RL for People

Emma Brunskill
Assistant Professor, Computer Science, Stanford
Thanks to Christoph Dann, Andrea Zanette, Phil Thomas, and Xinkun Nie for some figures & slides

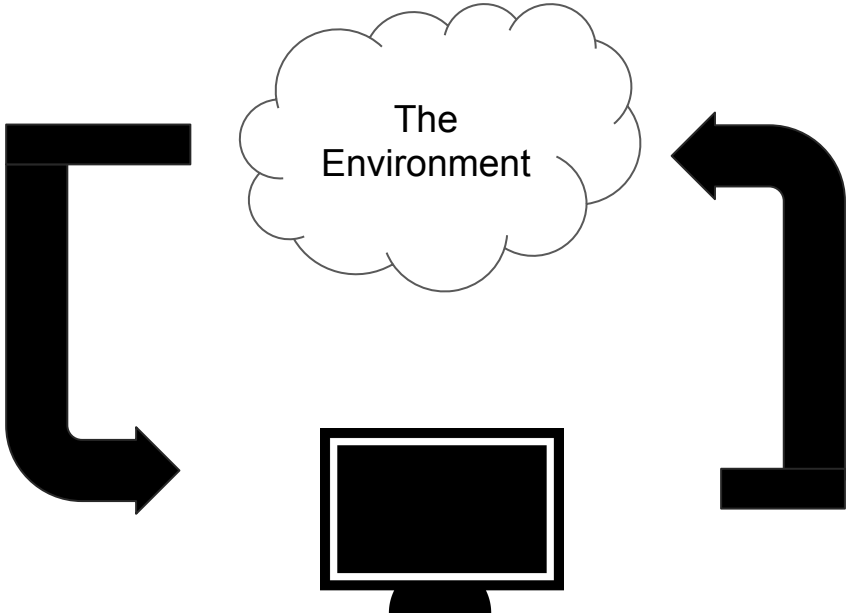
Sequential Decision Making Under Uncertainty

For a doctor does not deliberate whether he shall heal…
assume(s) the end and consider how and by what means it is obtained…
if it is achieved by one means …
consider how it will be achieved… till they come to the first cause…
and what is last in this order of analysis seems to be the first in the order of becoming

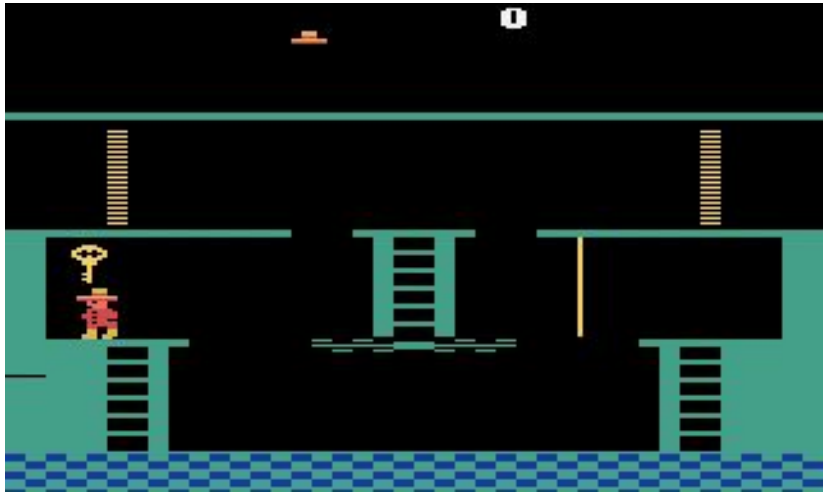
Aristotle

Learning to Make Good Sequences of Decisions Under Uncertainty

Learning to Make Good Sequences of Decisions Under Uncertainty ⇒ 1980s Reinforcement Learning



2010s: A New Era of Reinforcement Learning



Mastering the Game of Go Without Human Knowledge

- Silver et al. Nature 2017

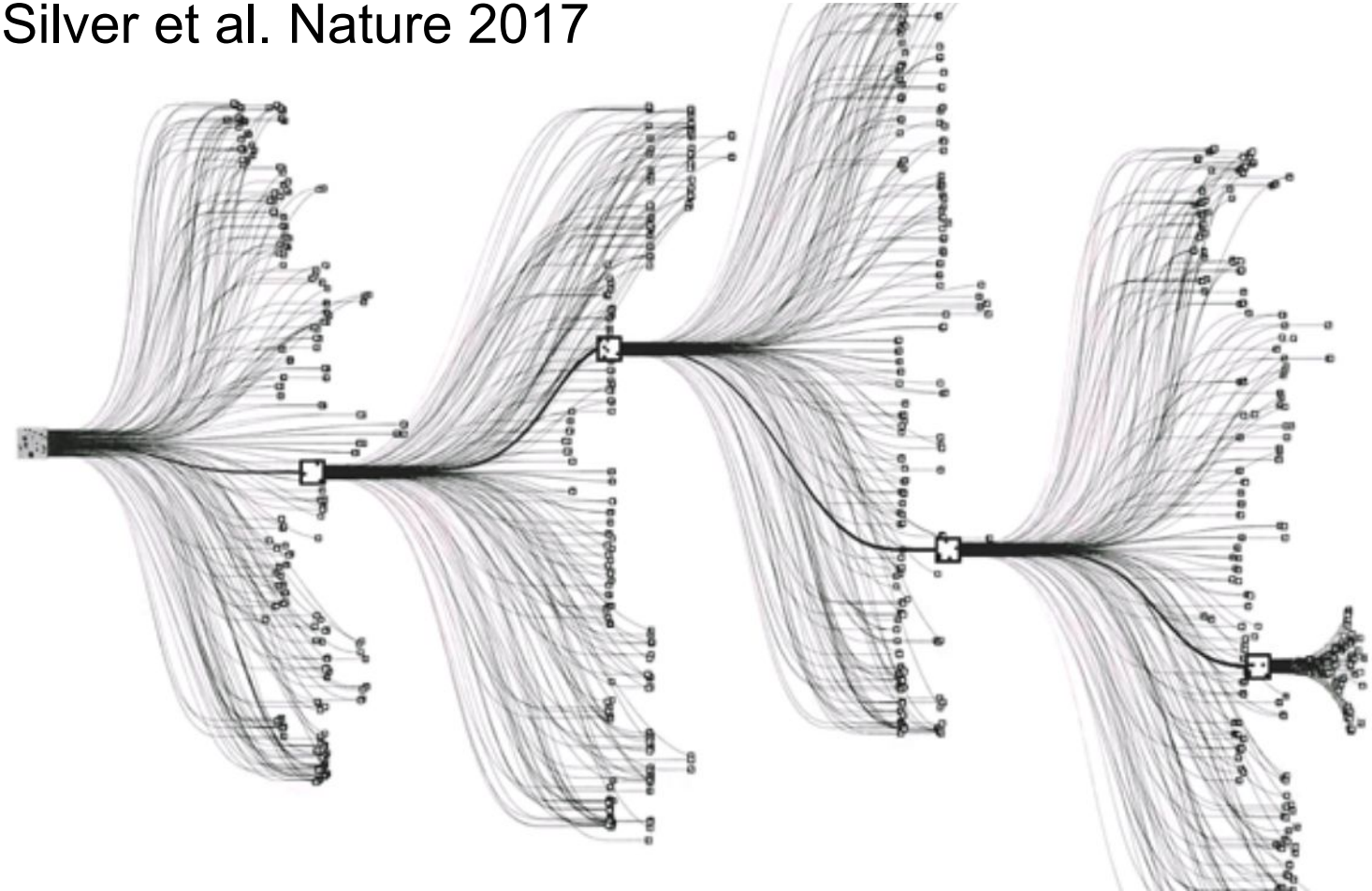


Image from DeepMind

Is Reinforcement Learning “Solved”?

