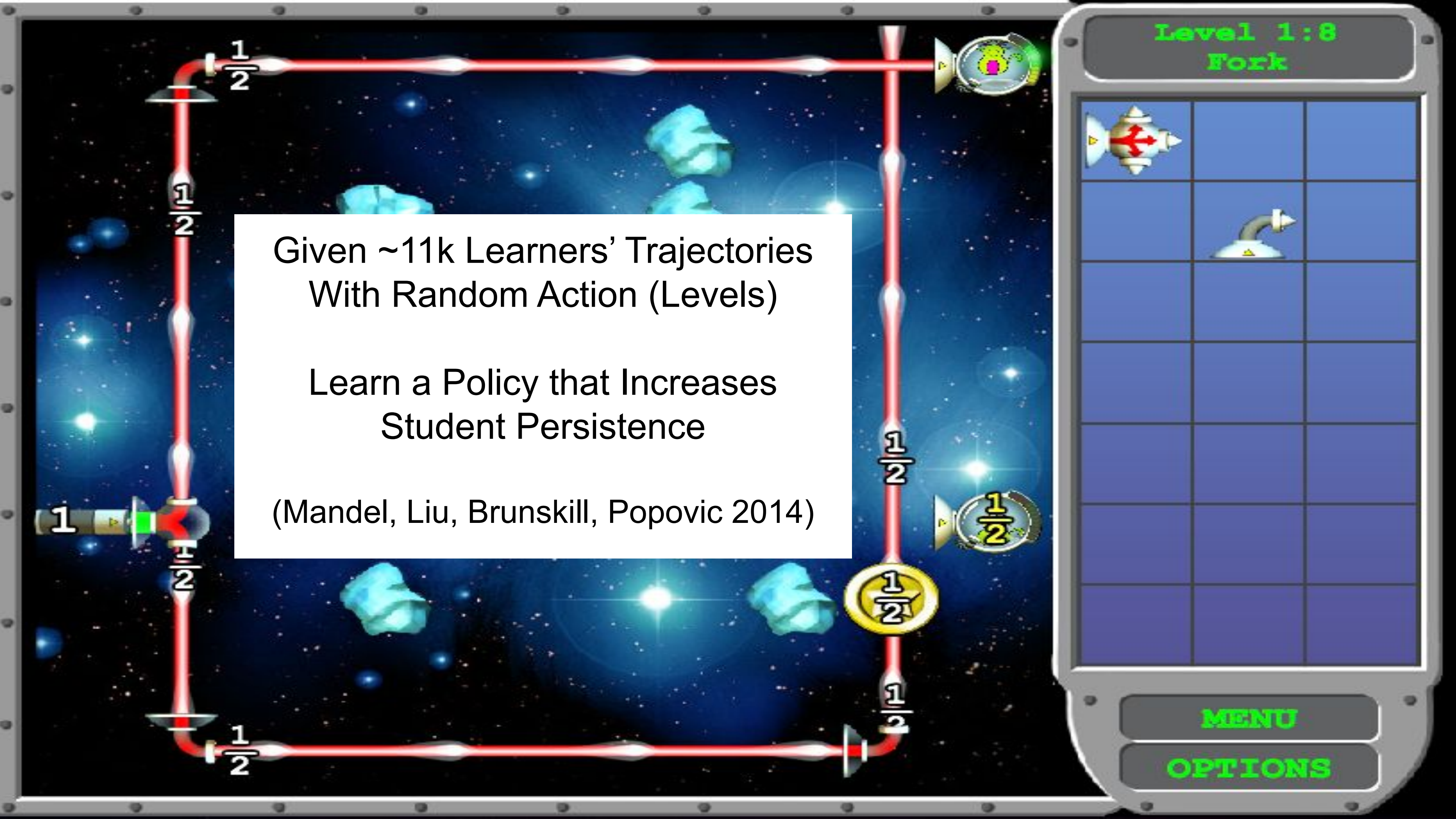
Can We Do Better than Imitation Learning?



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

Took > 30s

Took <= 30s



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

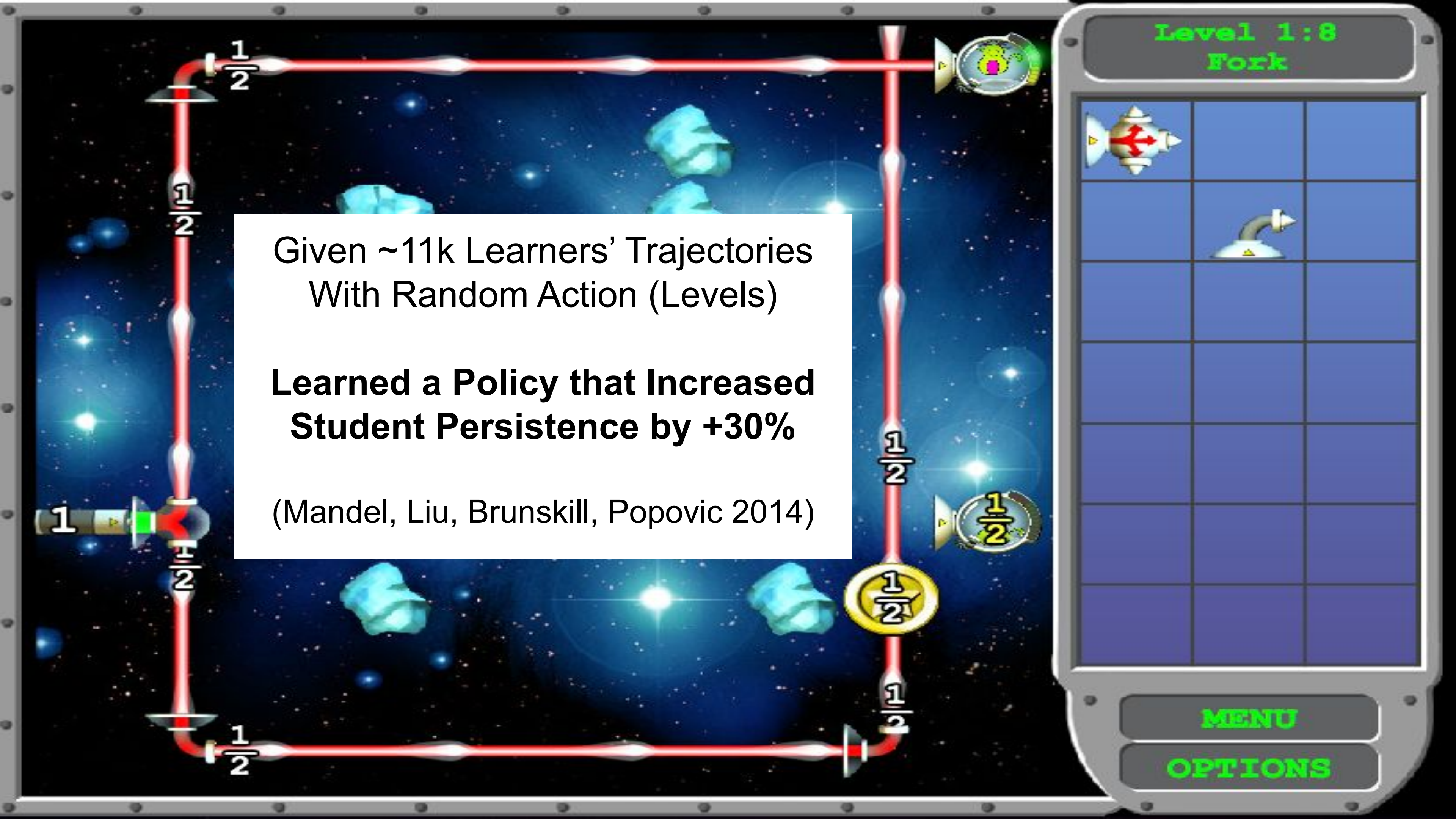
Learn a Policy that Increases
Student Persistence

(Mandel, Liu, Brunskill, Popovic 2014)

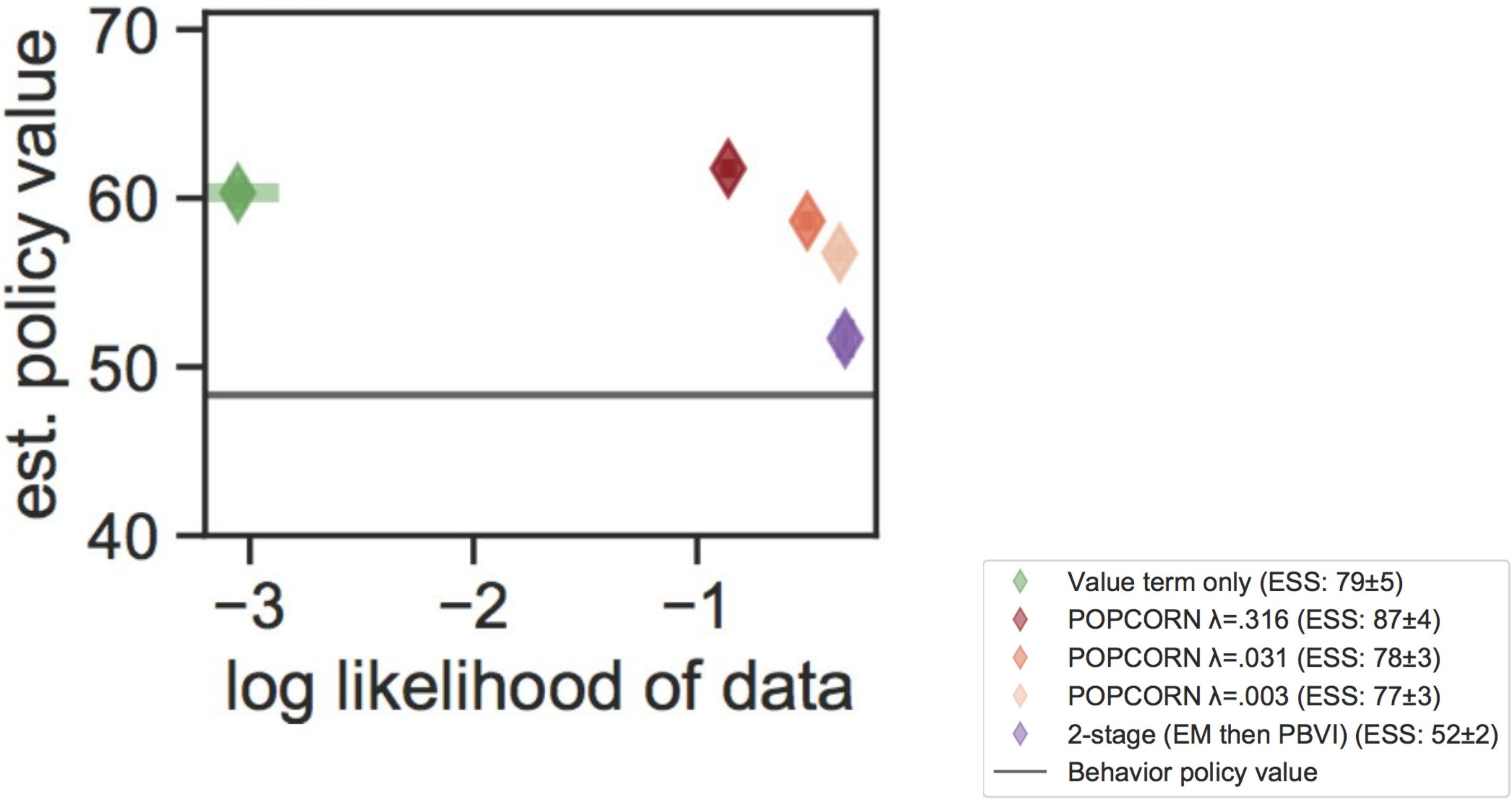
Given ~11k Learners' Trajectories
With Random Action (Levels)

Learned a Policy that Increased Student Persistence by +30%

(Mandel, Liu, Brunskill, Popovic 2014)

