Batch / Offline RL Policy Evaluation

Emma Brunskill
CS234
Winter 2023

Thanks to Phil Thomas for some figures

Refresh Your Understanding

Select all that are true about First-Visit Monte-Carlo (MC) Policy Evaluation:

- It does not rely on the Markov assumption.
- It requires the dynamics model to be known.
- It is always unbiased.
- It exhibits low variance.
- It uses importance sampling
- None of the above.

Refresh Your Understanding: Solutions

Select all that are true about First-Visit Monte-Carlo (MC) Policy Evaluation:

- It does not rely on the Markov assumption.
- It requires the dynamics model to be known.
- It is always unbiased.
- It exhibits low variance.
- It uses importance sampling
- None of the above.

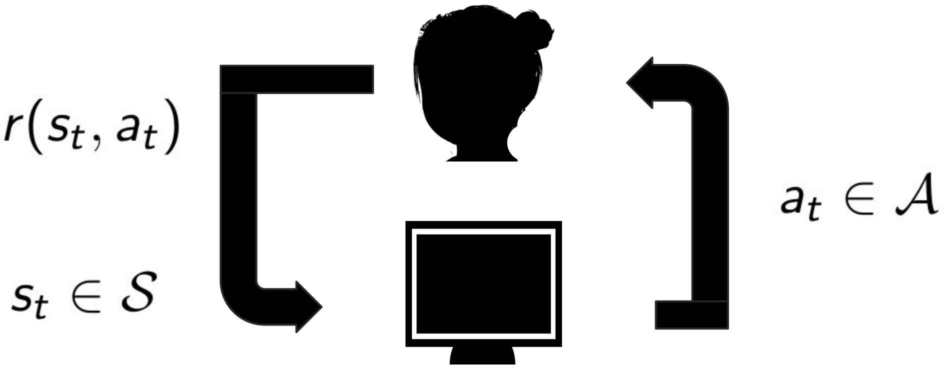
Outline for Today

1. Introduction and Setting
2. Offline batch evaluation using models
3. Offline batch evaluation using Q functions
4. Offline batch evaluation using importance sampling

Where We Are

- Fast reinforcement learning
- **Learning from offline data**
    - **Overview and Policy evaluation**
    - **Policy optimization**

Reinforcement Learning

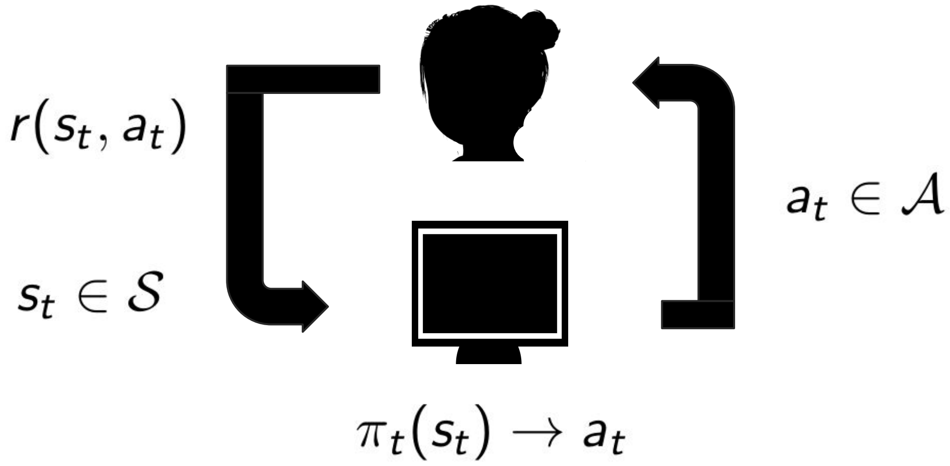


$$\pi_t(s_t) \rightarrow a_t$$

$$\underbrace{V^{\pi}(s)}_{\text{Value func.}} = \underbrace{r(s, \pi(s))}_{\text{Reward}} + \gamma \sum_{s'} \underbrace{p(s'|s, a)}_{\text{Dynamics}} V^{\pi}(s')$$

Only observed through samples (experience)

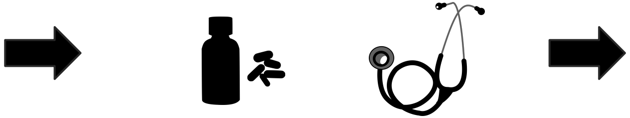
New Topic: Counterfactual / Batch RL



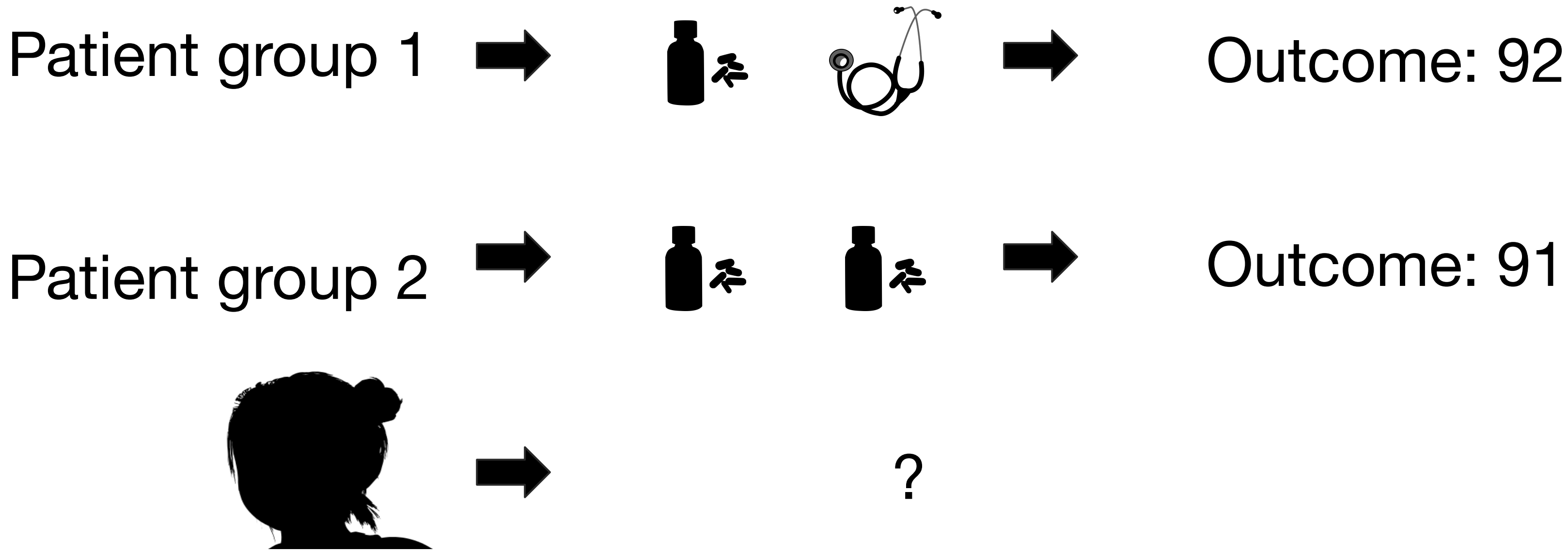
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

Outline for Today

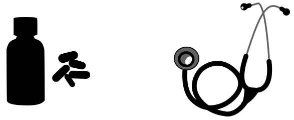
1. Introduction and Setting
2. Offline batch evaluation using models
3. Offline batch evaluation using Q functions
4. Offline batch evaluation using importance sampling