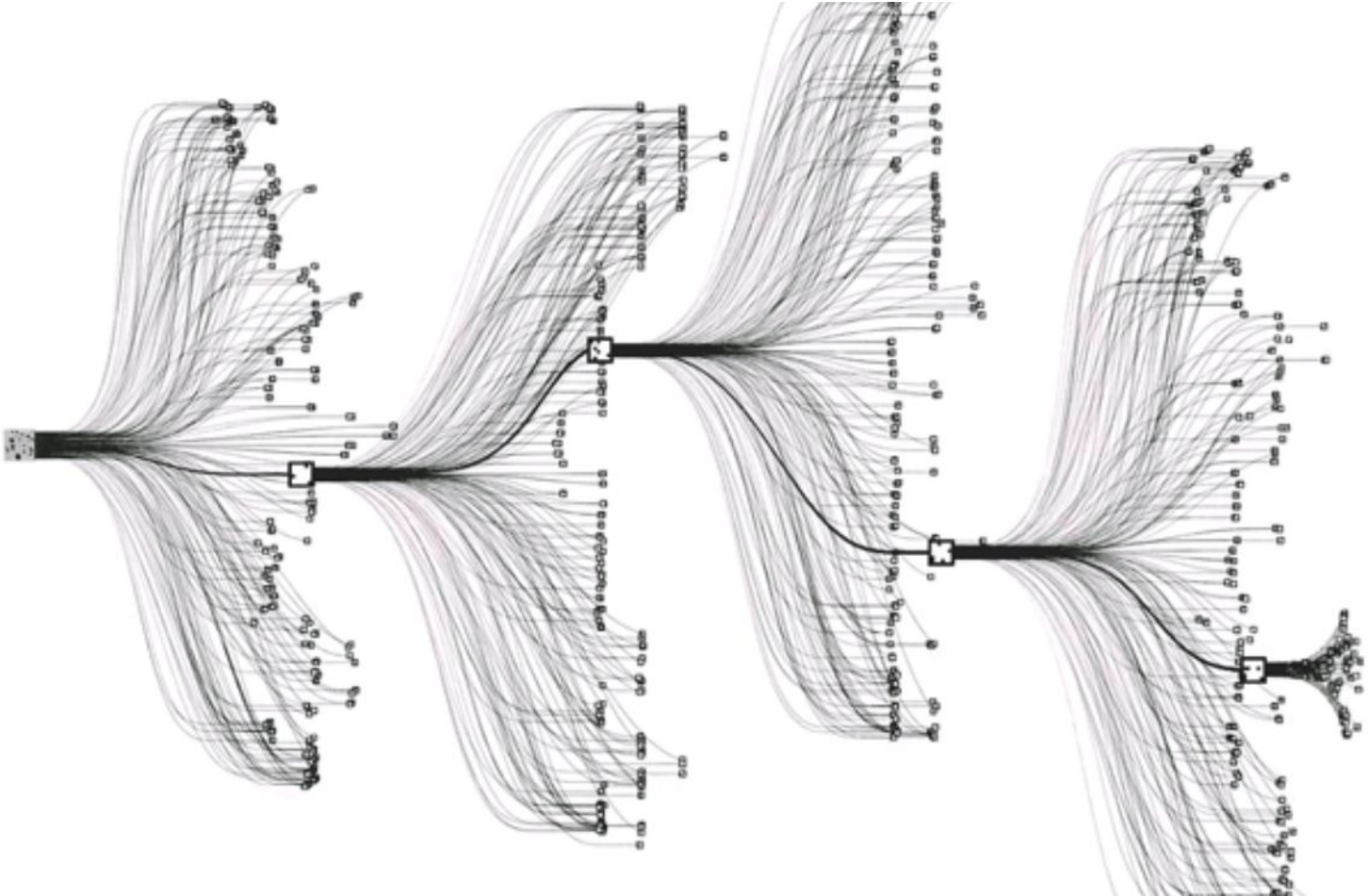


Image from DeepMind

Parallel Legacy of “RL” to Benefit People

https://web.stanford.edu/group/cslipublications/cslipublications/SuppesCorpus/Professional%20Photos/album/1960s/slides/5.html

VS

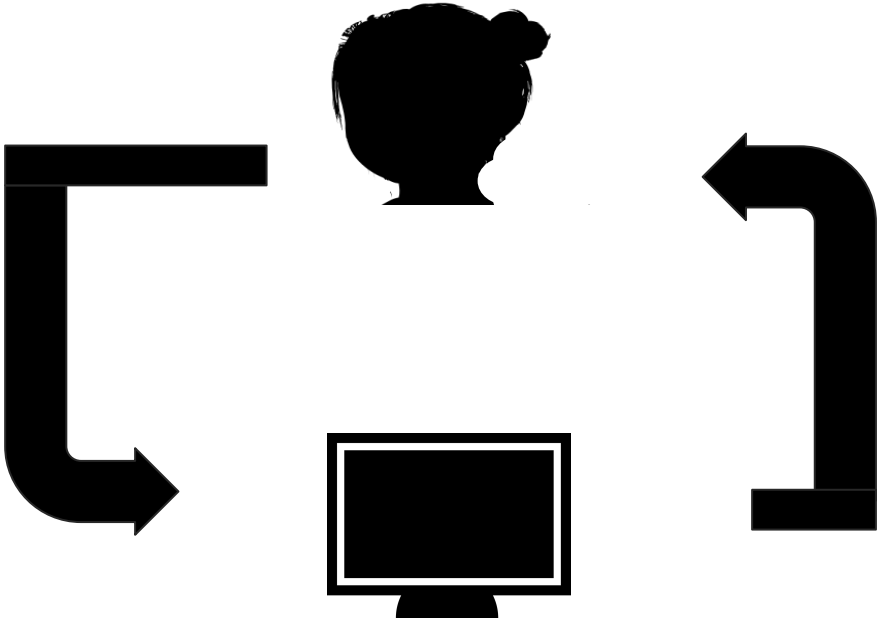
- Simulator of domain
- Enormous data to train
- Can always try out a new strategy in domain

VS

- No good simulator of human physiology, behavior & learning
- Gathering real data involves impacting real people

- Simulator of domain
- Enormous data to train
- Can always try out a new strategy in domain

Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions



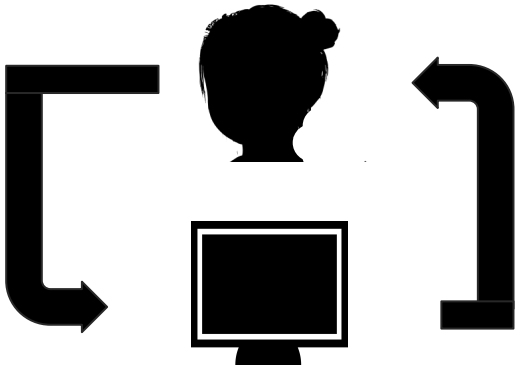
And if can learn to make good decisions faster, benefit more people

Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

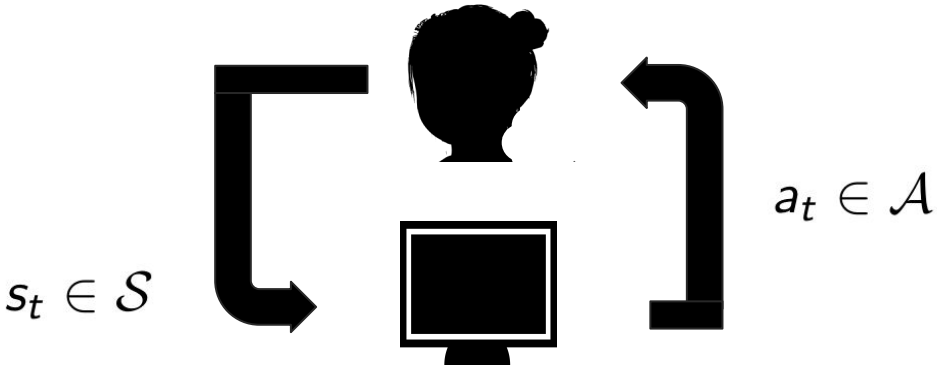
1. Exploration: how to quickly gather information to learn to make good decisions
2. Counterfactual /batch reinforcement learning: leveraging past data