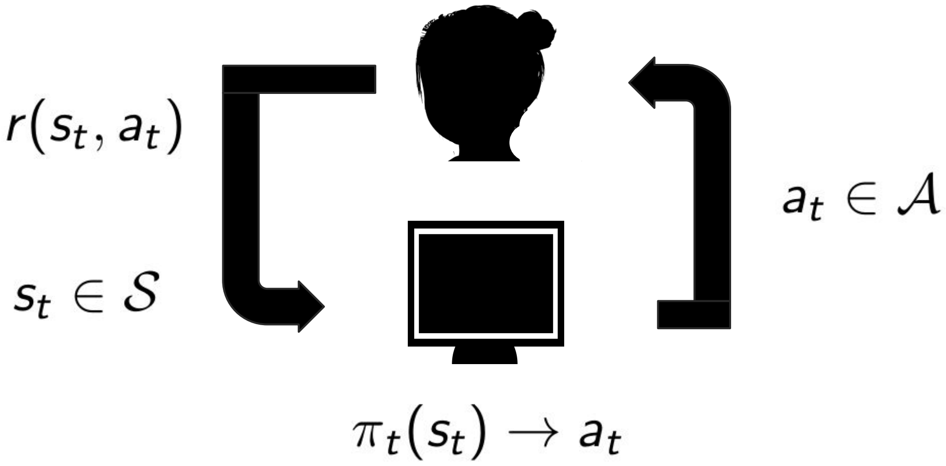


Encouraging Work on Observational Health Data (MIMIC) Hypotension

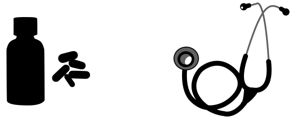


Futoma, Hughes, Doshi-Velez AISTATS 2020

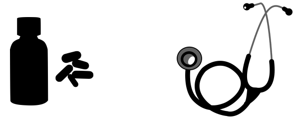
New Topic: Counterfactual / Batch RL



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

Patient group 1 ➡  ➡ Outcome: 92

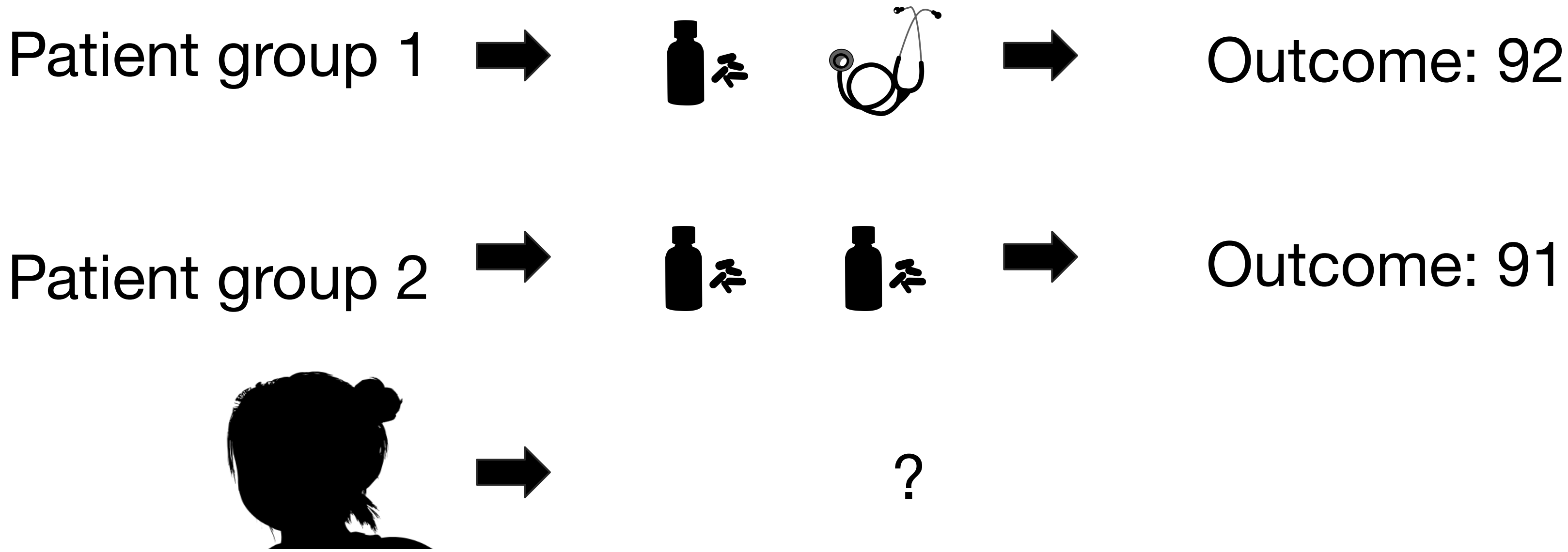
Patient group 2 ➡  ➡ Outcome: 91

Patient group 1  Outcome: 92

Patient group 2  Outcome: 91

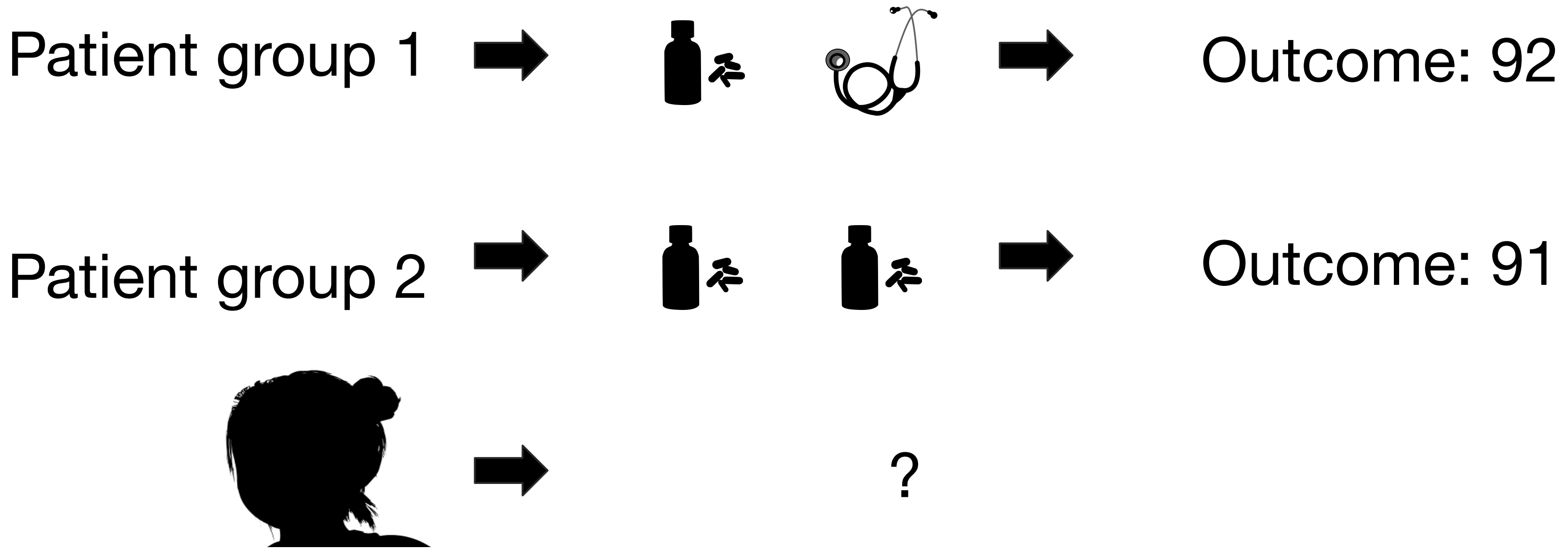
 ?

“What If?” Reasoning Given Past Data

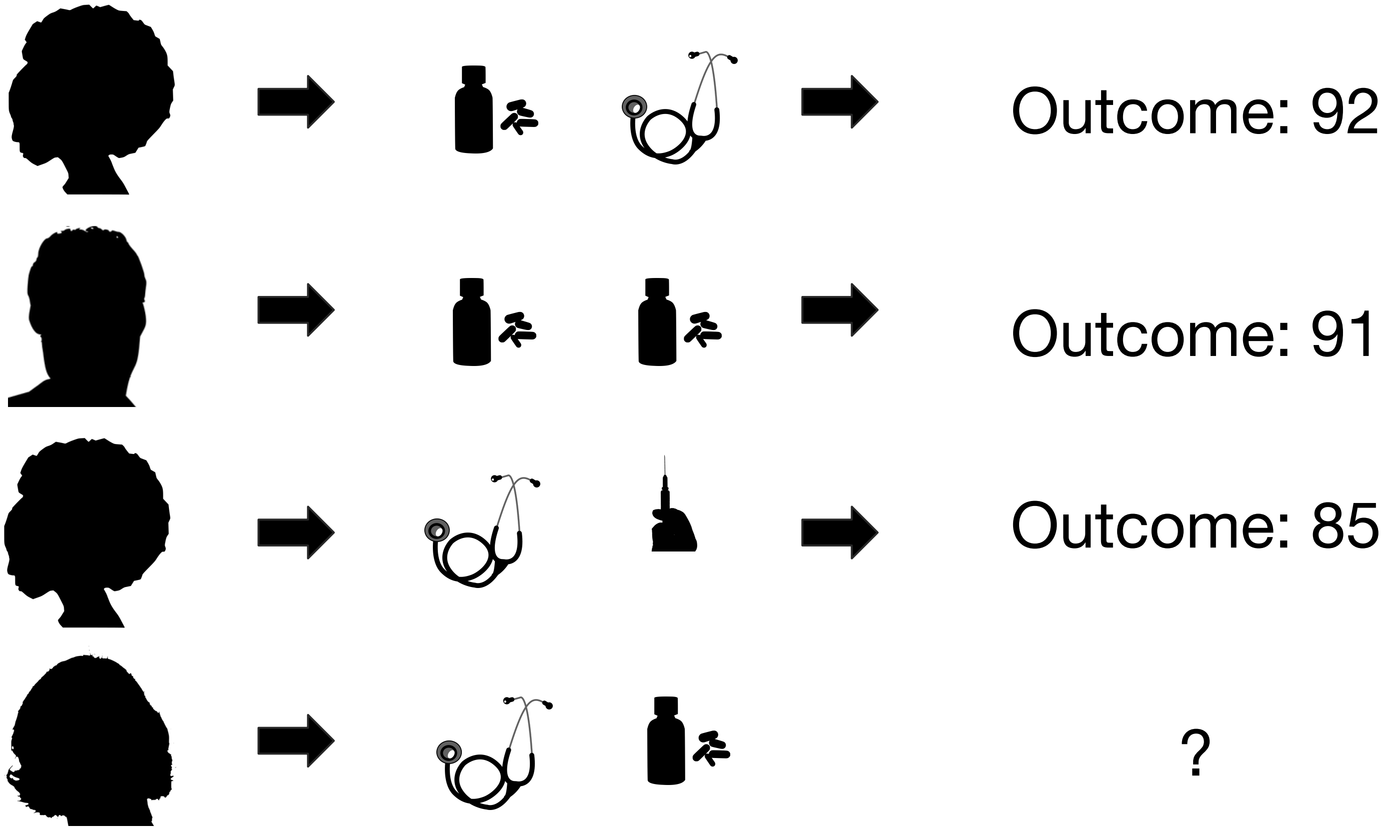


What information would you want to know in order to decide, given the above evidence, how best to treat new patient?