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

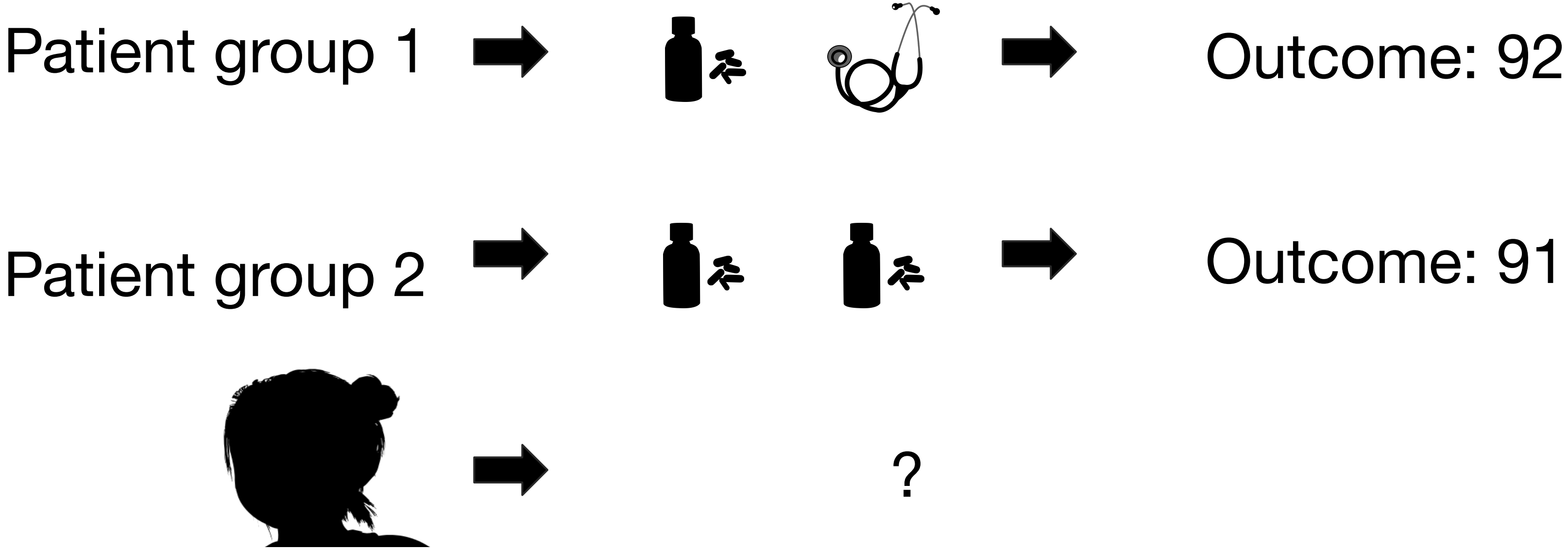
Patient group 1 ➡  ➡ Outcome: 92

Patient group 2 ➡  ➡ Outcome: 91

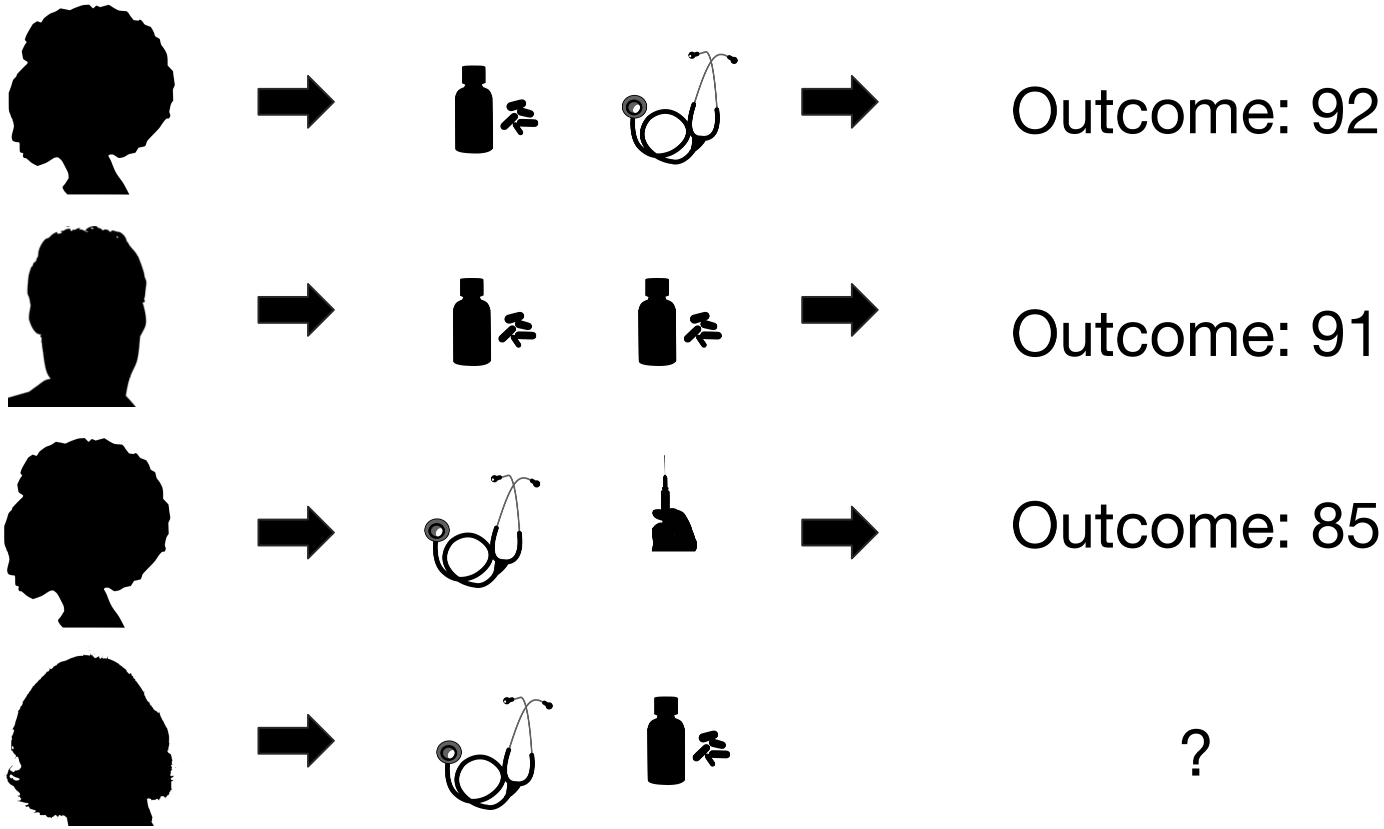
 ➡ ?

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

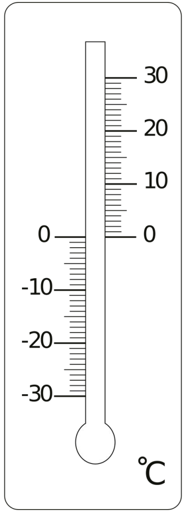
Data Is Censored in that Only Observe Outcomes for Decisions Made



Need for Generalization



Potential Applications



Off Policy Reinforcement Learning

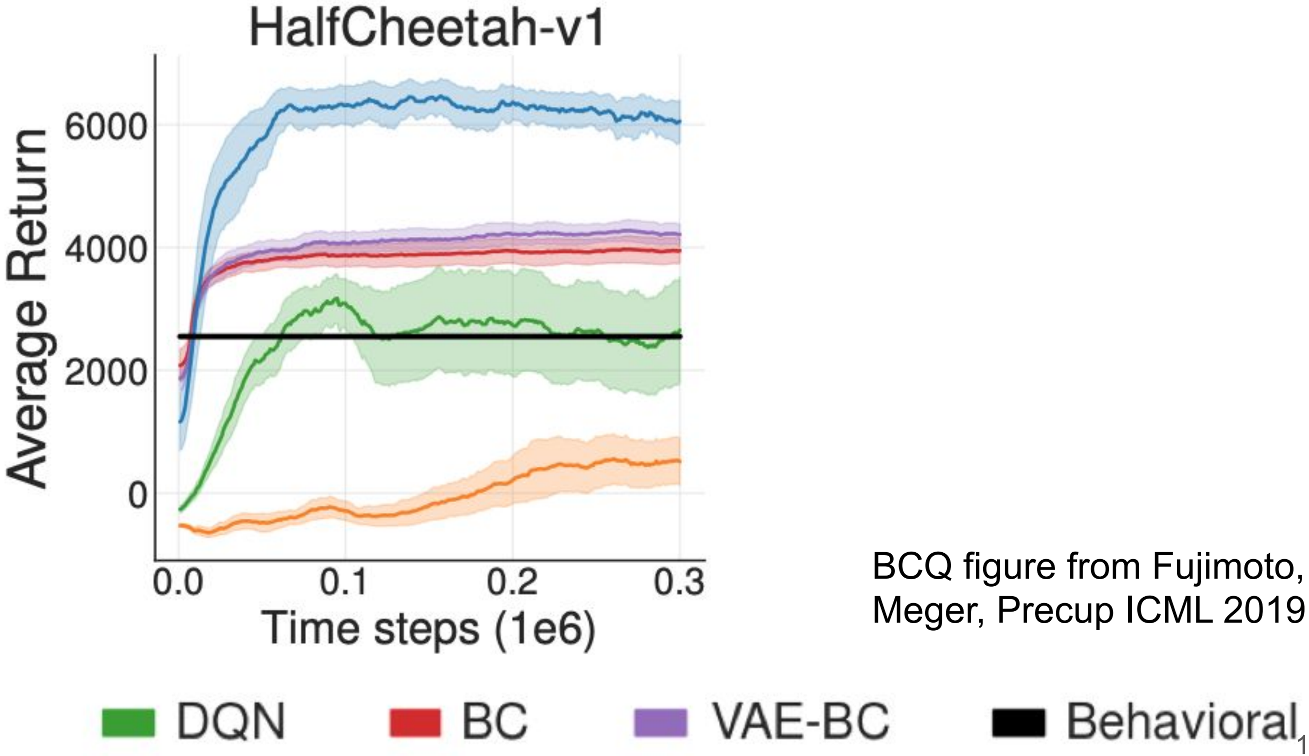
Watkins 1989
Watkins and Dayan 1992
Precup et al. 2000
Lagoudakis and Parr 2002
Murphy 2005
Sutton, Szepesvari and Maei 2009
Shortreed, Laber, Lizotte, Stroup, Pineau, & Murphy 2011
Degirs, White, and Sutton 2012
Mnih et al. 2015
Mahmood et al. 2014
Jiang & Li 2016
Hallak, Tamar and Mannor 2015
Munos,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