Background: Markov Decision Process

$s_t \in \mathcal{S}$

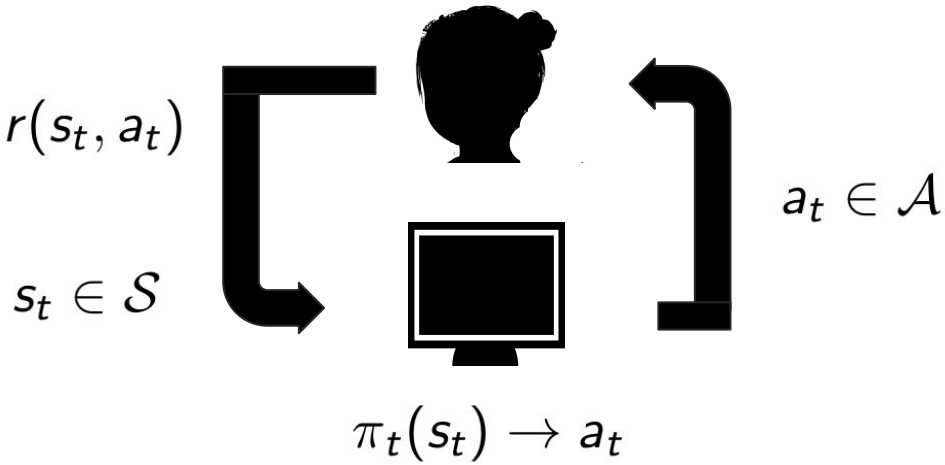


Background: Markov Decision Process

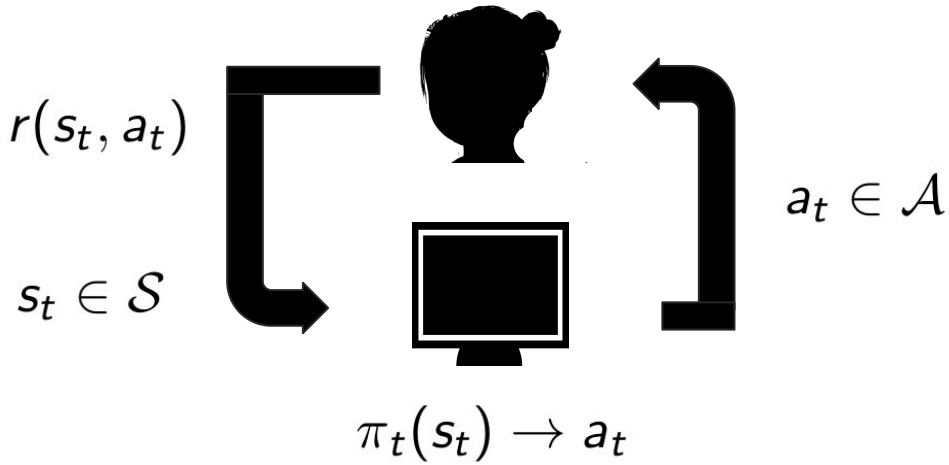


$$\pi_t(s_t) \rightarrow a_t$$

Background: Markov Decision Process

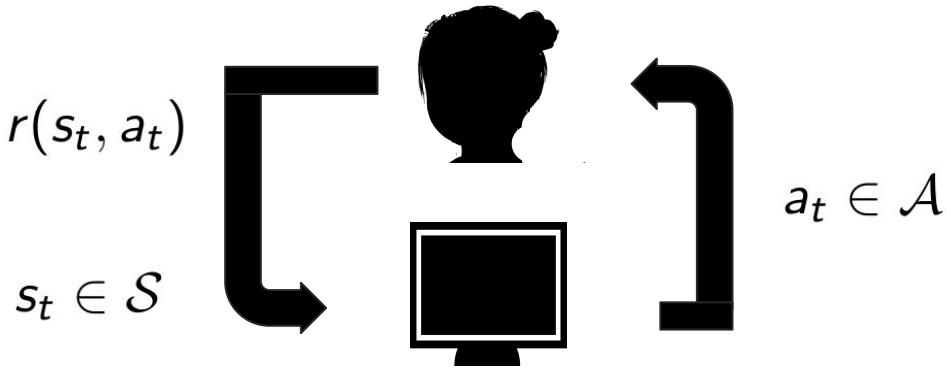


Background: Markov Decision Process Value Function



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

Background: Markov Decision Process Value Function

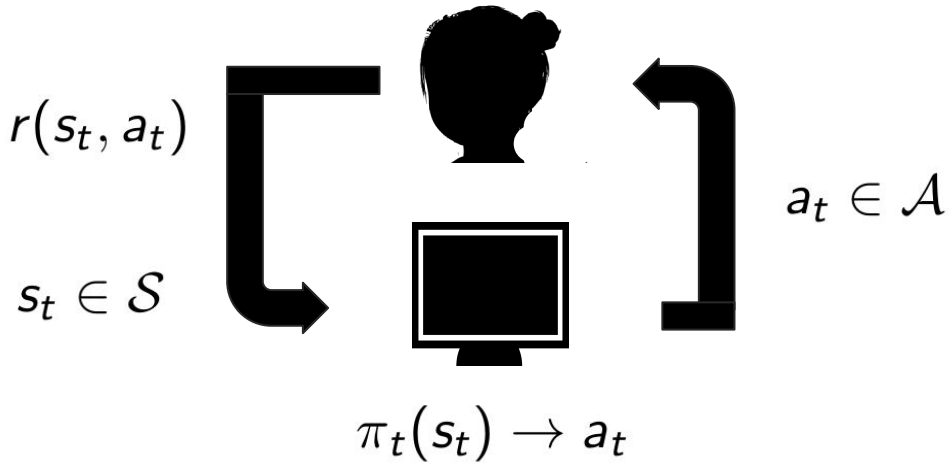


$$\pi_t(s_t) \rightarrow a_t$$

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

$$V^*(s) = \max_\pi V^\pi(s)$$

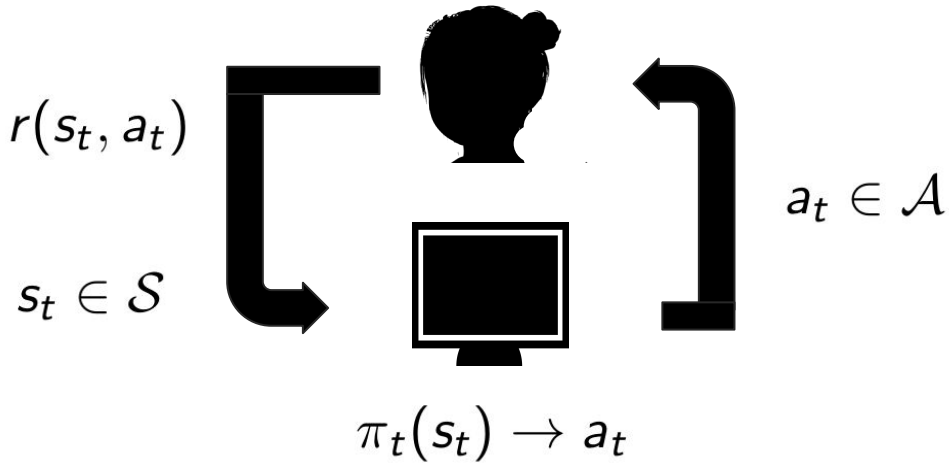
Background: Reinforcement Learning



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

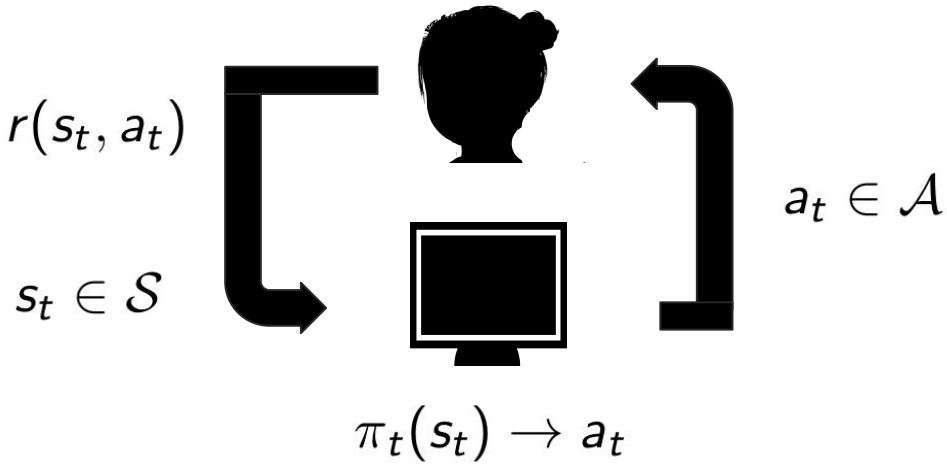
Only observed through samples (experience)

Background: Probably Approximately Correct RL



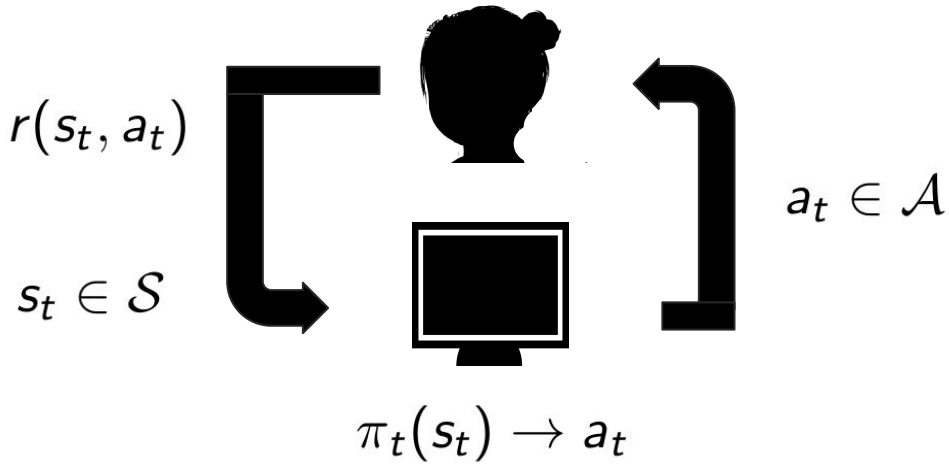
$$P\Big(\underbrace{\sum_t \mathbb{1}\left(V^{\pi_t}(s_t) < V^*(s_t) - \epsilon\right)}_{\text{Sample complexity}} \leq F(\epsilon, \delta, \mathcal{S}, \mathcal{A})\Big) \geq 1 - \delta$$

Background: Regret Bounds for RL



$$P\Big(\underbrace{\sum_t r(s_t, \pi^*(s_t)) - \sum_t r(s_t, \pi_t(s_t))}_{\text{Regret}} \leq F(\delta, \mathcal{S}, \mathcal{A}, T)\Big) \geq 1-\delta$$

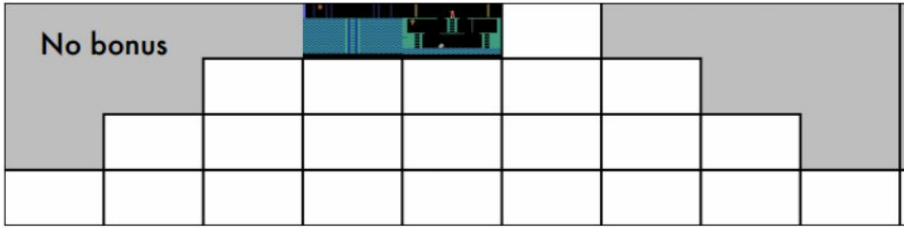
Impact of Strategic Exploration / Efficient RL

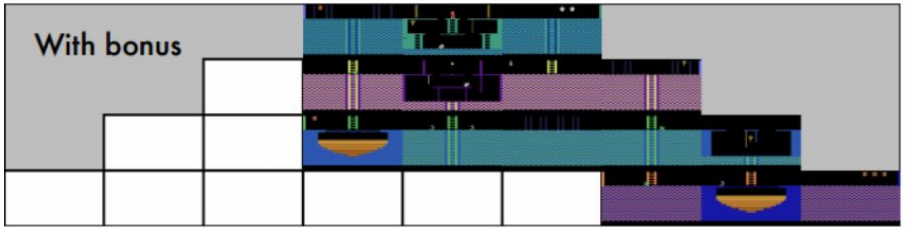


$$P\Big(\underbrace{\sum_t r(s_t, \pi^*(s_t)) - \sum_t r(s_t, \pi_t(s_t))}_{\text{Regret}} \leq F(\delta, \mathcal{S}, \mathcal{A}, T)\Big) \geq 1-\delta$$