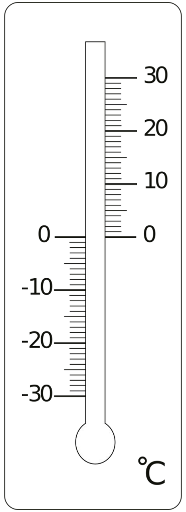
Data Is Censored in that Only Observe Outcomes for Decisions Made



Need for Generalization



Potential Applications



Off Policy Reinforcement Learning

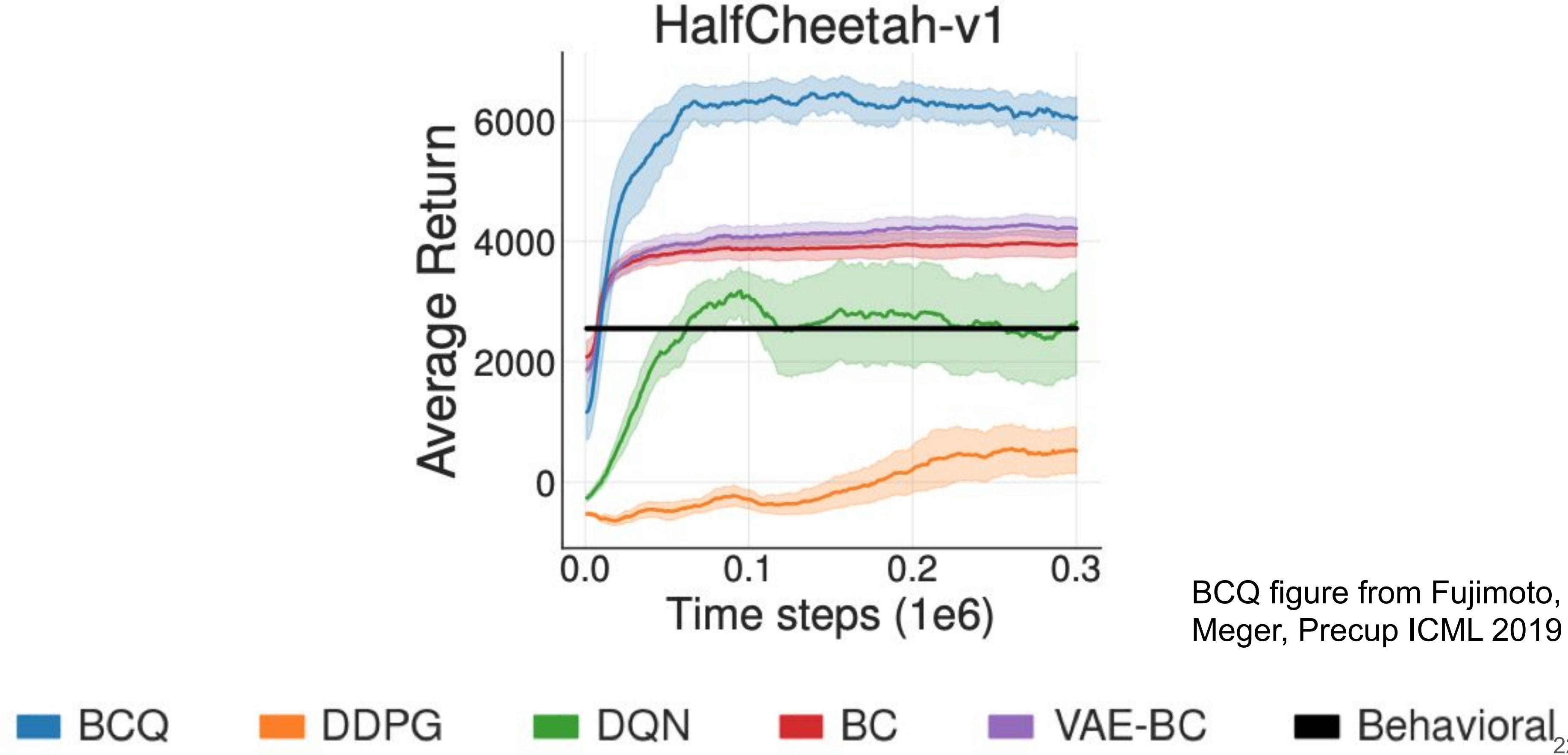
Watkins 1989
Watkins and Dayan 1992
Precup et al. 2000
Lagoudakis and Parr 2002
Murphy 2005
Sutton, Szepesvari and Maei 2009
Shortreed, Laber, Lizotte, Stroup, Pineau, & Murphy 2011
Degirs, White, and Sutton 2012
Mnih et al. 2015
Mahmood et al. 2014
Jiang & Li 2016
Hallak, Tamar and Mannor 2015
Munos, Stepleton, Harutyunyan and Bellemare 2016
Sutton, Mahmood and White 2016
Du, Chen, Li, Ziao, and Zhou 2016 **...**

Why Can’t We Just Use Q-Learning?

- Q-learning is an off policy RL algorithm
  - Can be used with data different than the state--action pairs would visit under the optimal Q state action values

- But deadly triad of bootstrapping, function approximation and off policy, and can fail

Important in Practice



BCQ figure from Fujimoto, Meger, Precup ICML 2019

Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
   a. Using models
   b. Using model free methods
   c. Use importance sampling
3. Offline policy learning / optimization

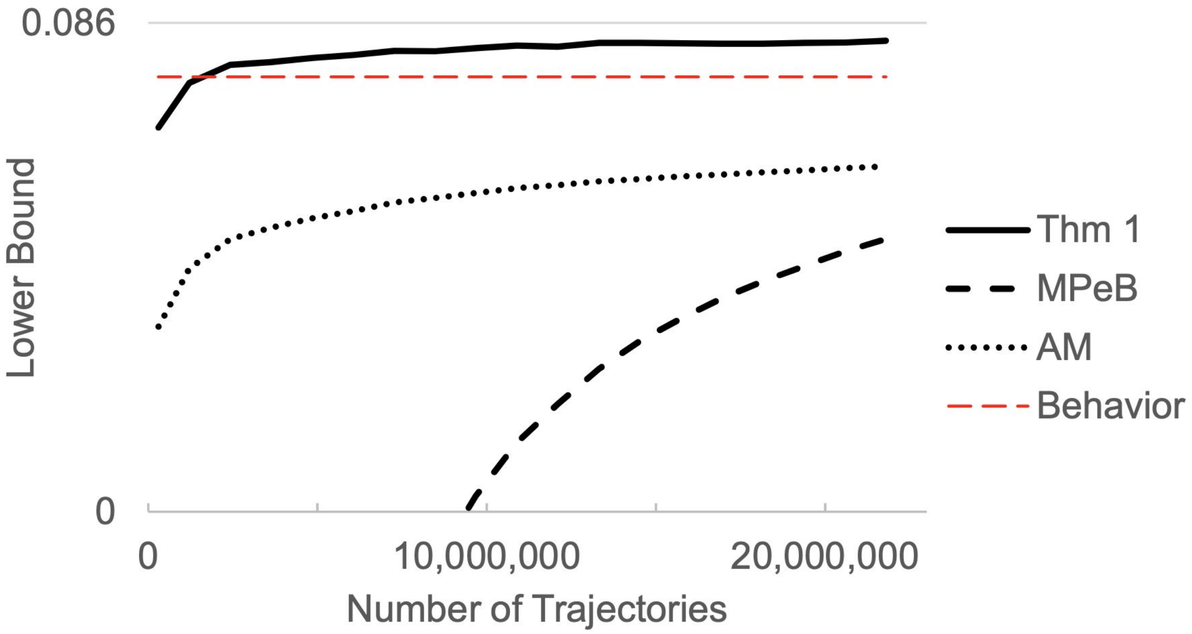
Batch Policy Evaluation: Estimate Performance of a Specific Decision Policy

$$\underbrace{\int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds}_{\text{Policy Evaluation}}$$

behavior policy

$\mathcal{D}$: 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 $\mathcal{D}$

Sample Efficient Methods Matter Policy Evaluation



Thomas, Philip, Georgios Theocharous, and Mohammad Ghavamzadeh. "High-confidence off-policy evaluation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1. 2015.

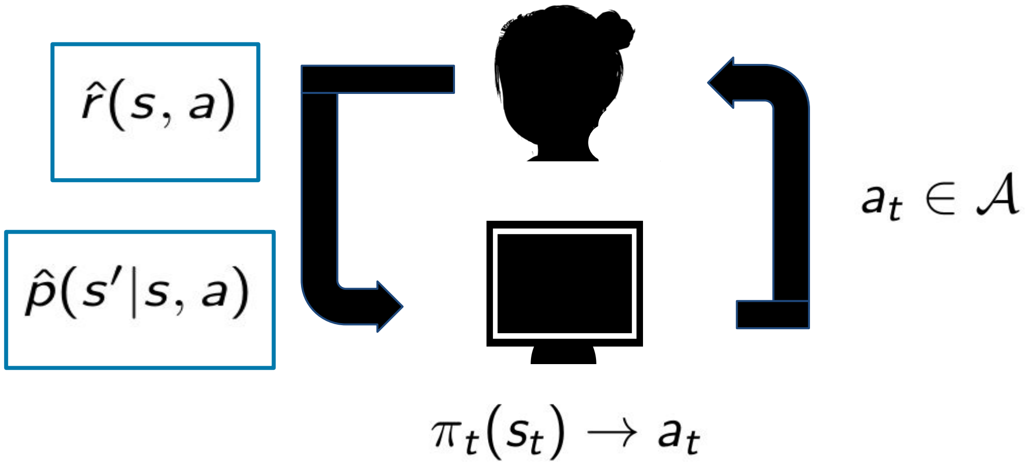
Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
    a. **Using models**
    b. Using model free methods
    c. Use importance sampling
3. Offline policy learning / optimization

Took > 30s

Took <= 30s

Learn Dynamics and Reward Models from Data



$\mathcal{D}$: 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 $\mathcal{D}$

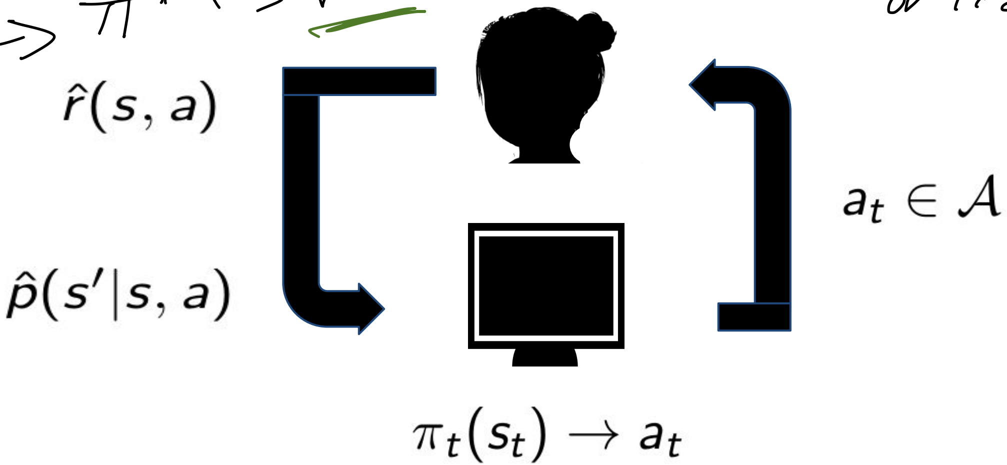
Learn Dynamics and Reward Models from Data, Evaluate Policy

$D \rightarrow \hat{p}(s'|s,a) \rightarrow \hat{\pi}^* \rightarrow V^{\hat{\pi}^*}$