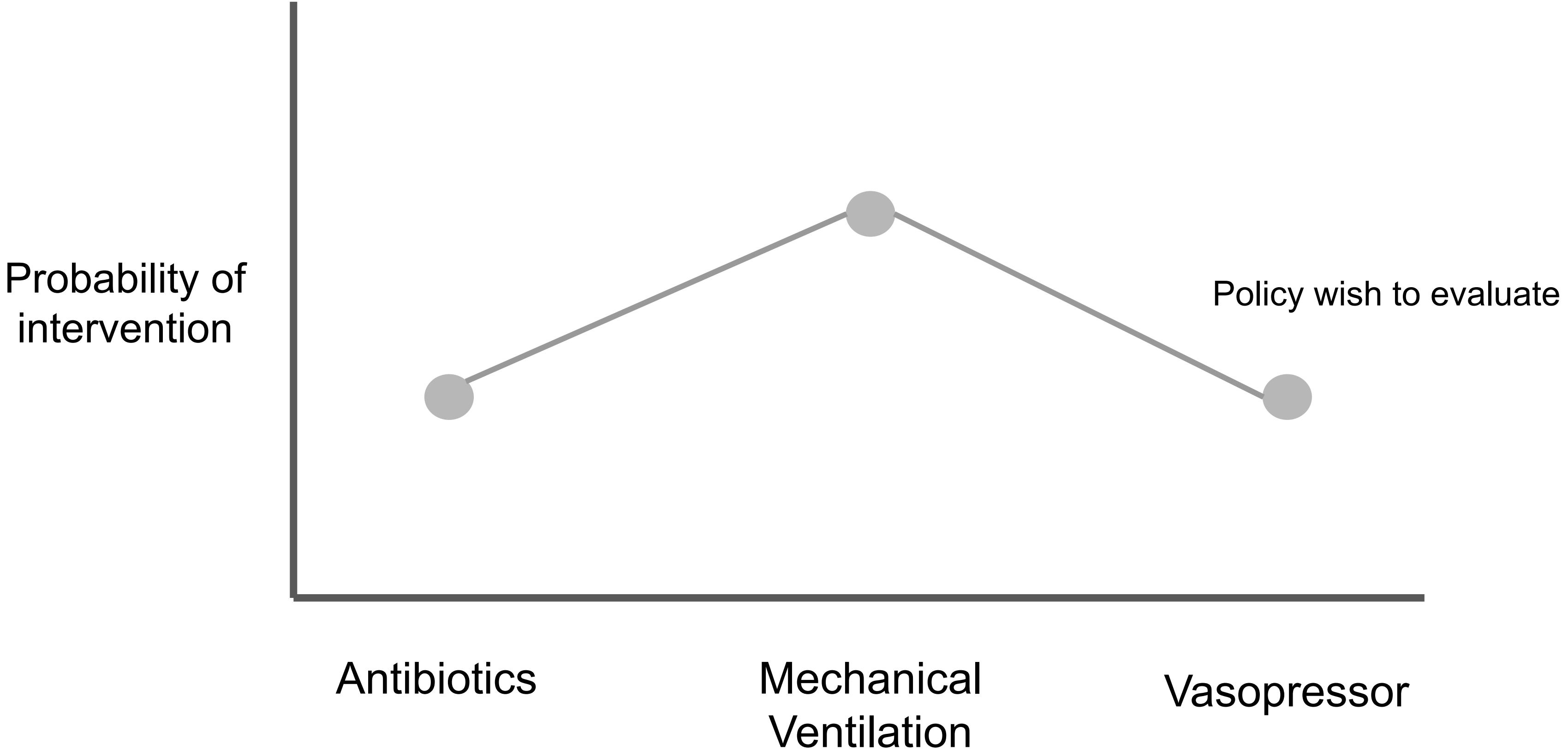
- But deadly triad of bootstrapping, function approximation and off policy, and can fail