- **Better decisions on more time steps ⇒ better policies for people**
- **Enable applications that can't have many poor decisions to benefit from RL**

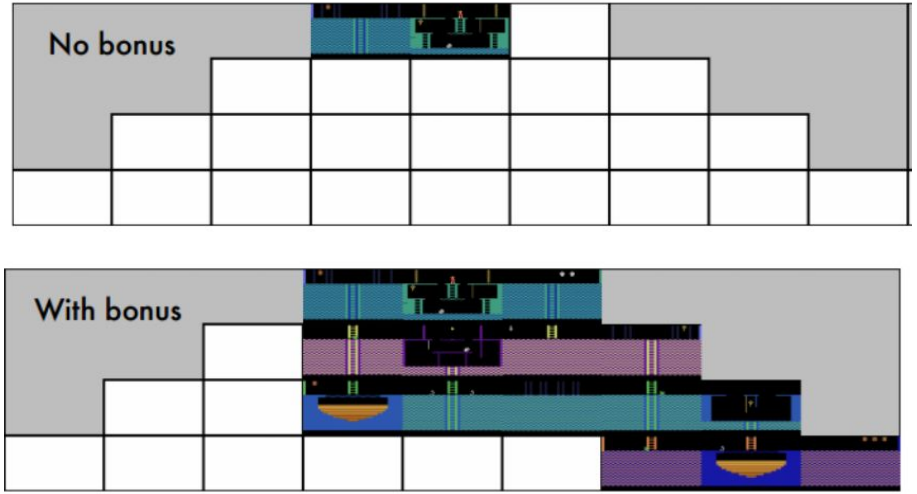
Strategic Exploration Can Be Essential

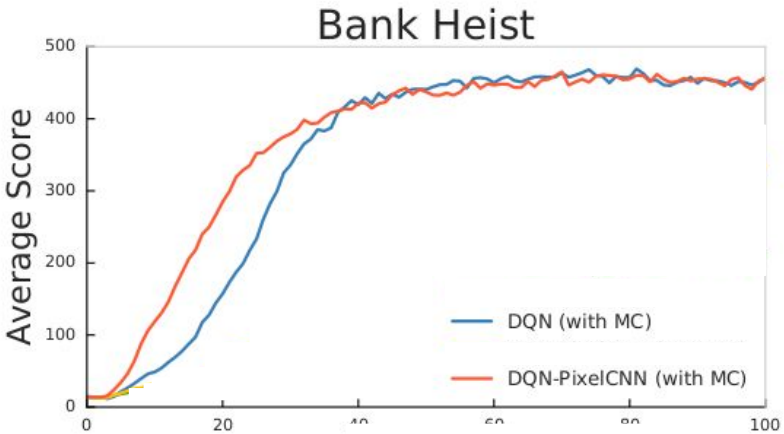




- **After 50 million actions**
- Top: Deep Q Learning (Mnih et al Nature 2015).
- Bottom: Strategic exploration. (Bellemare et al. 2017)

Understanding When it is Hard to Learn to Make Seq. of Decisions Under Uncertainty

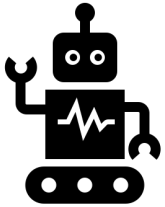


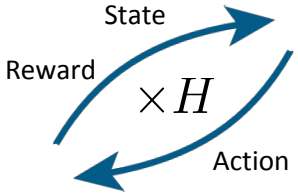


(Bellemare et al. ICML 2017)

- **After 50 million actions**
- Top: Deep Q Learning (Mnih et al Nature 2015).
- Bottom: Strategic exploration. (Bellemare et al. 2017)

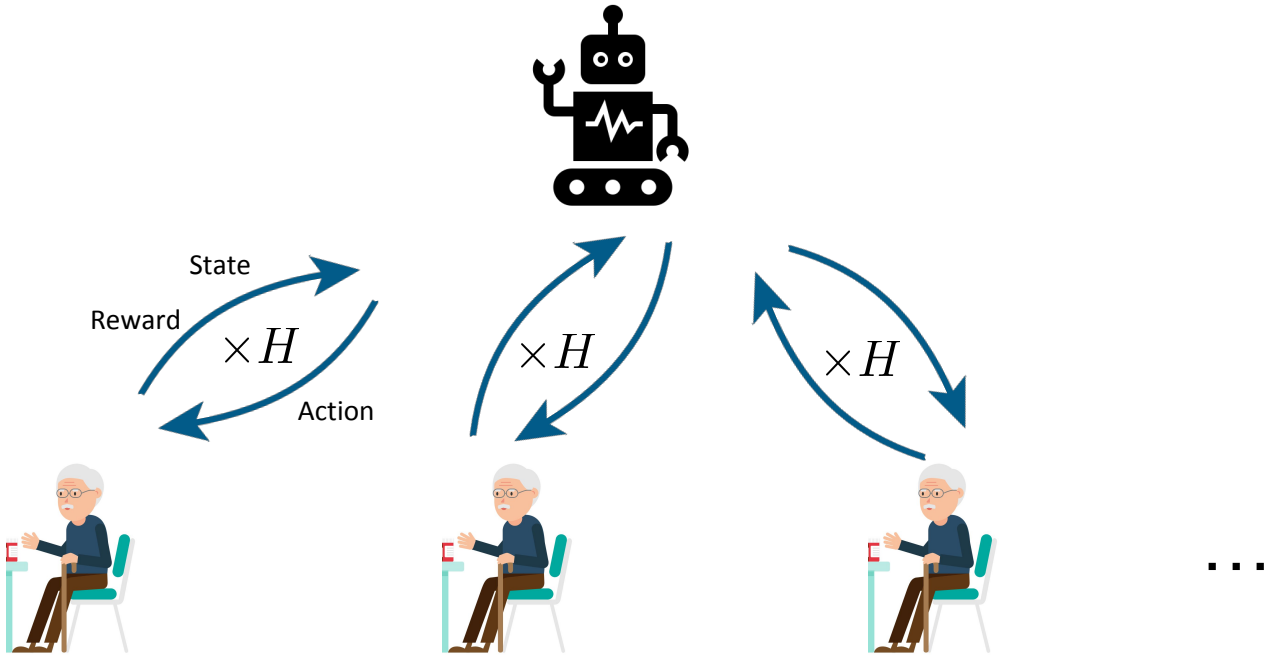
Tabular Episodic RL





Patient 1
Episode 1

Tabular Episodic RL



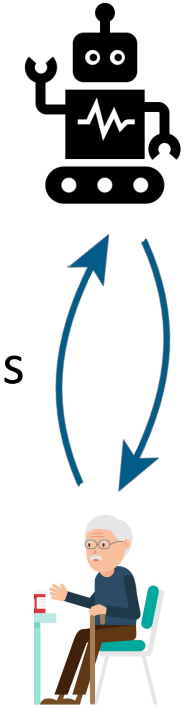
Patient 1
Episode 1

Patient 2
Episode 2

Patient 3
Episode 3

...