$+ \hat{R}$

or learn a new policy w/DQN or any other RL method

$\hat{V}^{\hat{\pi}^*}$



$$V^\pi \approx (I - \gamma \hat{P}^\pi)^{-1} \hat{R}^\pi$$

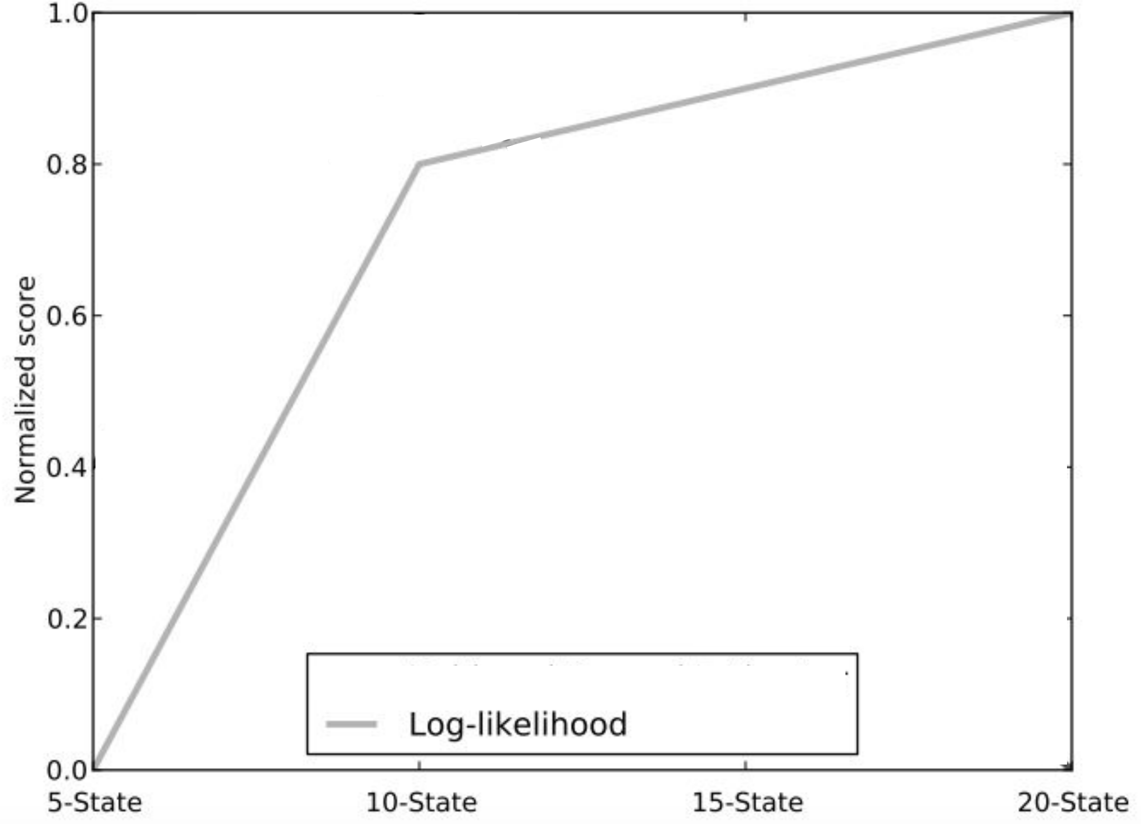
$$P^\pi(s'|s) = p(s'|s, \pi(s))$$

$\mathcal{D}$: 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 $\mathcal{D}$

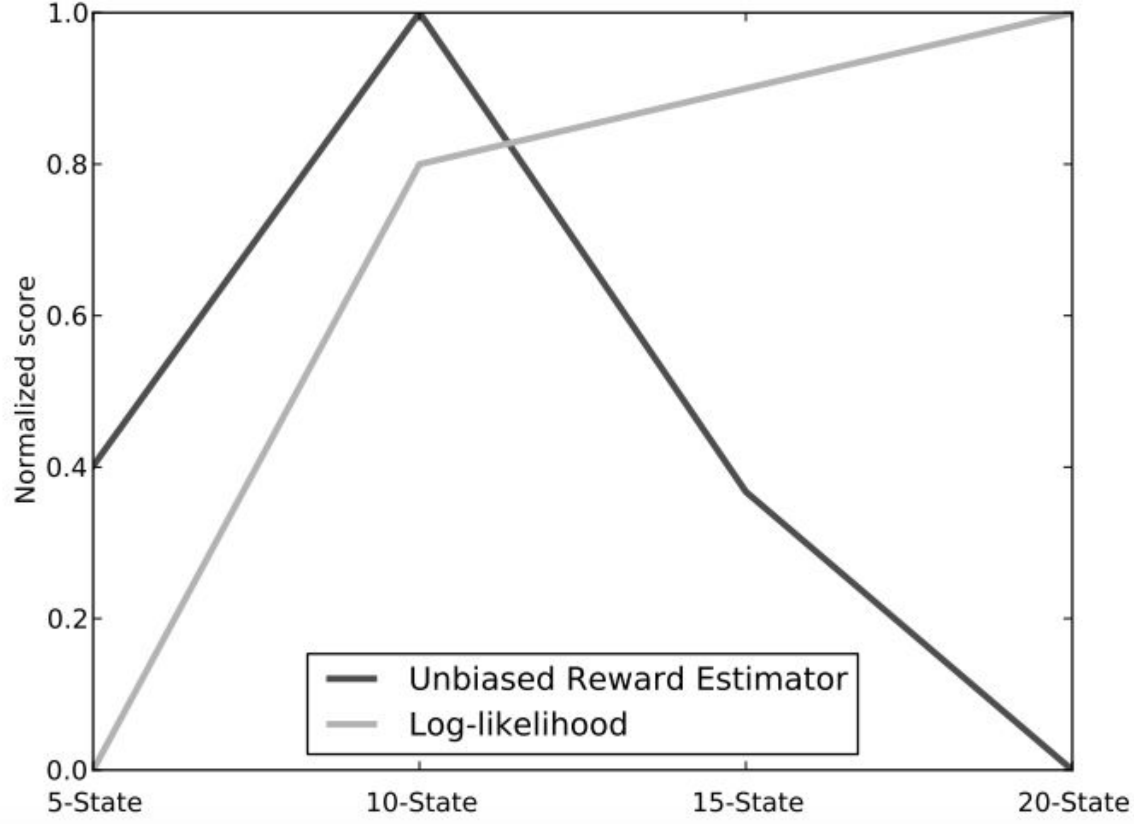
- Mannor, Simster, Sun, Tsitsiklis 2007

Better Dynamics/Reward Models for Existing Data (Improve likelihood)

held out

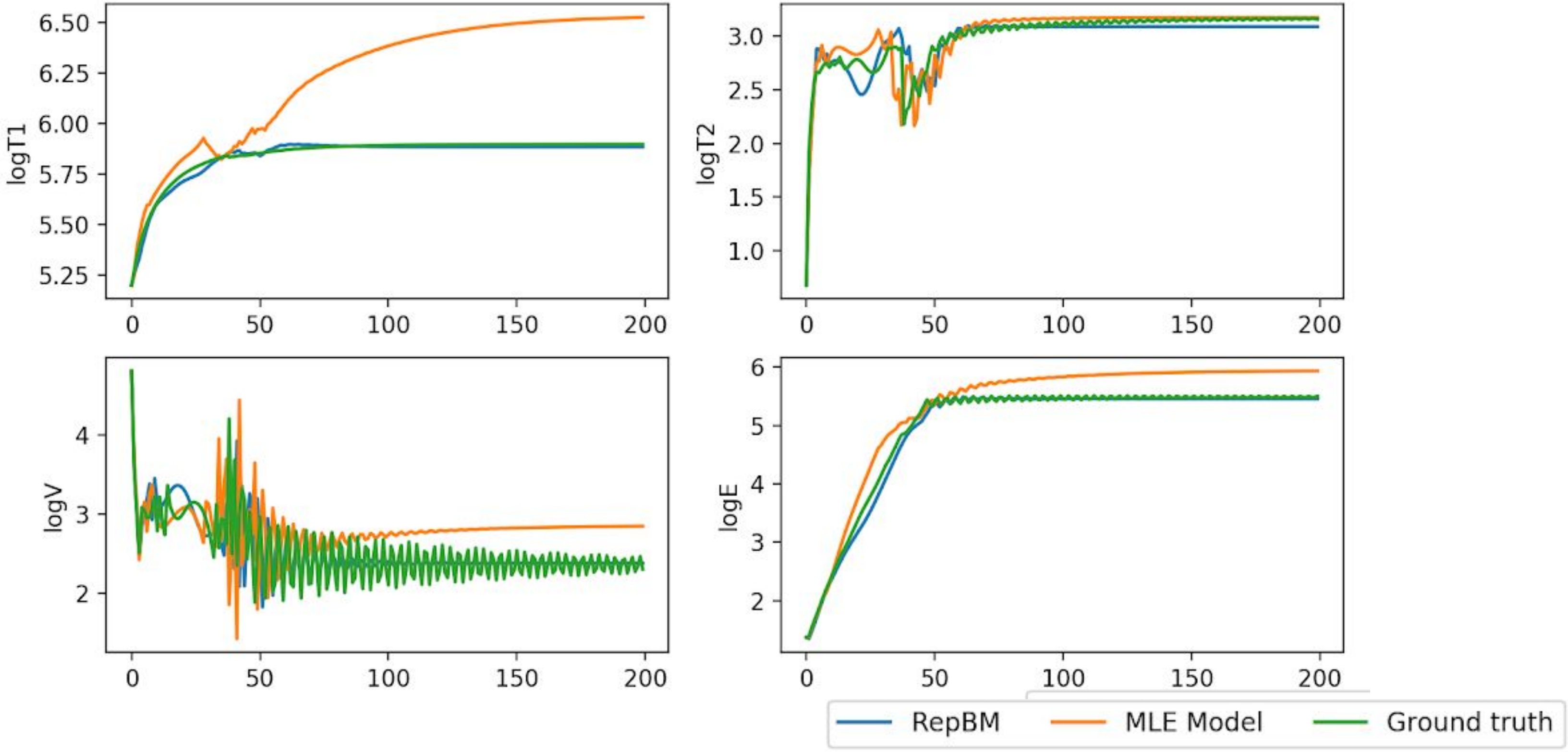


Better Dynamics/Reward Models for Existing Data, May **Not** Lead to Better Policies for Future Use → Bias due to Model **Misspecification**



Mandel, Liu, Brunskill, Popovic AAMAS 2014

Models Fit for Off Policy Evaluation Can Result in Better Estimates When Trained Under a **Different Loss Function**



Liu, Gottesman, Raghu, Komorowski, Faisal, Doshi-Velez, Brunskill NeurIPS 2018

Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
    a. Using models
    b. Using model free methods
    c. Use importance sampling
3. Offline policy learning / optimization

Model Free Value Function Approximation: $\underbrace{\text{Fitted Q Evaluation}}$

FQI (iteration)

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

$$\tilde{Q}^\pi(s_i, a_i) \quad = \quad r_i + \gamma V_\theta^\pi(s_{i+1})$$ target

$$\arg\min_\theta \textstyle\sum_i (Q_\theta^\pi(s_i, a_i) - \tilde{Q}^\pi(s_i, a_i))^2$$

Ernst 2005?

- Fitted Q evaluation, LSTD, ...

$\mathcal{D}$: 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 $\mathcal{D}$

Algorithm 3 Fitted Q Evaluation: $\texttt{FQE}(\pi, c)$

Input: Dataset $\mathrm{D} = \{x_i, a_i, x'_i, c_i\}_{i=1}^n \sim \pi_{\mathrm{D}}$. Function class F.
Policy π to be evaluated

1: Initialize $Q_0 \in \mathrm{F}$ randomly
2: **for** $k = 1, 2, \dots, K$ **do**
3: Compute target $y_i = c_i + \gamma Q_{k-1}(x'_i, \pi(x'_i)) \ \forall i$
4: Build training set $\widetilde{\mathrm{D}}_k = \{(x_i, a_i), y_i\}_{i=1}^n$
5: Solve a supervised learning problem:
$$Q_k = \arg\min_{f \in \mathrm{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i, a_i) - y_i)^2$$
6: **end for**

Output: $\widehat{C}^\pi(x) = Q_K(x, \pi(x)) \quad \forall x$

only for a fixed π

Let's assume we use a DNN for F.