Important in Practice

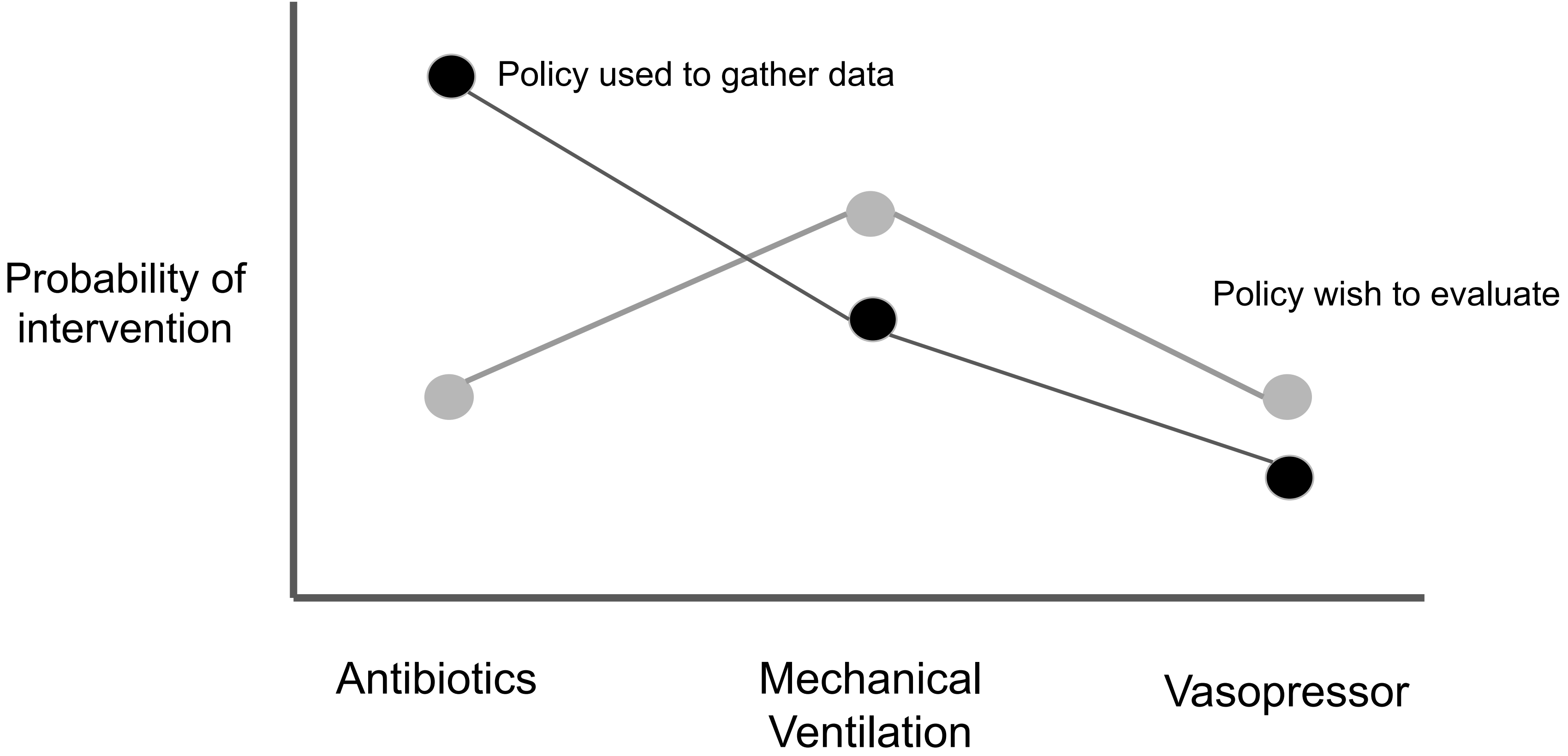


BCQ figure from Fujimoto, Meger, Precup ICML 2019

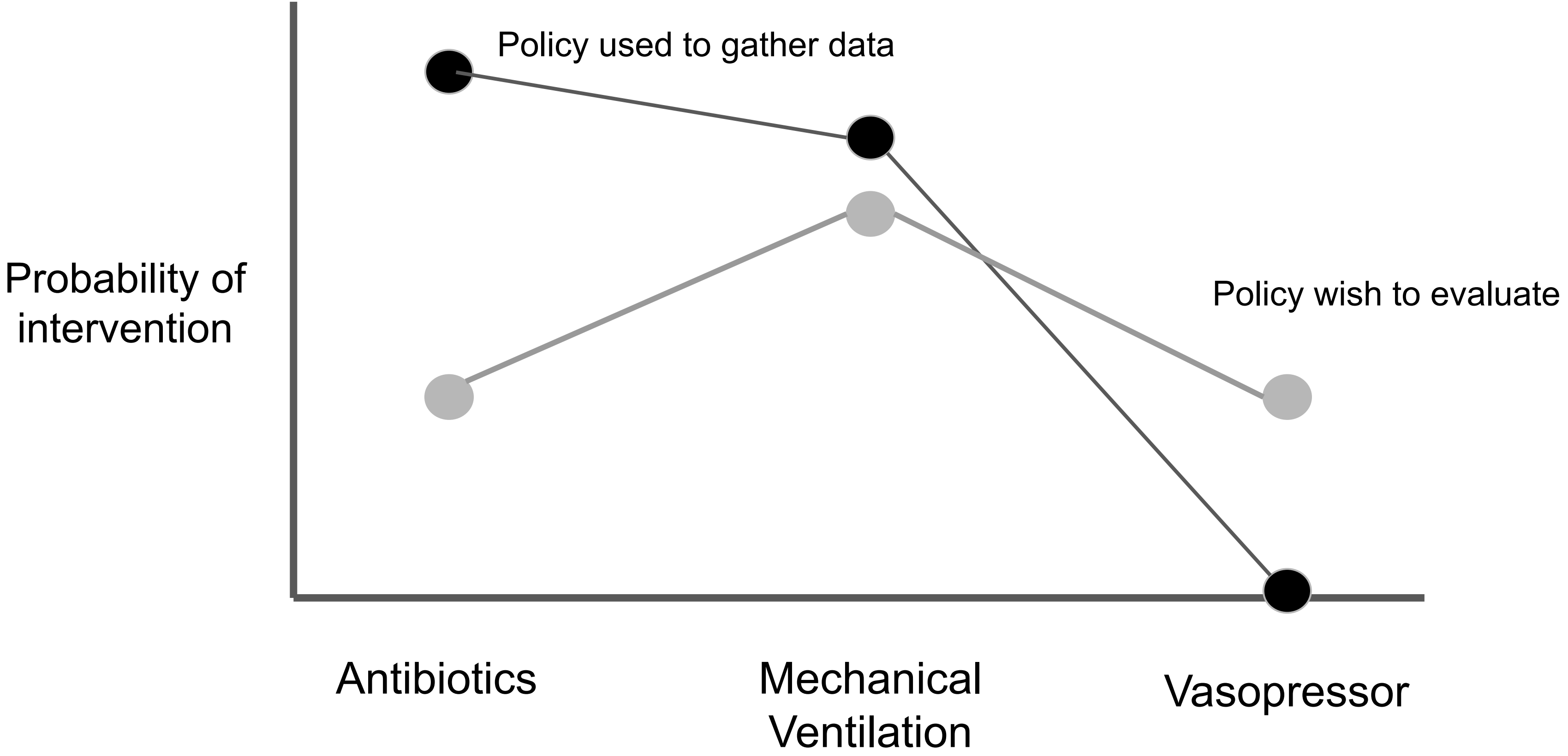
Challenge: Overlap Requirement



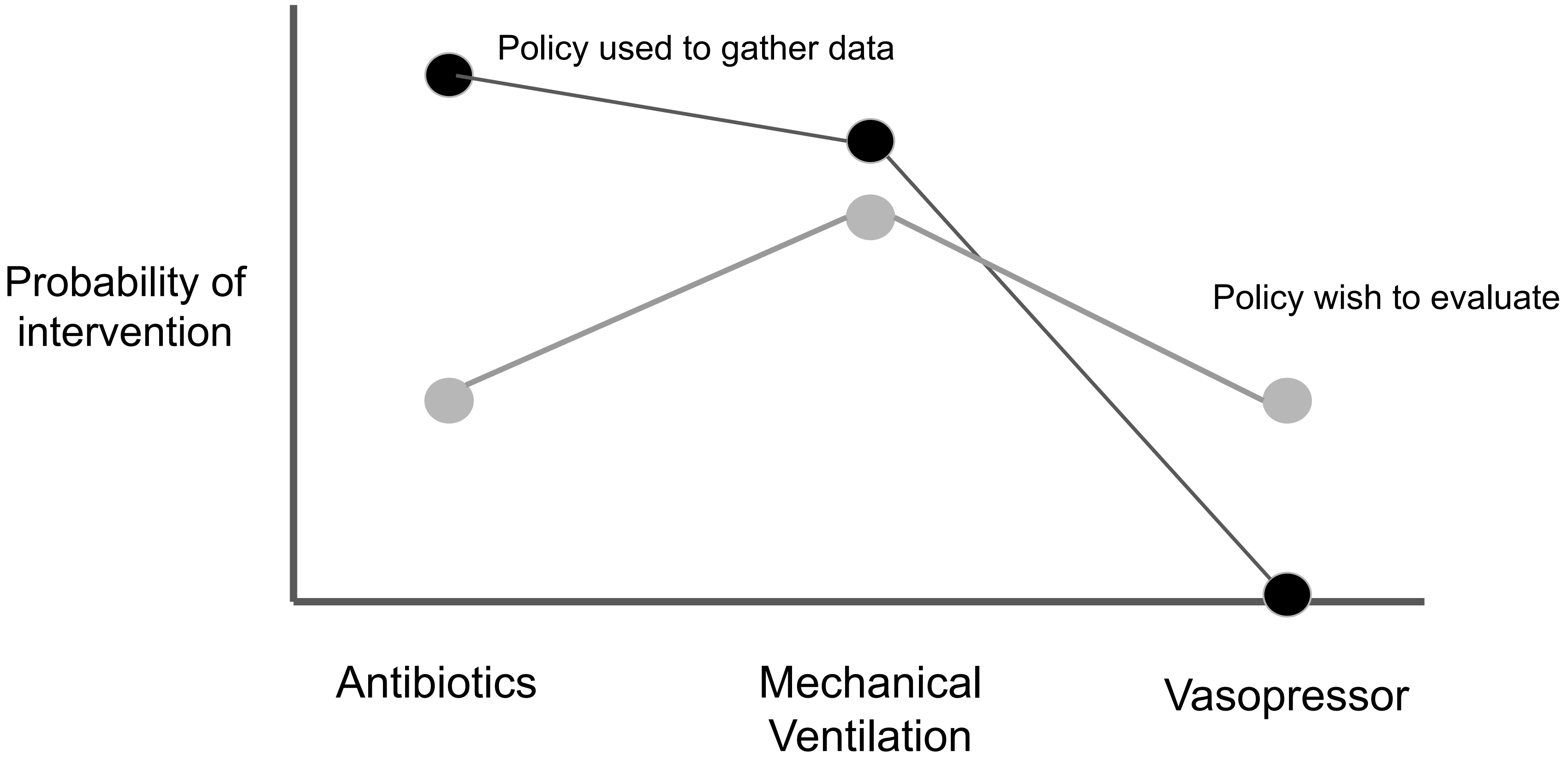
Overlap Requirement: Data Must Support Policy Wish to Evaluate



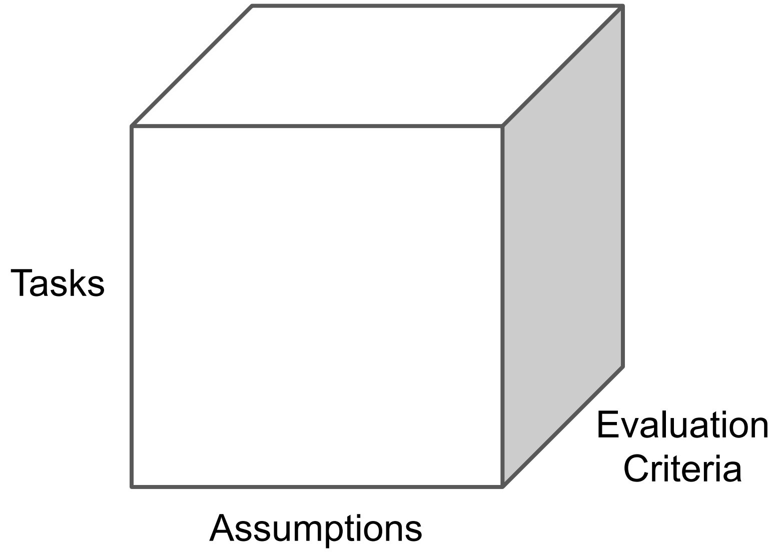
No Overlap for Vasopressor⇒ Can’t Do Off Policy Estimation for Desired Policy



How to Evaluate Sufficient Overlap in Real Data?



Offline / Batch Reinforcement Learning



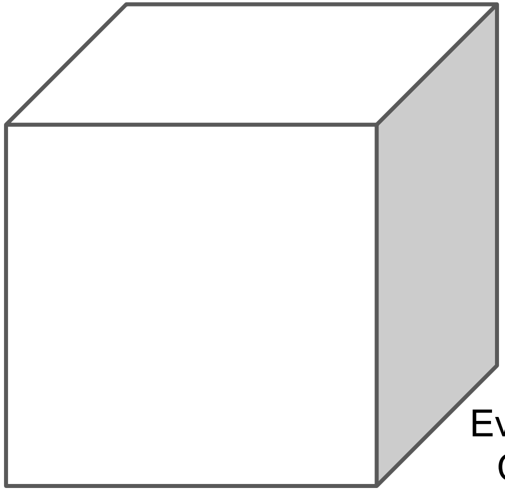
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Common Tasks: Off Policy Evaluation & Optimization

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

Assumptions

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

Common Assumptions

- Stationary process: Policy will be evaluated in or deployed in the same stationary decision process as the behavior policy operated in to gather data
- **Markov**
- Sequential ignorability (no confounding)

$$\{Y(A_{1:(t-1)}, a_{t:T}), S_{t'}(A_{1:(t-1)}, a_{t:(t'-1)})\}_{t'=t+1}^{T} \perp\!\!\!\perp A_t \,\Big|\, \mathcal{F}_t$$

- Overlap

$$\forall (s,a)\ \mu_e(s,a) > 0 \quad \rightarrow \mu_b(s,a) > 0$$

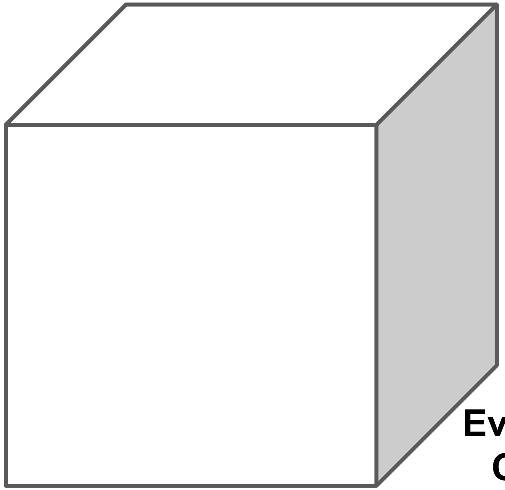
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Common Tasks: Off Policy Evaluation & Optimization