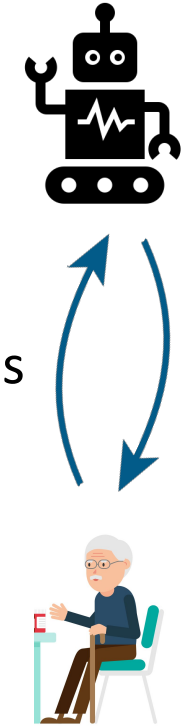
Goal: Policy that Optimizes Expected Reward Over H Steps



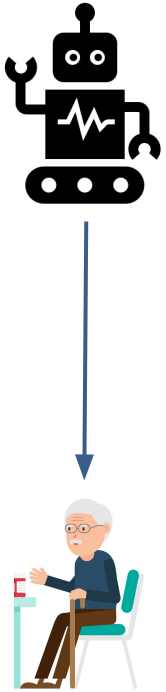
MDP Goal: Policy that Optimizes Expected Reward Over H Steps

Bandit Goal: Policy that Optimizes Expected Reward Over 1 Steps

H interactions

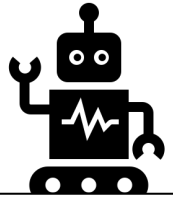


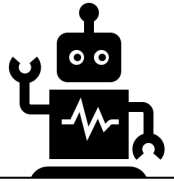
vs



MDP Goal: Policy that Optimizes Expected Reward Over H Steps

Bandit Goal: Policy that Optimizes Expected Reward Over 1 Steps

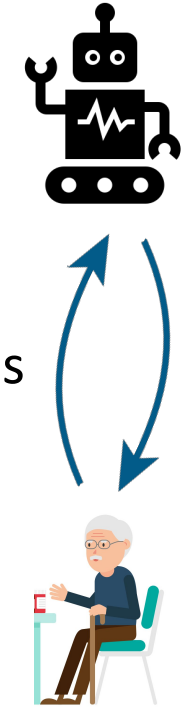




H interact

Is MDP setting provably harder as H increases? (COLT Open Question, Jiang & Agarwal 2018 conjecture “yes”)

Surprizingly, can give a partial answer of no



H interactions

Regret scales with Sqrt (SAK) + terms independent of K/T

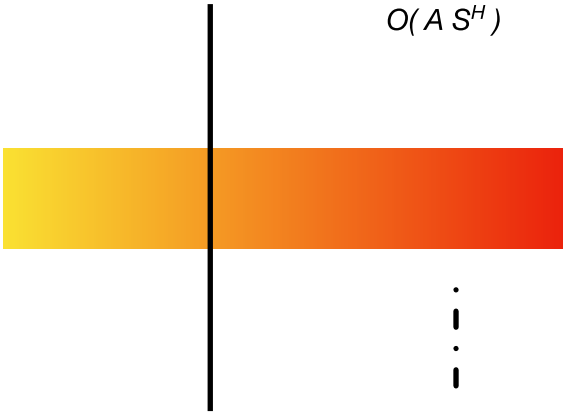
K = # of episodes agent has acted in, each of length H

Techniques to Show This Highlight Insights of Recent Progress in Tabular RL

No Intelligent Exploration

PAC

Regret



S: # states
A: # actions
T: # steps
H: time horizon

Lower Bound | **Efficient Exploration** | **No Intelligent Exploration**

PAC

$$\tilde{O}\left(\left(\frac{SAH^2}{\epsilon^2}+\frac{S^2AH^3}{\epsilon}\right)\ln\frac{1}{\delta}\right).$$

(Dann, Wei, Li, B. 2019)

(Dann & B 2015)

$$\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H^2}{\epsilon^2}\ln\frac{1}{\delta}\right)$$

(Kakade 2003; Strehl & Littman 2005)

$$\tilde{O}\left(\frac{S^2A}{\epsilon^3(1-\gamma)^6}\right)$$

O(A S^H)

Regret

$$\tilde{O}(\sqrt{HSAT})$$

(Azar et al. 2017)

$$\tilde{O}(S\sqrt{HAT})$$

(Dann, Lattimore, B 2017)

$$\tilde{O}(H\sqrt{SAT})$$

(Dann & B 2015)

$$\tilde{O}(HS\sqrt{AT})$$

(UCRL2, Jaksch et al. 2010)

O(T) (greedy or epsilon-greedy)

S: # states
A: # actions
T: # steps
H: time horizon

Problem Dependent Analysis

Lower Bound

$$\tilde{O}\left(\left(\frac{SAH^2}{\epsilon^2}+\frac{S^2AH^3}{\epsilon}\right)\ln\frac{1}{\delta}\right).$$

(Dann, Wei, Li, B. 2019)

Efficient Exploration

(Dann & B 2015)

$$\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|H^2}{\epsilon^2}\ln\frac{1}{\delta}\right)$$

(Kakade 2003; Strehl & Littman 2005)

$$\tilde{O}\left(\frac{S^2A}{\epsilon^3(1-\gamma)^6}\right)$$

No Intelligent Exploration

O(A S^H)

PAC

Regret

$$\tilde{O}(\sqrt{\mathbb{Q}^*SAT})$$

(Zanette & B 2019)

$$\tilde{O}(\sqrt{HSAT})$$

(Azar et al. 2017)

$$\tilde{O}(S\sqrt{HAT})$$

(Dann, Lattimore, B 2017)

$$\tilde{O}(H\sqrt{SAT})$$

(Dann & B 2015)

$$\tilde{O}(HS\sqrt{AT})$$

(UCRL2, Jaksch et al. 2010)

O(T) (greedy or epsilon-greedy)

$\mathbb{Q}^*$: problem dependent constant that does not need to be known

S: # states
A: # actions
T: # steps
H: time horizon

Early Work: Bound Uncertainty Over Dynamics Model Parameters

$$|Q^\star(s,a) - \widehat{Q}^\star(s,a)| = |p(s,a)^\top V^\star - \widehat{p}(s,a)^\top \widehat{V}^\star|$$

(Assuming no reward error)

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

Early Work: Bound Uncertainty Over Dynamics Model Parameters

$$|Q^\star(s,a) - \widehat{Q}^\star(s,a)| = |p(s,a)^\top V^\star - \widehat{p}(s,a)^\top \widehat{V}^\star| \lessapprox \frac{H}{\sqrt{n}}$$

(Hoeffding Inequality)

(Assuming no reward error)

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

Better: Bound Uncertainty *Over Expected Value*

$$|Q^\star(s,a) - \widehat{Q}^\star(s,a)| = |p(s,a)^\top V^\star - \widehat{p}(s,a)^\top \widehat{V}^\star| \lessapprox \frac{H}{\sqrt{n}}$$

(Assuming no reward error)