What is different vs DQN?

no max

Le, H., Voloshin, C., & Yue, Y. (2019, May). Batch policy learning under constraints. In *International Conference on Machine Learning*

Example Fitted Q Evaluation Guarantees

concentrability coefficient

shift action distrib "overlap"

$$d_F^\pi \quad = \quad \sup_{g\in F} \inf_{f\in F} ||f - B^\pi g||_\pi$$

Theorem 4.2 (Generalization error of FQE). *Under Assumption 1, for $\epsilon > 0$ & $\delta \in (0,1)$, after K iterations of Fitted Q Evaluation (Algorithm 3), for $n = O\Big(\frac{\overline{C}^4}{\epsilon^2}\big(\log\frac{K}{\delta} + \mathtt{dim}_{\mathrm{F}} \log\frac{\overline{C}^2}{\epsilon^2} + \log \mathtt{dim}_{\mathrm{F}}\big)\Big)$, we have with probability $1-\delta$:*

how much data

ϵ desired target accuracy

$$\left|\int_{s_0\in\rho} \hat{V}^\pi(s_0) - V^\pi(s_0)\right| \quad \leq \quad \frac{\gamma^{.5}}{(1-\gamma)^{1.5}}\left(\sqrt{\beta_u}(2d_F^\pi + \epsilon) + \frac{2\gamma^{K/2}\bar{C}}{(1-\gamma)^{.5}}\right)$$

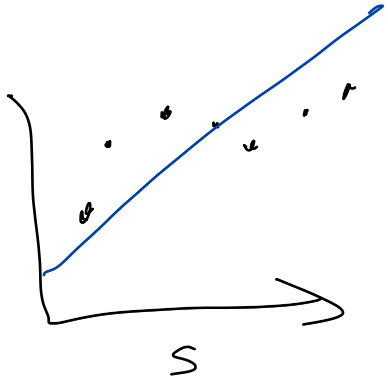
Le, Voloshin, Yue ICML 2019

$\mathcal{D}$: 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 $\mathcal{D}$

Model Free Policy Evaluation

- Challenge: still relies on Markov assumption
- Challenge: still relies on models being well specified or have no computable guarantees if there is misspecification

$$d_F^\pi = \sup_{g \in F} \inf_{f \in F} ||f - B^\pi g||_\pi$$



Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
    a. Using models
    b. Using model free methods
    c. Use importance sampling
3. Offline policy learning / optimization

Off Policy Evaluation With Minimal Assumptions

- Would like a method that doesn’t rely on models being correct or Markov assumption
- Monte Carlo methods did this for online policy evaluation
- We would like to do something similar
- Challenge: data distribution mismatch

Importance Sampling*

x = states $r(x)$ = reward of state

$p(x)$ = prob of reaching x under a policy

but no data from $p(x)$

$$\mathbb{E}_p[r] = \sum_x p(x)r(x)$$

$q(x)$ is a different policy

$$= \sum_x \frac{q(x)}{q(x)} p(x) r(x)$$

$$= \sum_x q(x) \left[\frac{p(x)}{q(x)} r(x) \right]$$

$$\approx \frac{1}{N} \sum_{\substack{i=1 \\ x \sim q(x)}}^{N} \frac{p(x_i)}{q(x_i)} r(x_i)$$

unbiased estimate

*Former CS234 student said this was his favorite idea of the class!

Importance Sampling: Can Compute Expected Value Under An Alternate Distribution!

$$\begin{aligned}
\mathbb{E}_p[r] &= \sum_x p(x) r(x) \\
&= \sum_x \frac{p(x)q(x)}{q(x)} r(x) \\
&\approx \frac{1}{N} \sum_{i=1, x\sim q}^{N} \frac{p(x_i)}{q(x_i)} r(x_i)
\end{aligned}$$

Importance Sampling is an Unbiased Estimator of True Expectation Under Desired Distribution If

$$\begin{aligned}\mathbb{E}_p[r] &= \sum_x p(x) r(x) \\ &= \sum_x \frac{p(x)q(x)}{q(x)} r(x) \\ &\approx \frac{1}{N} \sum_{i=1, x\sim q}^{N} \frac{p(x_i)}{q(x_i)} r(x_i)\end{aligned}$$

* consider healthcare
where may want to
consider patients
getting different
treatments

x, a, r

- The sampling distribution q(x) > 0 for all x s.t. p(x) > 0 (Coverage / overlap)
- No hidden confounding *

Check Your Understanding: Importance Sampling

We can use importance sampling to do batch bandit policy evaluation. Consider we have a dataset for ~~pulls~~ samples from 3 actions. Consider that

- Action 1 is a Bernoulli var where with probability 0.02 r= 100 else r = 0
- Action 2 is a Bernoulli var where with probability 0.55 r=2 else r = 0
- Action 3 is a Bernoulli var where with probability 0.5 r=1, else r = 0

Select all that are true.

behavior policy

- Data is sampled from π1 where with probability 0.8 it pulls action 3 else it pulls action 2. The policy we wish to evaluate, π2, pulls action 2 with probability 0.5 else it pulls action 1. π2 has higher true reward than pi1.
- We cannot use π1 to get an unbiased estimate of the average reward π2 using importance sampling.
- If rewards can be positive or negative, we can still get a lower bound on π2 using data from pi1 using importance sampling
- Not Sure

Check Your Understanding: Importance Sampling

We can use importance sampling to do batch bandit policy evaluation. Consider we have a dataset for pulls from 3 actions. Consider that

- Action 1 is a Bernoulli var where with probability 0.02 r= 100 else r = 0 $E[r(a_1)] = 2 + 0 = 2$
- Action 2 is a Bernoulli var where with probability 0.55 r=2 else r = 0 $E[r(a_2)] = .55 \cdot 2 = 1.1$
- Action 3 is a Bernoulli var where with probability 0.5 r=1, else r = 0 $E[r(a_3)] = .5$

π_1 $.8\, r(a_3) + .2\, r(a_2)$

Select all that are true.

reward $.8 \cdot .5 + .2 \cdot 1.1 \approx .42$ ish

- T Data is sampled from π1 where with probability 0.8 it pulls action 3 else it pulls action 2. The policy we wish to evaluate, π2, pulls action 2 with probability 0.5 else it pulls action 1. π2 has higher true reward than pi1. π_2 $.5\, r(a_2) + .5 \cdot r(a_1) = .5 \cdot 1.1 + .5 \cdot 2 \approx 1.5$ true
- T We cannot use π1 to get an unbiased estimate of the average reward π2 using importance sampling.
- F If rewards can be positive or negative, we can still get a lower bound on π2 using data from pi1 using importance sampling
  → if rewards ≥ 0 you can do this
- Not Sure

inverse propensity weighting

Importance Sampling for RL Policy Evaluation

$$V^\pi(s) = \sum_\tau \overbrace{p(\tau|\pi, s)}^{\text{prob of traj}} R(\tau)$$

trajectories

reward of ~~a~~ traj τ

$$= \sum_\tau p(\tau|\pi, s) \frac{p(\tau|\pi_b, s)}{p(\tau|\pi_b, s)} R(\tau)$$

$$= \sum_\tau p(\tau|\pi_b, s) \left[\frac{p(\tau|\pi, s)}{p(\tau|\pi_b, s)} R(\tau) \right]$$

$\mathcal{D}$: 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 $\mathcal{D}$

Importance Sampling for RL Policy Evaluation

$$
\begin{aligned}
V^\pi(s) &= \sum_\tau p(\tau|\pi, s) R(\tau) \\
&= \sum_\tau p(\tau|\pi_b, s) \frac{p(\tau|\pi, s)}{p(\tau|\pi_b, s)} R_\tau \\
&\approx \frac{1}{N} \sum_{i=1, \tau_i \sim \pi_b}^{N} \frac{p(\tau_i|\pi, s)}{p(\tau_i|\pi_b, s)} R_{\tau_i}
\end{aligned}
$$

$$p(\tau_i|\pi, s) = \prod_{t=1}^{H} p(s_{t+1}|s_t, a_t)\,\pi(a_t|s_t)$$

$=$

$\mathcal{D}$: 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 $\mathcal{D}$

Importance Sampling for RL Policy Evaluation: Don't Need to Know Dynamics Model!

$$
\begin{aligned}
V^{\pi}(s) &= \sum_{\tau} p(\tau|\pi, s) R(\tau) \\
&= \sum_{\tau} p(\tau|\pi_b, s) \frac{p(\tau|\pi, s)}{p(\tau|\pi_b, s)} R_\tau \\
&\approx \sum_{i=1, \tau_i \sim \pi_b}^{N} \frac{p(\tau_i|\pi, s)}{p(\tau_i|\pi_b, s)} R_{\tau_i} \\
&= \sum_{i=1, \tau_i \sim \pi_b}^{N} R_{\tau_i} \prod_{t=1}^{H_i} \frac{\cancel{p(s_{i,t+1}|s_{it}, a_{it})} p(a_{it}|\pi, s_{it})}{\cancel{p(s_{i,t+1}|s_{it}, a_{it})} p(a_{it}|\pi_b, s_{it})} \\
&= \sum_{i=1, \tau_i \sim \pi_b}^{N} R_{\tau_i} \prod_{t=1}^{H_i} \frac{p(a_{it}|\pi, s_{it})}{p(a_{it}|\pi_b, s_{it})}
\end{aligned}
$$

$\mathcal{D}$: 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 $\mathcal{D}$

- First used for RL by Precup, Sutton & Singh 2000. Recent work includes: Thomas, Theocharous, Ghavamzadeh 2015; Thomas and Brunskill 2016; Guo, Thomas, Brunskill 2017; Hanna, Niekum, Stone 2019

Importance Sampling

- Does not rely on Markov assumption
- Requires minimal assumptions
- Provides unbiased estimator
- Similar to Monte Carlo estimator but corrects for distribution mismatch

Optional Check Your Understanding: Importance Sampling 2

Select all that you'd guess might be true about importance sampling

- It 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
- It is likely to be high variance
- Not Sure

Per Decision Importance Sampling (PDIS)

- Leverage temporal structure of the domain (**similar to policy gradient**)

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

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

Importance Sampling Variance

- Importance sampling, like Monte Carlo estimation, is generally high variance
- Importance sampling is particularly high variance for estimating the return of a policy in a sequential decision process

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

- Variance can generally scale exponentially with the horizon
  a. Concentration inequalities like Hoeffding scale with the largest range of the variable
  b. The largest range of the variable depends on the product of importance weights
  c. **Optional Check your understanding: for a H step horizon with a maximum reward in a single trajectory of 1, and if p(a|s, pi_b) = .1 and p(a|s, pi) = 1 for each time step, what is the maximum importance-weighted return for a single trajectory?**

$$R_{\tau_i} \prod_{t=1}^{H_i} \frac{p(a_{it}|\pi, s_{it})}{p(a_{it}|\pi_b, s_{it})}$$

Extensions

- Leveraging Markov structure to break curse of horizon.
  - Marginalized importance sampling (state-action distribution)
  - Dai, Nachum, Chow, Li (dualdice, coindice) 2019/2020
  - Liu, Li, Tang, Zhou Neurips 2018
- Doubly robust estimation (Jiang and Li 2016; Thomas and Brunskill 2016)
- Blended estimators (Thomas and Brunskill 2016)

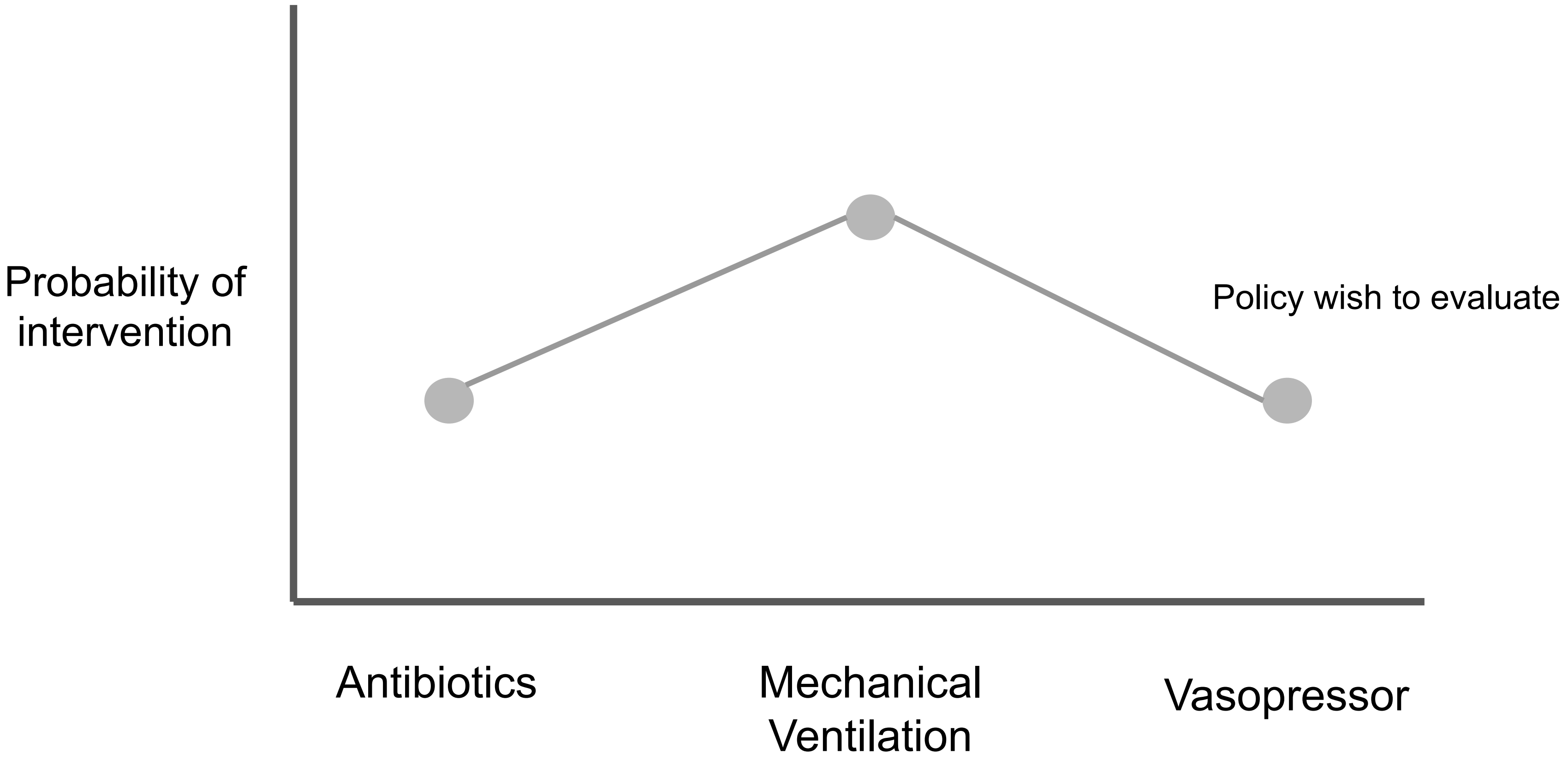
Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
   a. Using models
   b. Using model free methods
   c. Use importance sampling
3. **Offline policy learning / optimization**