Tasks

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

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



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

Off Policy Reinforcement Learning

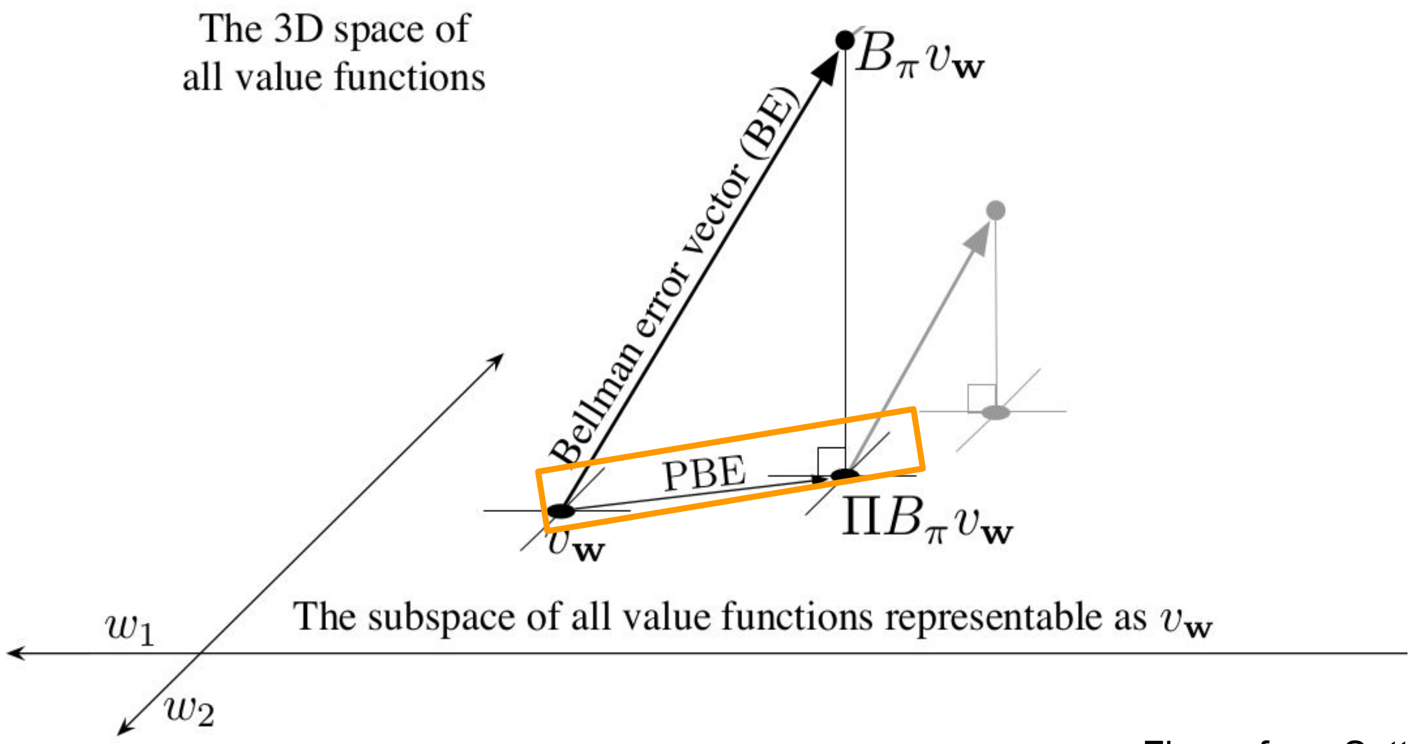
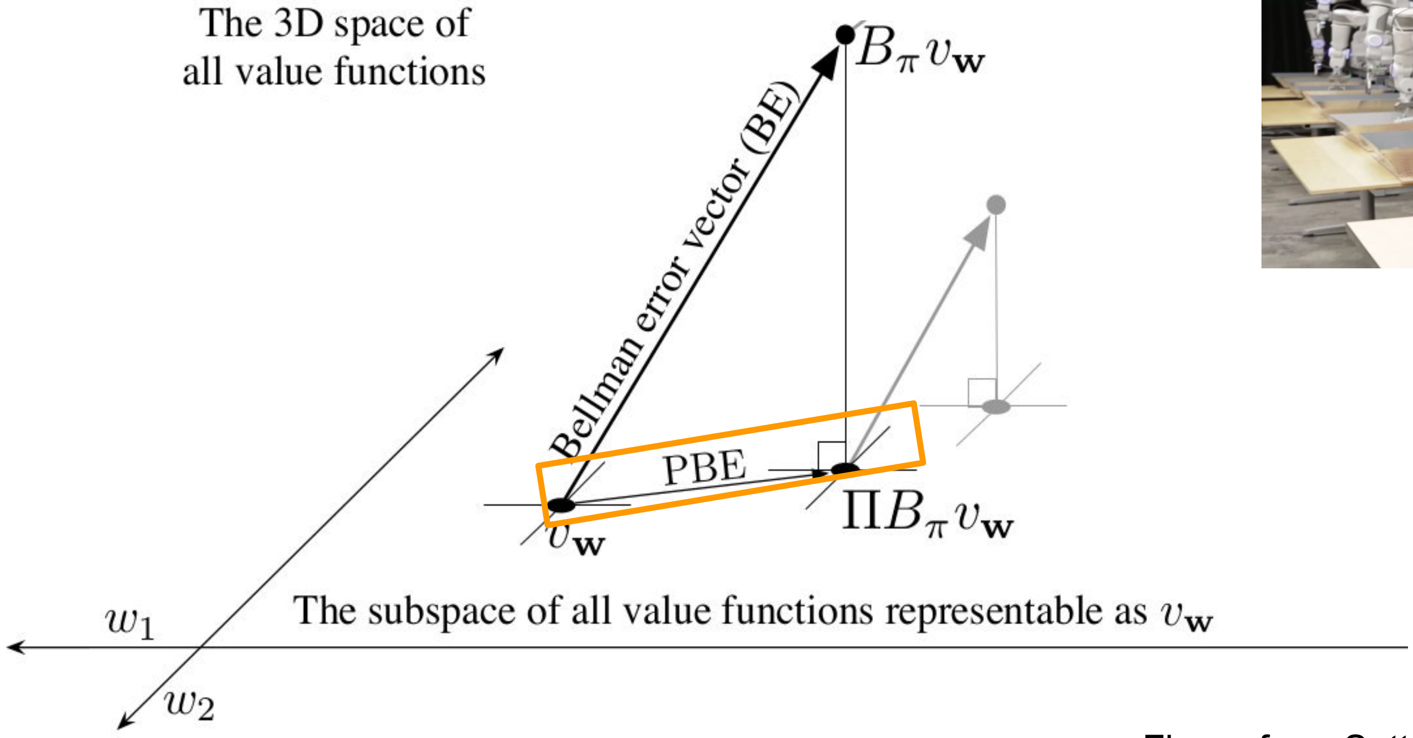