(Hoeffding Inequality)

$$\lessapprox \frac{\sigma_{s,a}^{V^\star}}{\sqrt{n}} + \frac{H}{n}$$

(Bernstein Inequality)

$$\sigma_{s,a}^{V^\star} = Var_{s^+ \sim p(s,a)} V^\star$$

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

Better: Bound Uncertainty Over Expected Value

And Use to Create New Optimism Bonuses Used for Decision Making

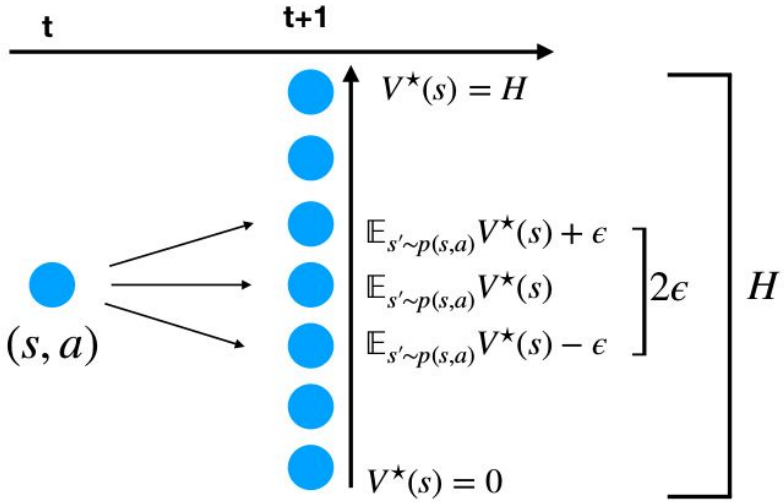
$$|Q^\star(s,a) - \widehat{Q}^\star(s,a)| = |p(s,a)^\top V^\star - \widehat{p}(s,a)^\top \widehat{V}^\star| \lessapprox \frac{H}{\sqrt{n}}$$

(Hoeffding Inequality)

(Assuming no reward error)

$$\lessapprox \frac{\sigma_{s,a}^{V^\star}}{\sqrt{n}} + \frac{H}{n}$$

(Bernstein Inequality)



$$\sigma_{s,a}^{V^\star} = Var_{s^+\sim p(s,a)} V^\star$$

$$\mathbb{Q}^* = \max_{s,a} Var_{s'\sim p(s'|s,a)} V^*(s')$$

"Environmental norm" Maillard et al NeurIPS 2014

for $k = 1, 2, \ldots$ **do**

  for $t = H, H-1, \ldots, 1$ **do**

    for $s \in \mathcal{S}$ **do**

      for $a \in \mathcal{A}$ **do**

        $\hat{p} = \frac{p_{sum}(\cdot,s,a)}{n_k(s,a)}$

        $b_k^{pv} = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}} \left(\frac{4J+B_p}{\sqrt{n_k(s,a)}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$

        $Q(a) = \min\{H - t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s,a) + \hat{p}^\top \overline{V}_{t+1} + b_k^{pv}\}$

      end for

for $k = 1, 2, \ldots$ **do**

  for $t = H, H-1, \ldots, 1$ **do**

    for $s \in \mathcal{S}$ **do**

      for $a \in \mathcal{A}$ **do**

        $\hat{p} = \frac{p_{sum}(\cdot,s,a)}{n_k(s,a)}$

        $b_k^{pv} = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}} \left(\frac{4J+B_p}{\sqrt{n_k(s,a)}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$

        $Q(a) = \min\{H - t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s,a) + \hat{p}^\top \overline{V}_{t+1} + b_k^{pv}\}$

      end for

Bonus term depends on upper and lower bounds on value function on next time step

Empirical estimates of dynamics and reward

for $k = 1, 2, \ldots$ **do**

  for $t = H, H-1, \ldots, 1$ **do**

    for $s \in \mathcal{S}$ **do**

      for $a \in \mathcal{A}$ **do**

        $\hat{p} = \frac{p_{sum}(\cdot,s,a)}{n_k(s,a)}$

        $b_k^{pv} = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}} \left(\frac{4J+B_p}{\sqrt{n_k(s,a)}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$

        $Q(a) = \min\{H - t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s,a) + \hat{p}^\top \overline{V}_{t+1} + b_k^{pv}\}$

      end for

Bonus term depends on upper and lower bounds on value function on next time step

Empirical estimates of dynamics and reward

Do Bellman backup w/ bonuses

for $k = 1, 2, \ldots$ **do**

  for $t = H, H-1, \ldots, 1$ **do**

    for $s \in \mathcal{S}$ **do**

      for $a \in \mathcal{A}$ **do**

        $\hat{p} = \frac{p_{sum}(\cdot,s,a)}{n_k(s,a)}$

        $b_k^{pv} = \phi(\hat{p}(s,a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s,a)}} \left(\frac{4J+B_p}{\sqrt{n_k(s,a)}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$

        $Q(a) = \min\{H - t, \hat{r}_k(s, \tilde{\pi}_k(s,t)) + b_k^r(s,a) + \hat{p}^\top \overline{V}_{t+1} + b_k^{pv}\}$

      end for

      $\tilde{\pi}_k(s,t) = \arg\max_a Q(a)$

      $\overline{V}_t(s) = Q(\tilde{\pi}_k(s,t))$

      $b_k^{pv} = \phi(\hat{p}(s, \tilde{\pi}_k(s,t)), \underline{V}_{t+1}) + \frac{1}{\sqrt{n(s,\tilde{\pi}_k(s,t))}} \left(\frac{4J+B_p}{\sqrt{n_k(s,\tilde{\pi}_k(s,t))}} + B_v \|\overline{V}_{t+1} - \underline{V}_{t+1}\|_{2,\hat{p}} \right)$

      $\underline{V}_t(s) = \max\{0, \hat{r}_k(s, \tilde{\pi}_k(s,t)) - b_k^r(s, \tilde{\pi}_k(s,t)) + \hat{p}^\top \underline{V}_{t+1} - b_k^{pv}\}$

    end for

  end for

  Evaluate policy $\tilde{\pi}_k$ and update MLE estimates $\hat{p}(\cdot,\cdot)$ and $\hat{r}(\cdot,\cdot)$

end for

Bonus term depends on upper and lower bounds on value function on next time step

Do Bellman backup w/ bonuses

Empirical estimates of dynamics and reward

Main Result

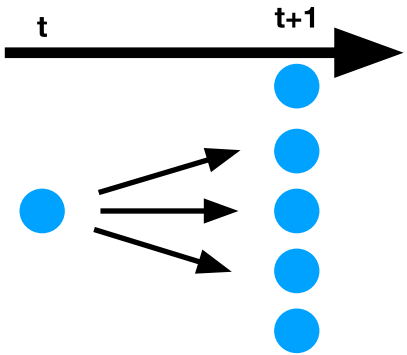
An algorithm with a (high probability) regret bound:

$$\tilde{O}\left(\min\left[\sqrt{\mathbb{Q}^* SAT}, \sqrt{\frac{G^2}{H} SAT}\right] + S^{1.5} A H^2 (H + \sqrt{S})\right)$$