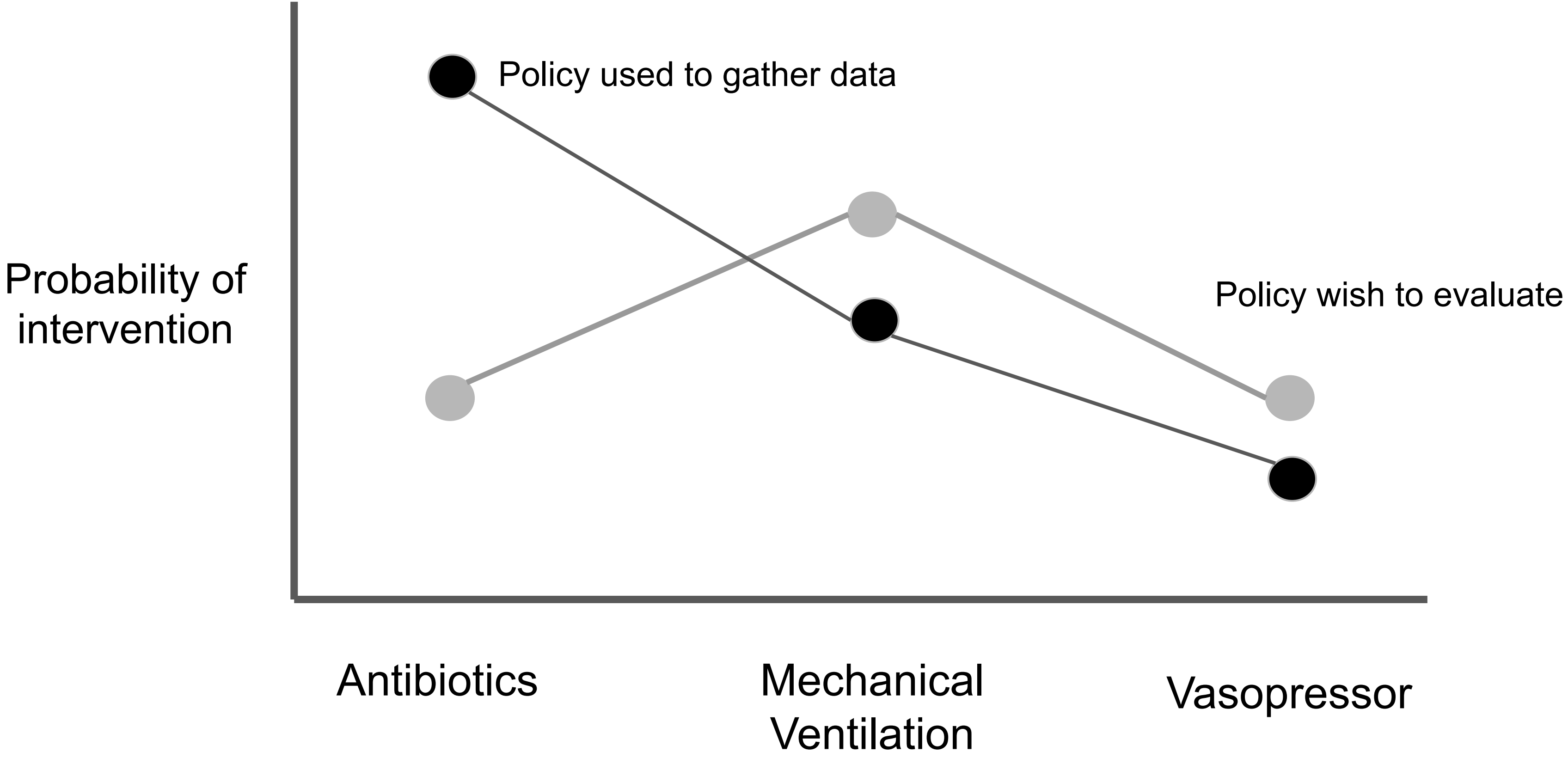
$$\arg\max_{\pi} \int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds$$

$\mathcal{D}$: 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 $\mathcal{D}$

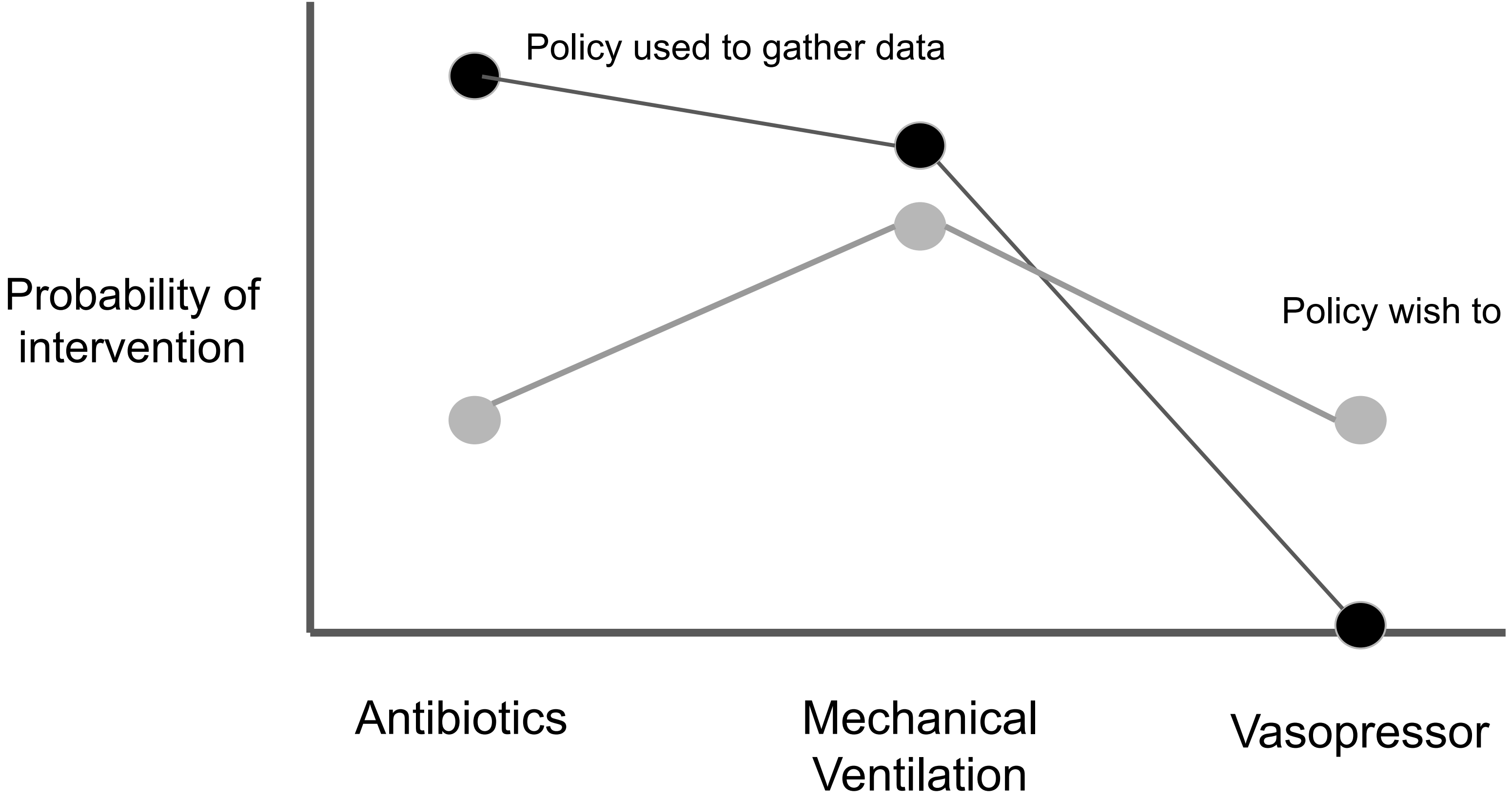
Challenges in Offline Policy Optimization



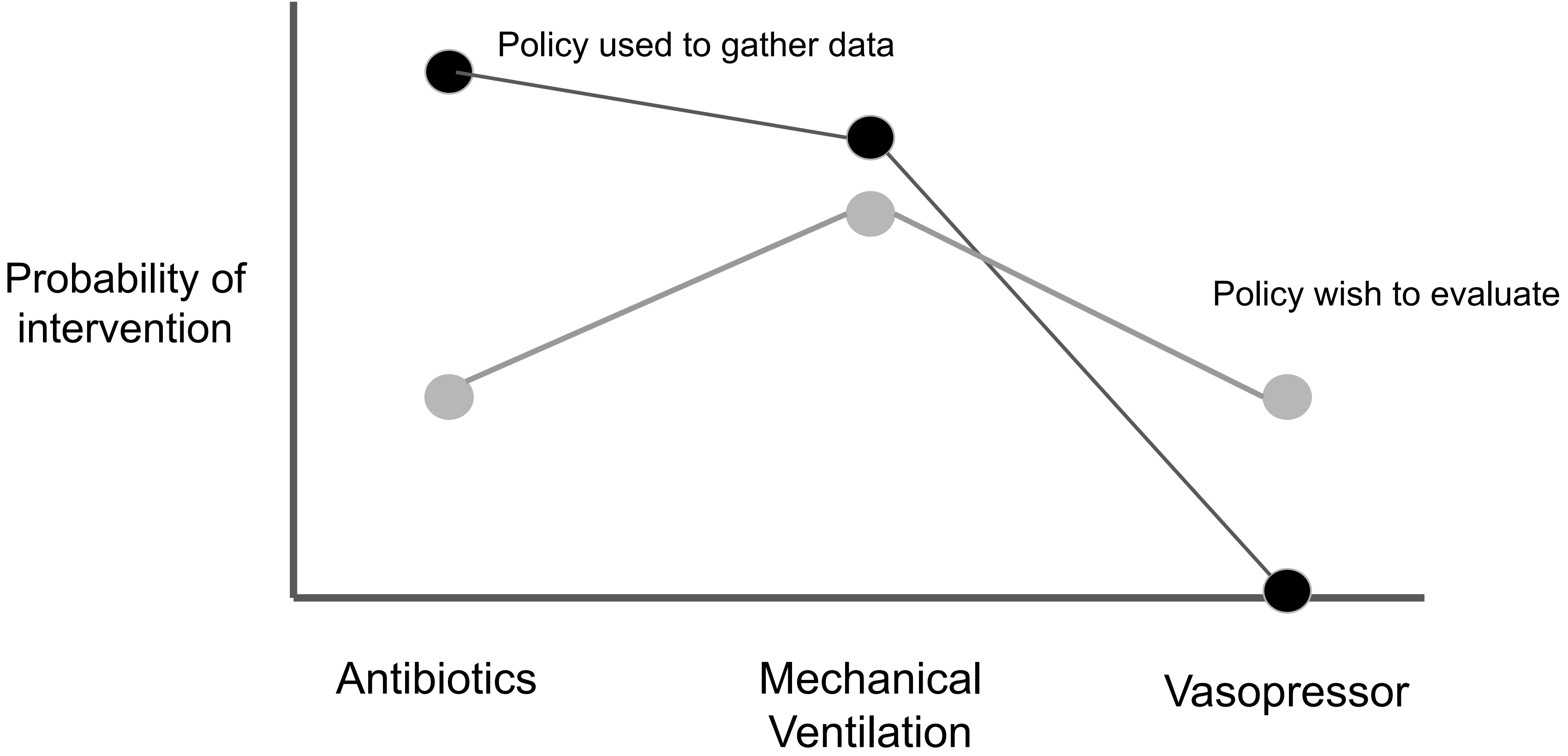
Overlap Requirement: Data Must Support Policy Wish to Evaluate