Figure from Sutton & Barto 2018

Off Policy Reinforcement Learning



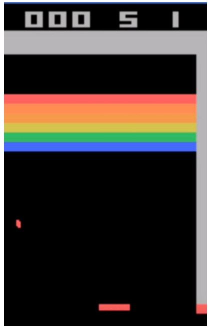


Figure from Sutton & Barto 2018

Batch Off Policy Reinforcement Learning

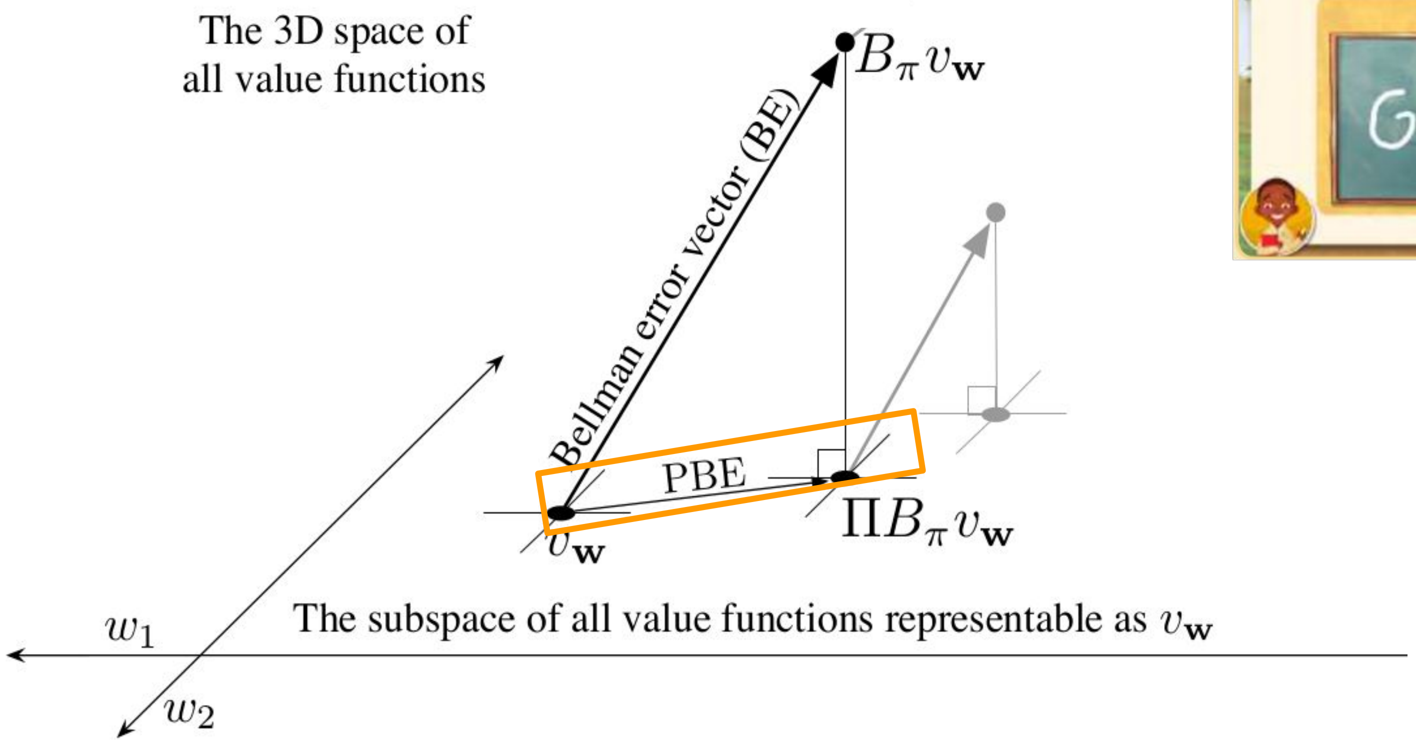


Figure from Sutton & Barto 2018

Batch Off Policy Reinforcement Learning

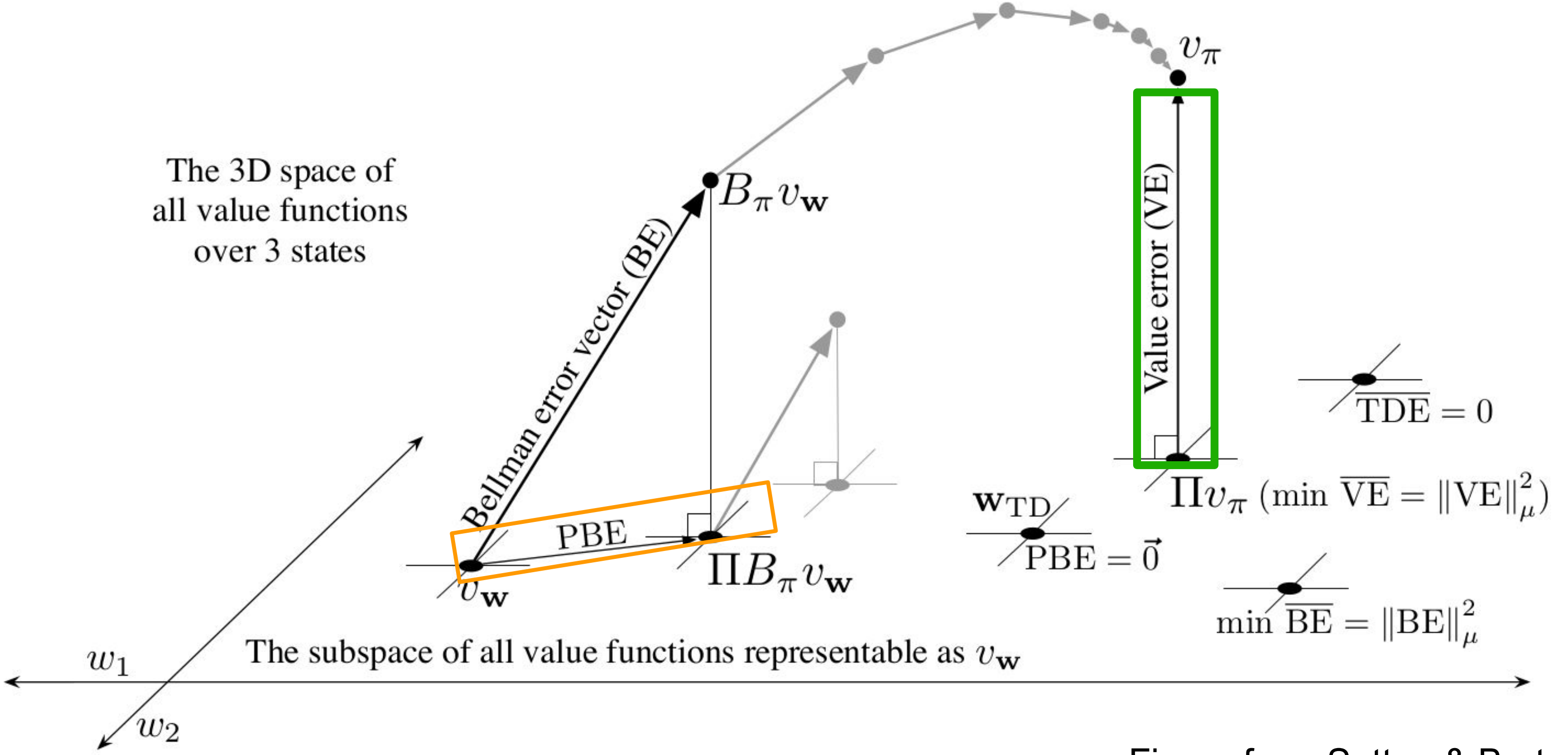


Figure from Sutton & Barto 2018

Common Evaluation Criteria for Off Policy Evaluation

- Computational efficiency
- Performance accuracy

$$\forall \mathcal{D}_i \in \{\mathcal{D}_1 \sim \mathcal{M}_1, \mathcal{D}_2 \sim \mathcal{M}_2, \ldots, \mathcal{D}_K \sim \mathcal{M}_K\} \quad \frac{1}{|\rho|} \sum_{s_0 \in \rho} (\hat{V}^{\pi}_{\mathcal{M}_i}(s_0, \mathcal{D}_i) - V^{\pi}_{\mathcal{M}_i}(s_0))^2$$

$$\lim_{|\mathcal{D}| \to \infty} \frac{1}{|\rho|} \sum_{s_0 \in \rho} \hat{V}^{\pi}(s_0, \mathcal{D}) \to \frac{1}{|\rho|} \sum_{s_0 \in \rho} V^{\pi}(s_0)$$

$$\frac{1}{|\rho|} \sum_{s_0 \in \rho} \hat{V}^{\pi}(s_0, \mathcal{D}) \leq \frac{1}{|\rho|} \sum_{s_0 \in \rho} V^{\pi}(s_0) - f(n, \ldots)$$

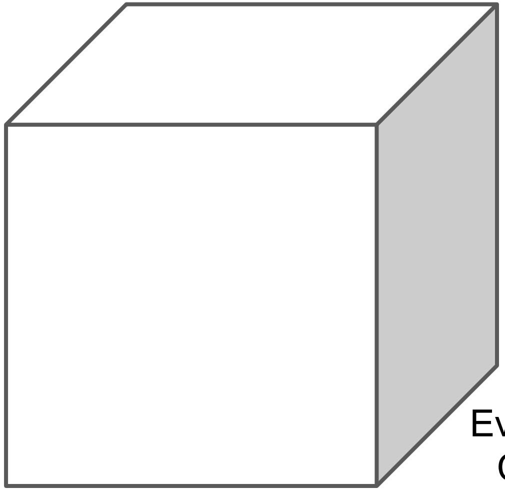
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \to a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

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

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

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

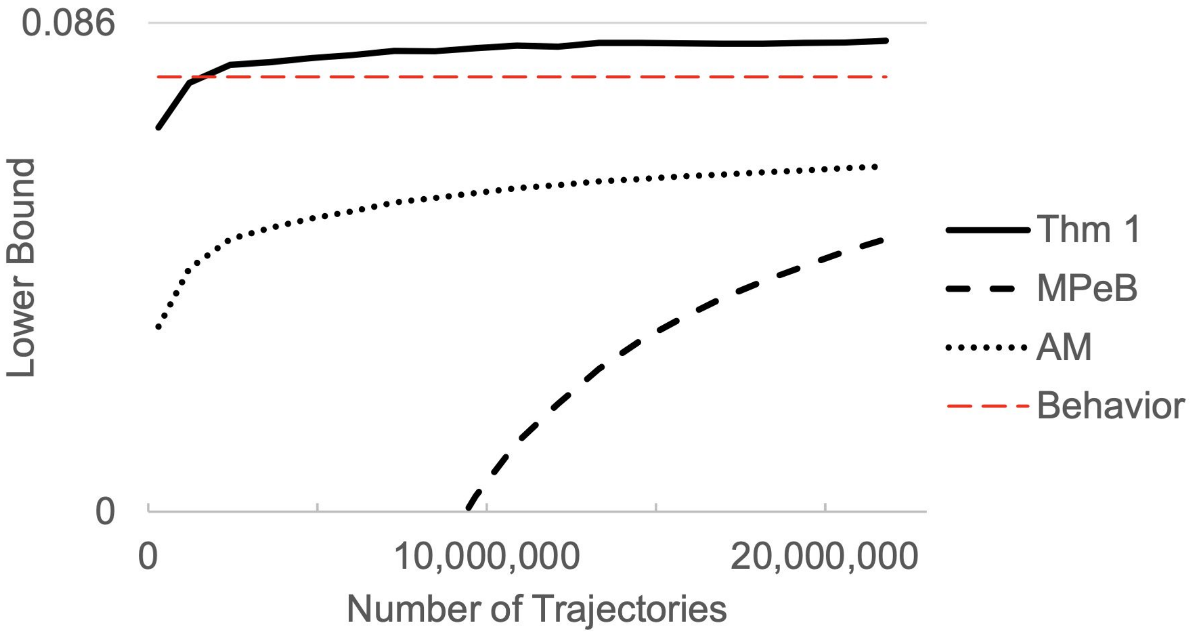
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Batch Policy Evaluation: Estimate the Performance of a Particular Decision Policy

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

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

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.