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



Seen Data Distribution Shift Challenge Before. PPO. DPO. RLHF…



Offline Policy Optimization Up to ~ 2020

- Algorithms often assume overlap
    - Off policy estimation: for policy of interest
    - Off policy optimization: for all policies including optimal one (“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

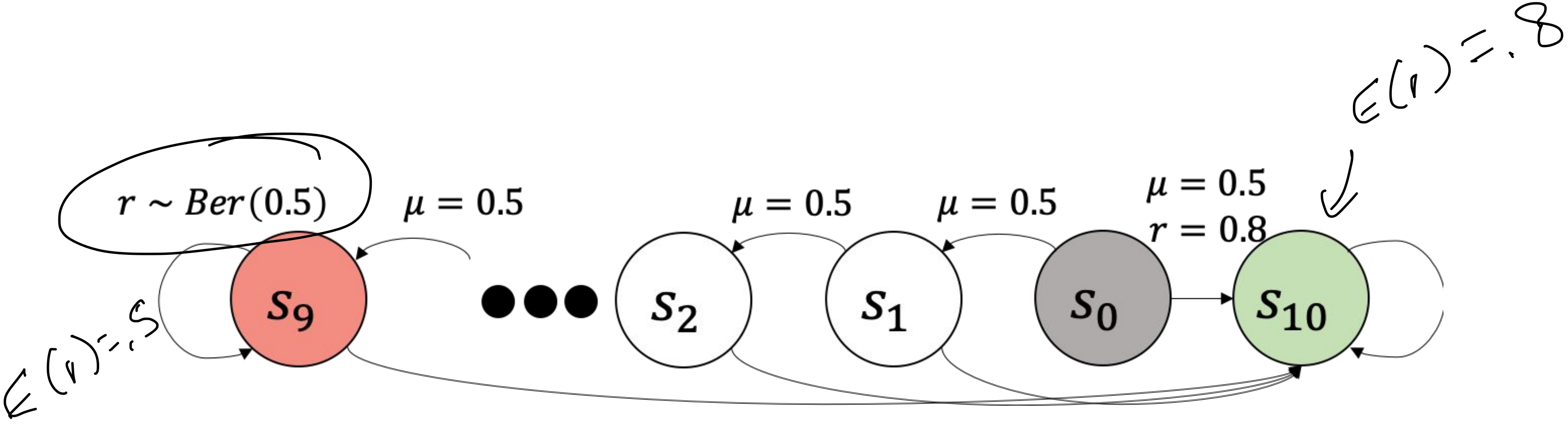
Doing the Best with What We've Got: Off Policy Optimization Without Full Data Coverage

- Restrict off policy optimization to those with overlap in data
  - We've seen related ideas before: KL constraint or PPO clipping
- Computationally tractable algorithm
- Simple idea: assume **pessimistic outcomes** for areas of state--action space with insufficient overlap/support

Common challenge that's attracted growing interest before 2020 but…

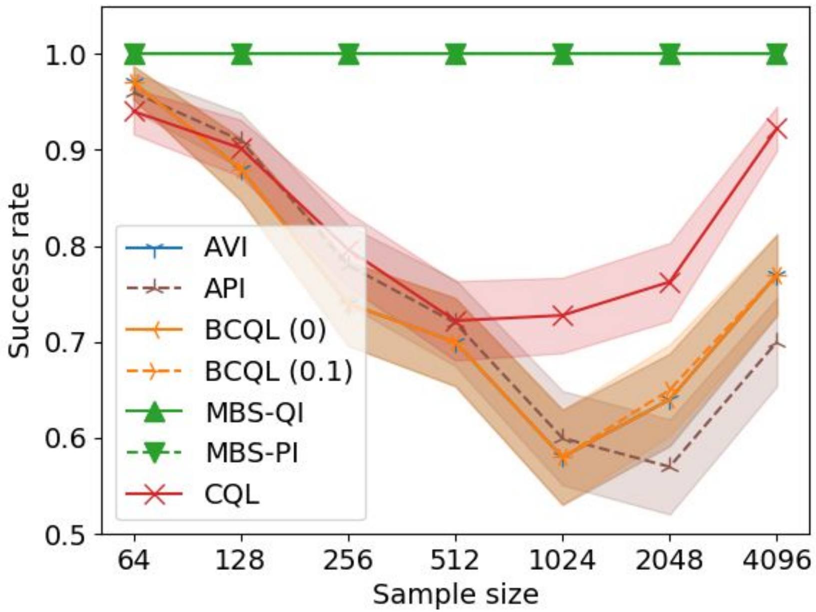
Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Illustrative Examples

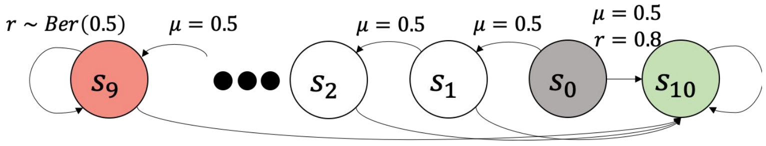


Liu, Swaminathan, Agarwal, Brunskill NeurIPS 2020

Recent Conservative Batch Reinforcement Learning Are Insufficient



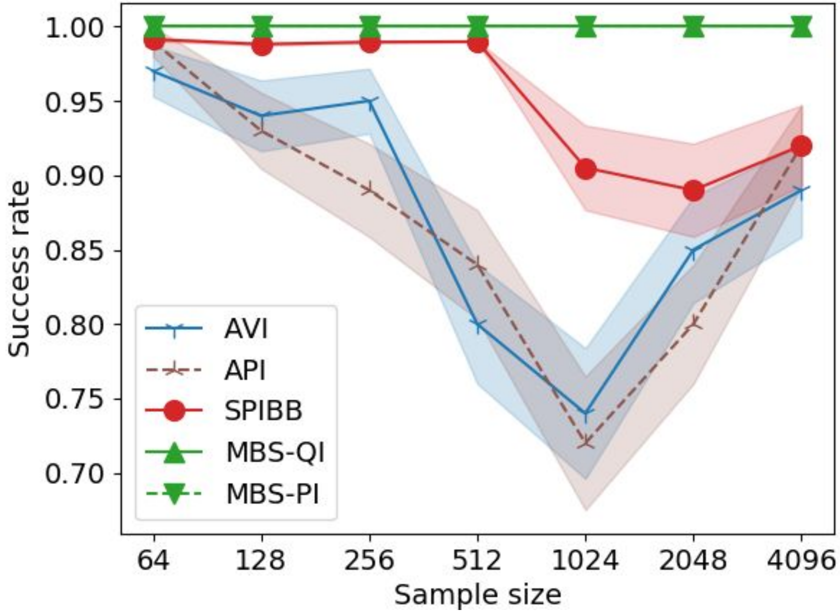
Success rate: #(getting the optimal policy)/#(trials)



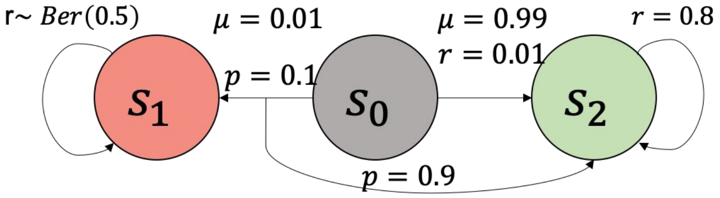
Reasons why baselines fail:

- Many baselines focus on penalty/constraints that are based on dist(π(a|s), π_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 *ν*(s,a) distribution possible under some policy in this MDP

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

$$\Downarrow$$

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

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

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

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

$$\Downarrow$$

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

*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 $\Pr(\zeta(s,a) = 0 | \pi) \leq \epsilon_{\zeta}$}

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.

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

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:

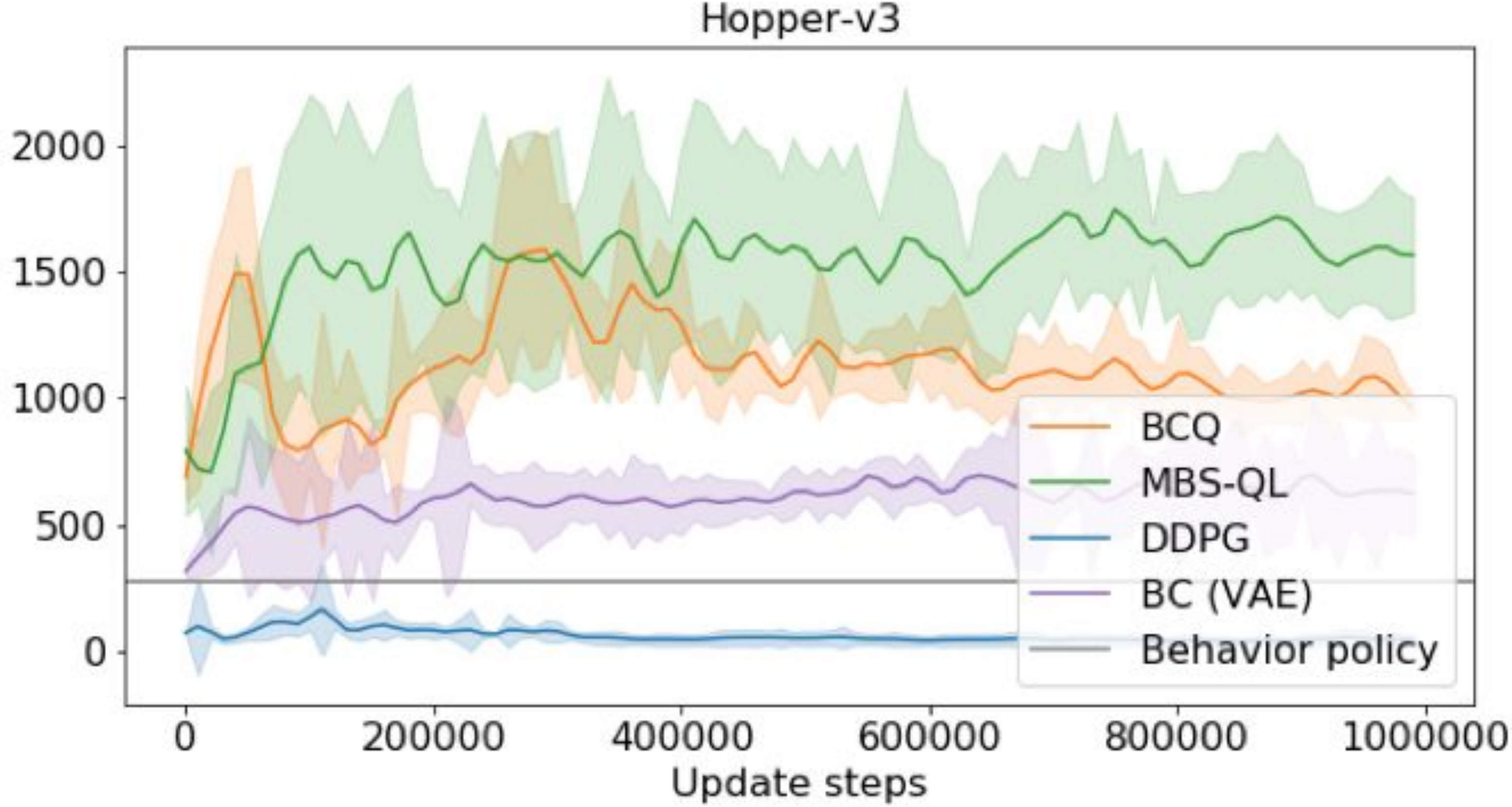
$$O\left(\frac{V_{\max}}{(1-\gamma)^3 b}\sqrt{\frac{\ln(|\mathcal{F}||\Pi|/\delta)}{n}}\right) + \frac{V_{\max}\epsilon_{\zeta}}{1-\gamma}$$

- 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.