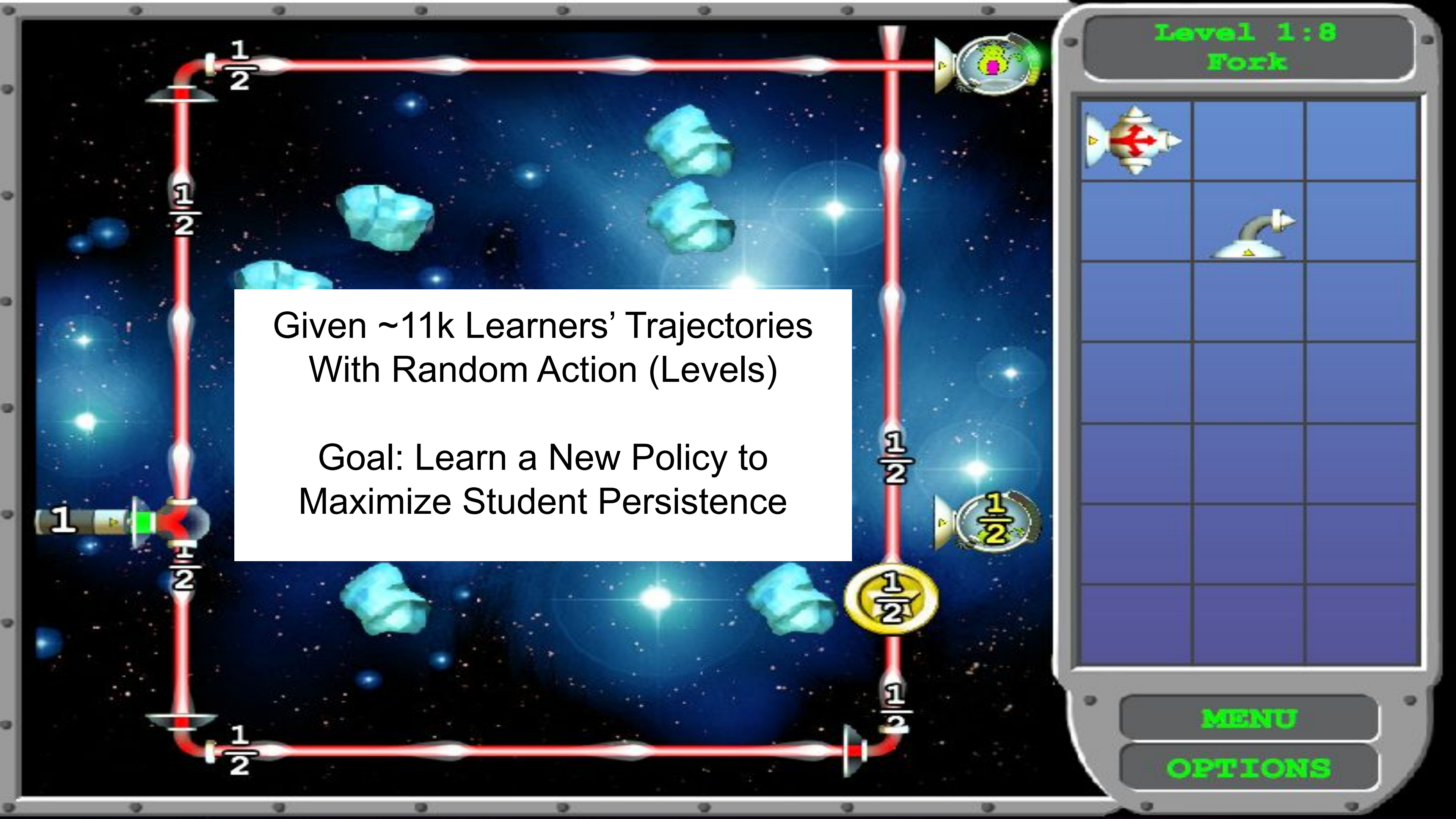
Outline

1. Introduction and Setting
2. **Offline batch evaluation using models**
3. Offline batch evaluation using Q functions
4. Offline batch evaluation using importance sampling
5. Safe batch RL


Level 1:8
Fork
MENU
OPTIONS

Took > 30s

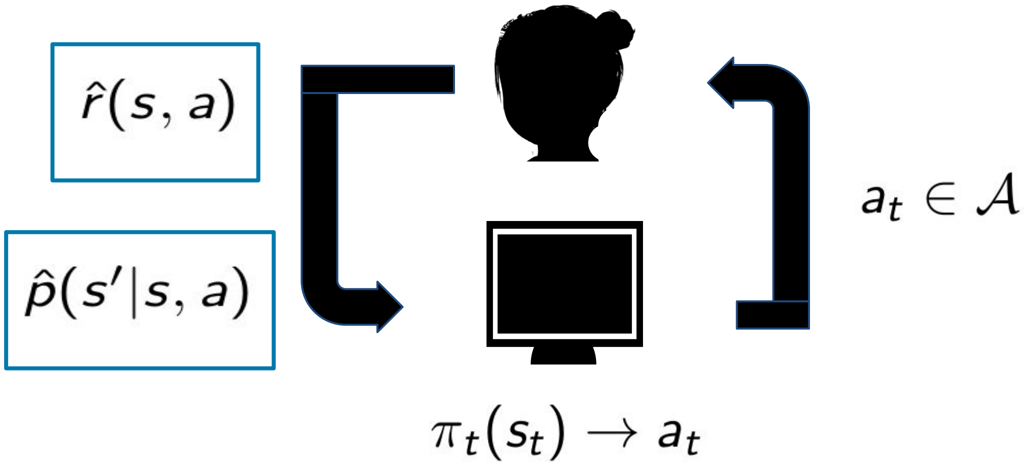
Took <= 30s



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

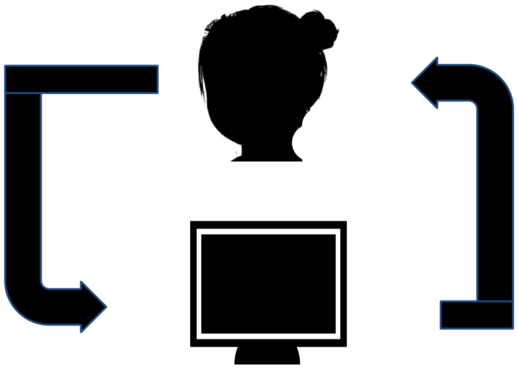
Goal: Learn a New Policy to
Maximize Student Persistence

Learn Dynamics and Reward Models from Data



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

Learn Dynamics and Reward Models from Data, Evaluate Policy



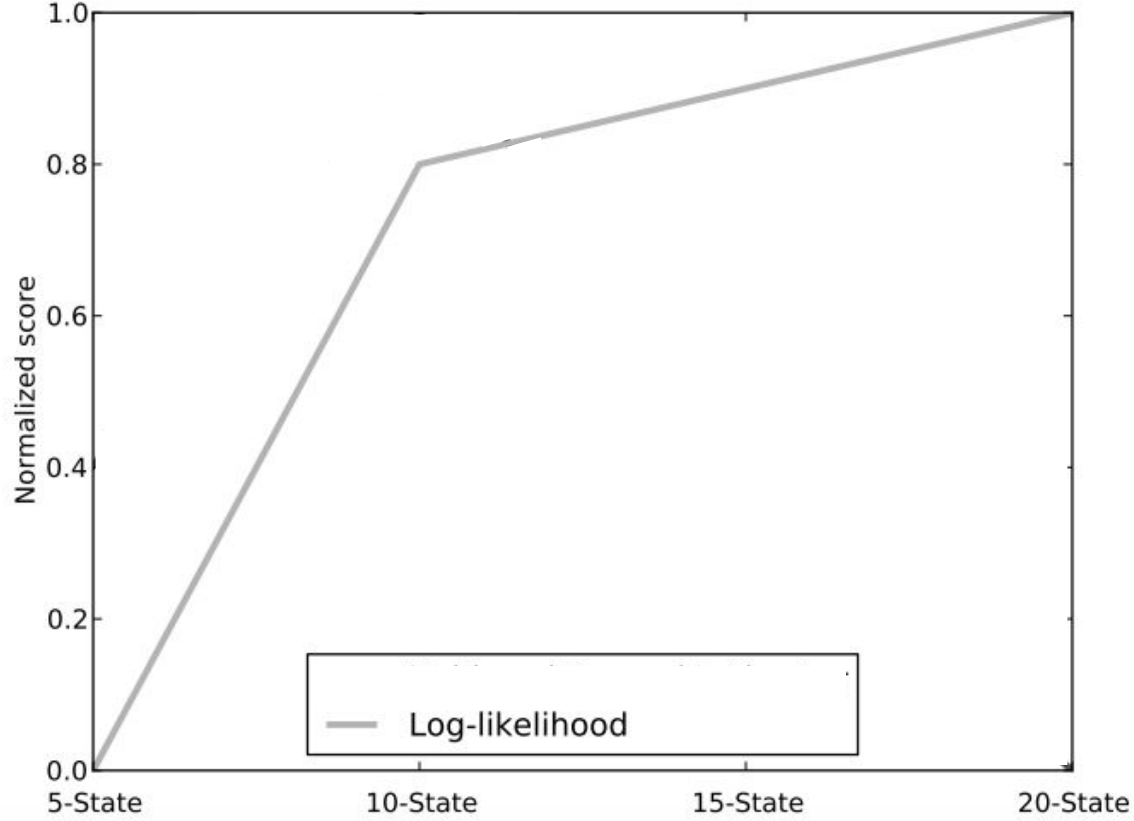
$$V^{\pi} \approx (I - \gamma \hat{P}^{\pi})^{-1} \hat{R}^{\pi}$$

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

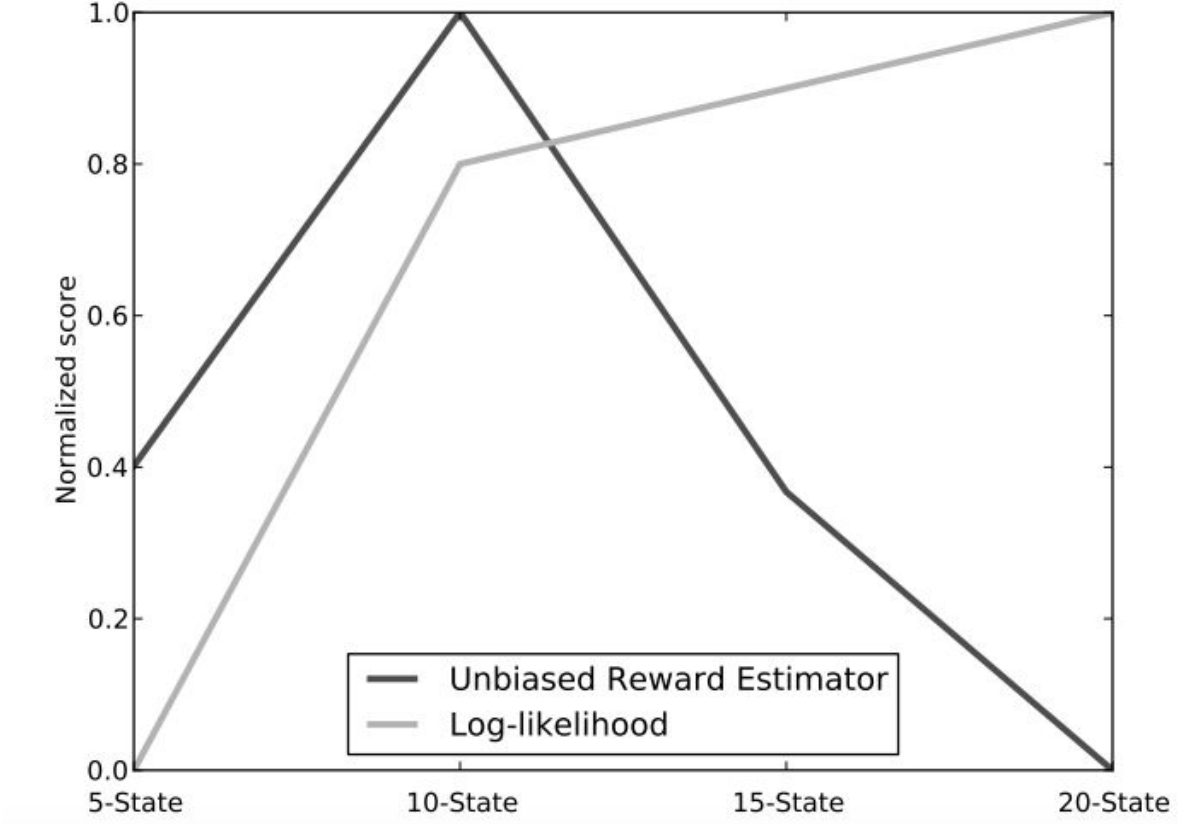
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

- Mannor, Simster, Sun, Tsitsiklis 2007

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

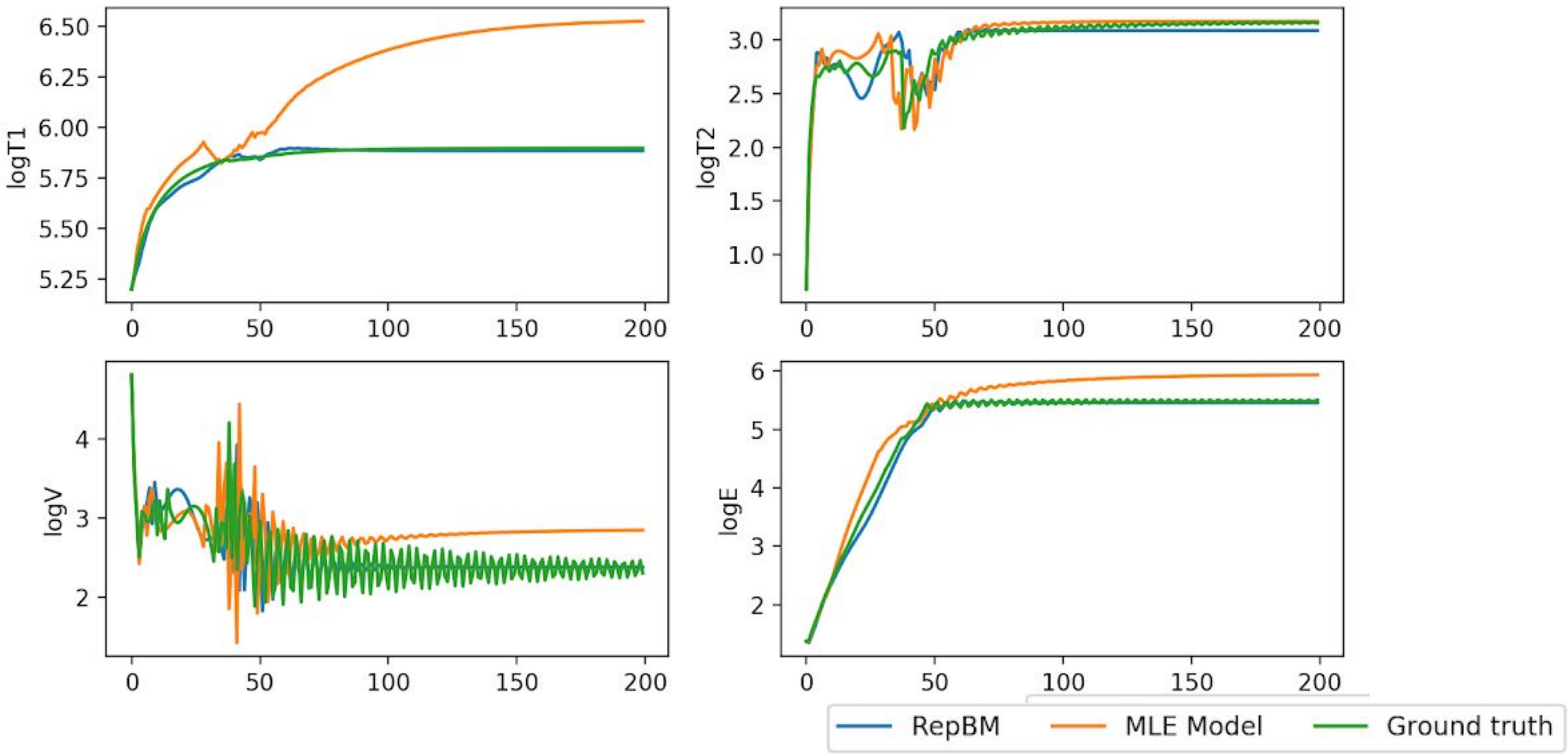


Better Dynamics/Reward Models for Existing Data, May **Not** Lead to Better Policies for Future Use → Bias due to Model **Misspecification**



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

1. Introduction and Setting
2. Offline batch evaluation using models
3. **Offline batch evaluation using Q functions**
4. Offline batch evaluation using importance sampling

Model Free Value Function Approximation: Fitted Q Evaluation

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

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

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

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

- Fitted Q evaluation, LSTD, ...

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** $k = 1, 2, \ldots, K$ **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: $\widehat{C}^\pi(x) = Q_K(x, \pi(x)) \quad \forall x$

Let's assume we use a DNN for F.

What is different vs DQN?

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

Example Fitted Q Evaluation Guarantees

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

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 $n = O\Big(\frac{\overline{C}^4}{\epsilon^2}(\log\frac{K}{\delta} + \mathtt{dim_F}\log\frac{\overline{C}^2}{\epsilon^2} + \log\mathtt{dim_F})\Big)$, we have with probability $1-\delta$:*

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

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

Le, Voloshin, Yue ICML 2019

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

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

Outline

1. Introduction and Setting
2. Offline batch evaluation using models
3. Offline batch evaluation using Q functions
4. **Offline batch evaluation using importance sampling**

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*

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

***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

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

- 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 from 3 arms. Consider that arm 1 is a Bernoulli where with probability .98 we get 0 and with probability 0.02 we get 100. Arm 2 is a Bernoulli where with probability 0.55 the reward is 2 else the reward is 0. Arm 3 has a probability of yielding a reward of 1 with probability 0.5 else it gets 0. Select all that are true.

- Data is sampled from pi1 where with probability 0.8 it pulls arm 3 else it pulls arm 2. The policy we wish to evaluate, pi2, pulls arm 2 with probability 0.5 else it pulls arm 1. pi2 has higher true reward than pi1.
- We cannot use pi1 to get an unbiased estimate of the average reward pi2 using importance sampling.
- If rewards can be positive or negative, we can still get a lower bound on pi2 using data from pi1 using importance sampling
- Now assume pi1 selects arm1 with probability 0.2 and arm2 with probability 0.8. We can use importance sampling to get an unbiased estimate of pi2 using data from pi1.
- Still with the same pi1, it is likely with N=20 pulls that the estimate using IS for pi2 will be higher than the empirical value of pi1.
- Not Sure

Check Your Understanding: Importance Sampling Answers

We can use importance sampling to do batch bandit policy evaluation. Consider we have a dataset for pulls from 3 arms. Consider that arm 1 is a Bernoulli where with probability .98 we get 0 and with probability 0.02 we get 100. Arm 2 is a Bernoulli where with probability 0.55 the reward is 2 else the reward is 0. Arm 3 has a probability of yielding a reward of 1 with probability 0.5 else it gets 0. Select all that are true.

- Data is sampled from pi1 where with probability 0.8 it pulls arm 3 else it pulls arm 2. The policy we wish to evaluate, pi2, pulls arm 2 with probability 0.5 else it pulls arm 1. pi2 has higher true reward than pi1. **(True)**
- We cannot use pi1 to get an unbiased estimate of the average reward pi2 using importance sampling. **(True, pi1 never pulls arm 1 which is taken by pi2)**
- If rewards can be positive or negative, we can still get a lower bound on pi2 using data from pi1 using importance sampling **(False, only if rewards are positive)**
- Now assume pi1 selects arm1 with probability 0.2 and arm2 with probability 0.8. We can use importance sampling to get an unbiased estimate of pi2 using data from pi1. (True)
- Still with the same pi1, it is likely with N=20 pulls that the estimate using IS for pi2 will be higher than the empirical value of pi1. (False)

Importance Sampling for RL Policy Evaluation

$$V^{\pi}(s) \quad = \quad \sum_{\tau} p(\tau|\pi, s) R(\tau)$$

$\mathcal{D}$: 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

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

$\mathcal{D}$: 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

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

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \to a$
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!

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

$\mathcal{D}$: 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

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

Check Your Understanding: Importance Sampling 2 Answers

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

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. **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?**

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

Outline

1. Introduction and Setting
2. Offline batch evaluation using models
3. Offline batch evaluation using Q functions
4. Offline batch evaluation using importance sampling

Recent Directions

- 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
- Blended estimators

What You Should Know

- Be able to define and apply importance sampling for off policy policy evaluation
- Define some limitations of IS (variance)
- Define why we might want to do batch offline RL policy evaluation and potential applications
- Be aware of the main potential limitations of model and model free methods

Class Organization

- Fast reinforcement learning
- **Learning from offline data**
  - **Overview and Policy evaluation**
  - **Policy optimization**