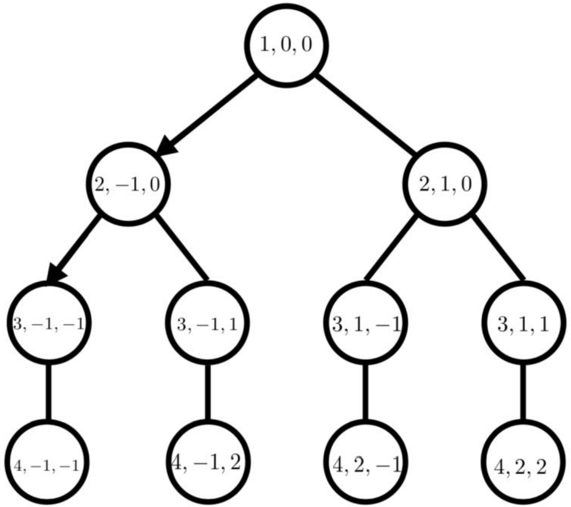
Can Do Get Substantially Better Solutions, With Same Data



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 Conservative **Model-Based** Offline RL

- Yu, Thomas, Yu, Ermon, Zou, Levine, Finn & Ma (NeurIPS 2020)
- 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 $s \to a$
S_0: Set of initial states
$\hat{V}^\pi(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Concurrent Work Conservative Offline RL

- Yu, Thomas, Yu, Ermon, Zou, Levine, Finn & Ma (NeurIPS 2020)
- 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
- **Conservative Q Learning (CQL) (Kumar et al.) continues to be popular**

$\mathcal{D}$: 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
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

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

- Pessimistic approaches do quite well, different methods win in different areas
- MBS has stronger theory results

Pessimistic Offline Policy Learning

- Restrict off policy optimization to those with overlap in data
- Simple idea: assume pessimistic outcomes for areas of state--action space with insufficient overlap/support
  - In model
  - In Q function

PPO constrained updates

Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
   a. Using models
   b. Using model free methods
   c. Use importance sampling
3. **Offline policy learning / optimization**

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

$\mathcal{D}$: 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 $\mathcal{D}$

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

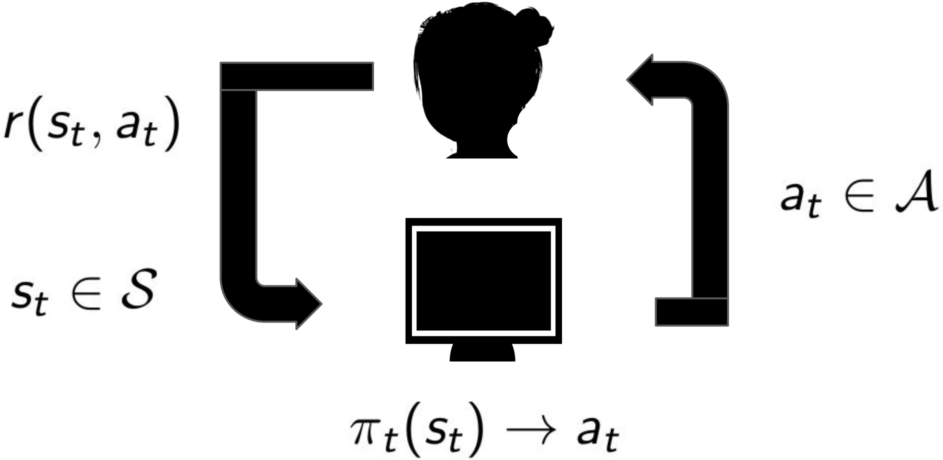
$$\arg\max_{a \in \mathcal{A}} f(a)$$

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

Thomas, Castro da Silva, Barto, Giguere, Brun, Brunskill. Science 2019

Offline RL

with Constraints on Future Performance of Policy



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

An Algorithm for Offline RL with Safety Constraints

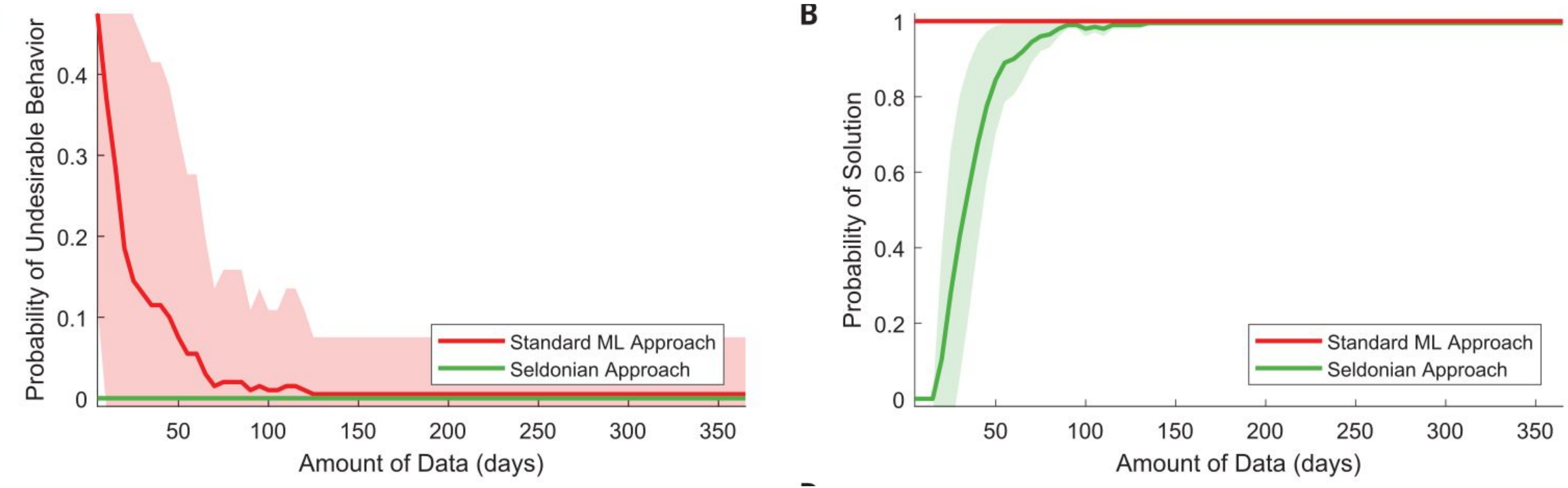
- Take in desired behavior constraints g and confidence level & data
- Given a finite set of decision policies, for each policy i
  - Compute generalization bound for each constraint
  - If passes all with desired confidence*, Safe(i) = true
- Estimate performance f of all policies that are safe
- Return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

Thomas, Castro da Silva, Barto, Giguere, Brun, Brunskill. Science 2019

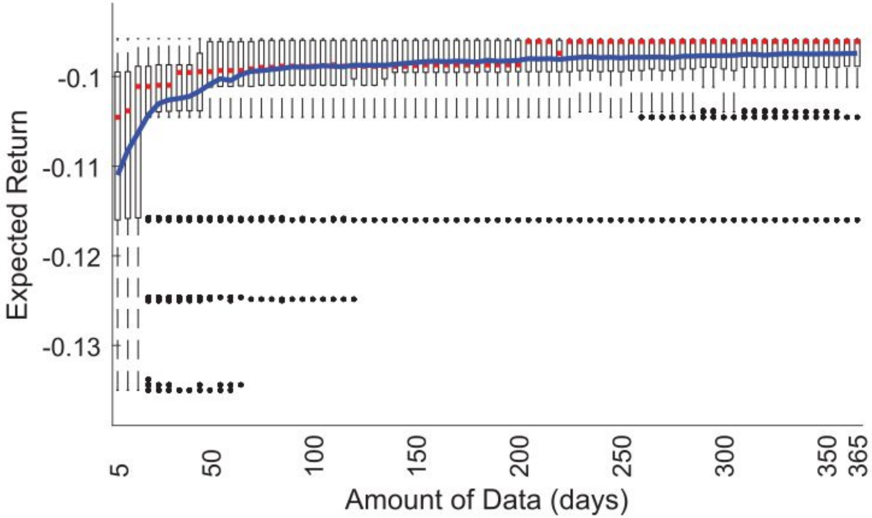
Diabetes Insulin Management

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

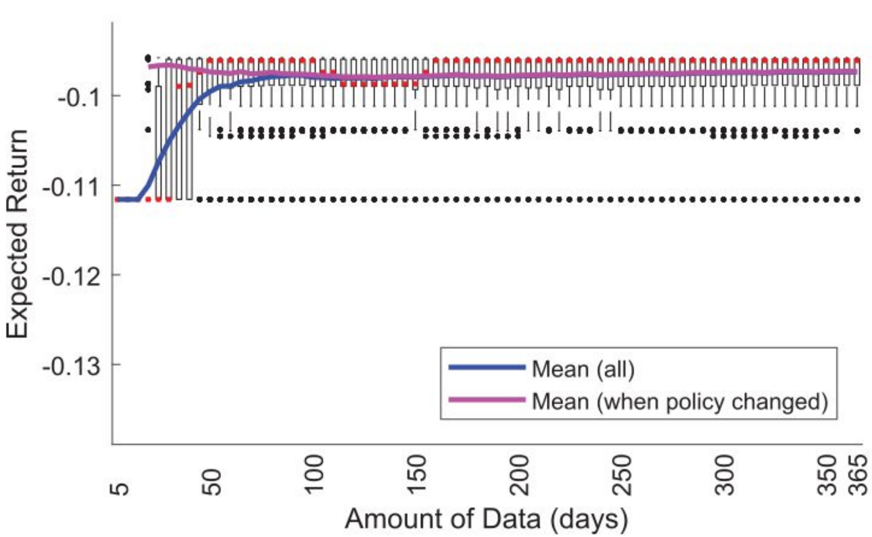
Personalized Insulin Dosage: Safe Batch Policy Improvement



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



Standard RL



Our Safe Batch RL

Outline for Today

1. Introduction and Setting
2. Offline batch policy evaluation
   a. Using models
   b. Using model free methods
   c. Use importance sampling
3. Offline policy learning / optimization

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

$\mathcal{D}$: 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 $\mathcal{D}$

What You Should Know/ Be Able to Do

- Define and apply importance sampling for off policy policy evaluation
- Describe limitations of model and model free off policy evaluation
- Define some limitations of IS (variance)
- Explain when and why offline RL may outperform imitation learning
- Describe the idea of pessimism under uncertainty and why it is useful
- Provide application examples where offline RL and offline policy evaluation would be useful

Optional Check Your Understanding: Importance Sampling 2 Solutions

Select all that you'd guess might be true about importance sampling

- It 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)
- It is likely to be high variance (True)
- Not Sure

Importance Sampling Variance. Optional CYU Solutions

- Importance sampling, like Monte Carlo estimation, is generally high variance
- Importance sampling is particularly high variance for estimating the return of a policy in a sequential decision process

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

- Variance can generally scale exponentially with the horizon
    a. Concentration inequalities like Hoeffding scale with the largest range of the variable
    b. The largest range of the variable depends on the product of importance weights
    c. **Optional Check your understanding: for a H step horizon with a maximum reward in a single trajectory of 1, and if p(a|s, pi_b) = .1 and p(a|s, pi) = 1 for each time step, what is the maximum importance-weighted return for a single trajectory?**

$$R_{\tau_i} \prod_{t=1}^{H_i} \frac{p(a_{it}|\pi, s_{it})}{p(a_{it}|\pi_b, s_{it})}$$

Solution: 1 / (.1)^H = 10^H