Problem dependent constants G & $\mathbb{Q}^*$

Algorithm is <u>not</u> given G or $\mathbb{Q}^*$

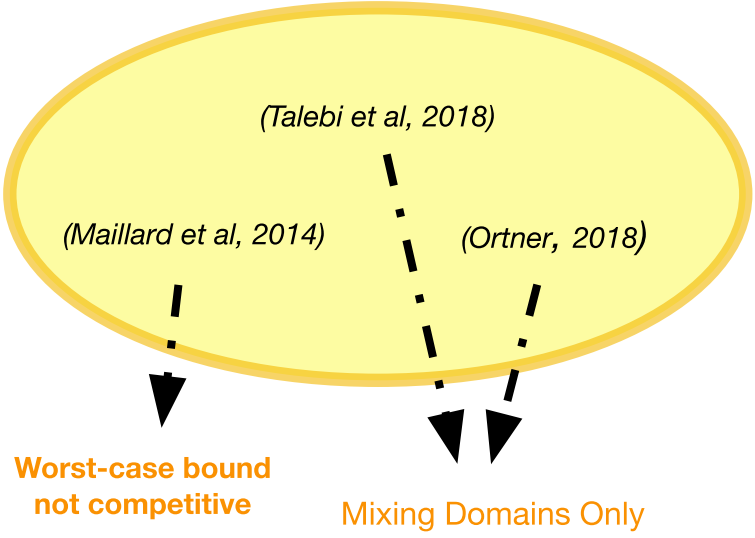
Minimax Optimality in dominant terms



$$\mathbb{Q}^* = \max_{s,a} Var_{s' \sim p(s'|s,a)} V^*(s')$$

$$G = \max_{s_0, a_0, \ldots, a_{H-1}, s_h} \sum_{i=1}^{H} r_i$$

Vs Prior Instance Dependent Bounds



[REGAL] *(Bartlett et al, 2010)*
[SCAL] *(Fruit et al, 2018)*

$$\tilde{O}(HS\sqrt{AT})$$

Intuition: H upper bounds the value function (it scales the problem)

Idea: if the span is know pass this info to the algorithm

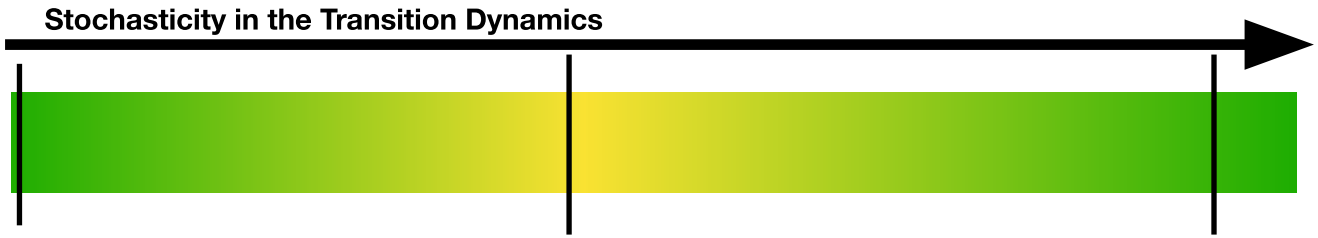
$$\tilde{O}(span(V^*)S\sqrt{AT})$$

Need to know domain properties

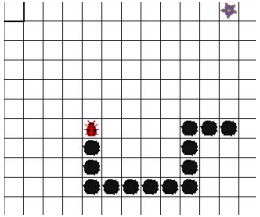
Our result: matches minimax worst case, doesn't require mixing (they are in harder infinite H setting)

Our result: doesn't require us to know span, tighter S dependence

Enhancing Understanding of When it Is Hard to Learn to Act Well

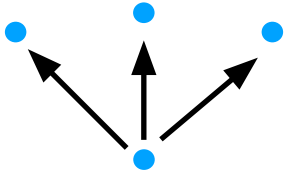


Deterministic MDP



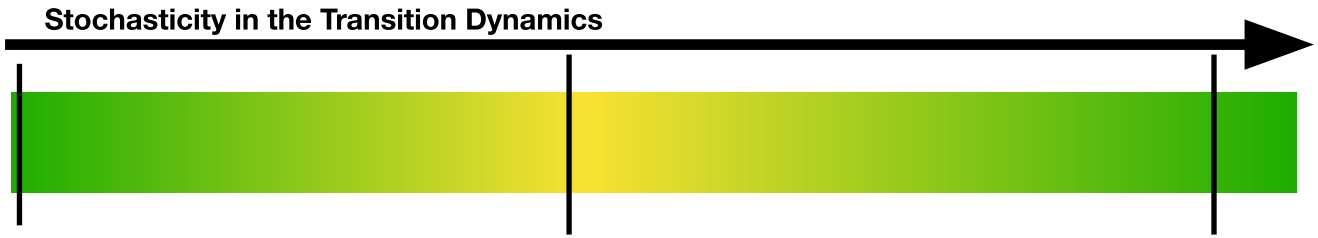
$$\tilde{O}(SAH^2)$$

Bandit Like Structure

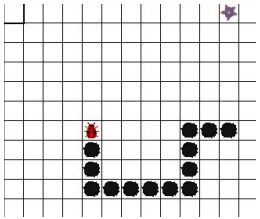


$$\tilde{O}(\sqrt{SAT} + [\ldots])$$

Enhancing Understanding of When it Is Hard to Learn to Act Well

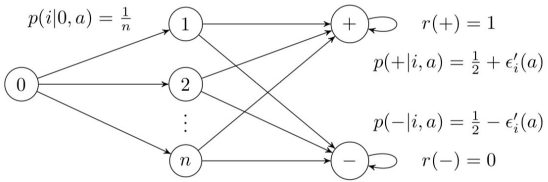


Deterministic MDP



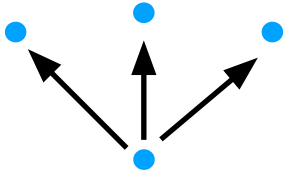
$$\tilde{O}(SAH^2)$$

Hard Instances Inducing the Lower Bound



$$\tilde{O}(\sqrt{HSAT} + [\ldots])$$

Bandit Like Structure



$$\tilde{O}(\sqrt{SAT} + [\ldots])$$

Answers part of COLT 2018 open question (Agarwal and Jiang):
No horizon dependence in regret bound for their setting

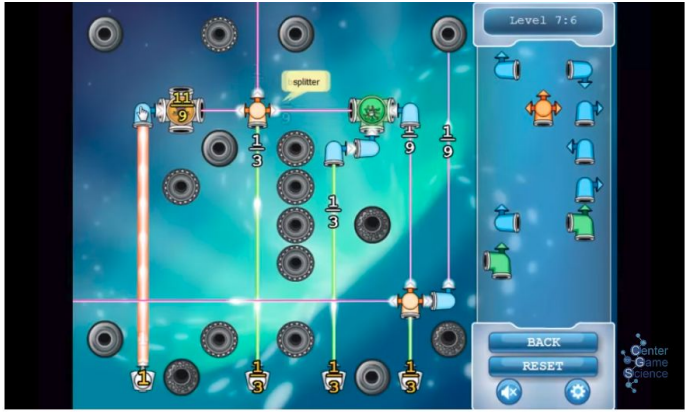
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

1. Exploration: how to quickly gather information to learn to make good decisions
    a. Minimax bounds for worst case MDP tabular for regret & PAC
    b. Instance dependent bounds for tabular MDP that suggest exploration is easier than expected in many cases

Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

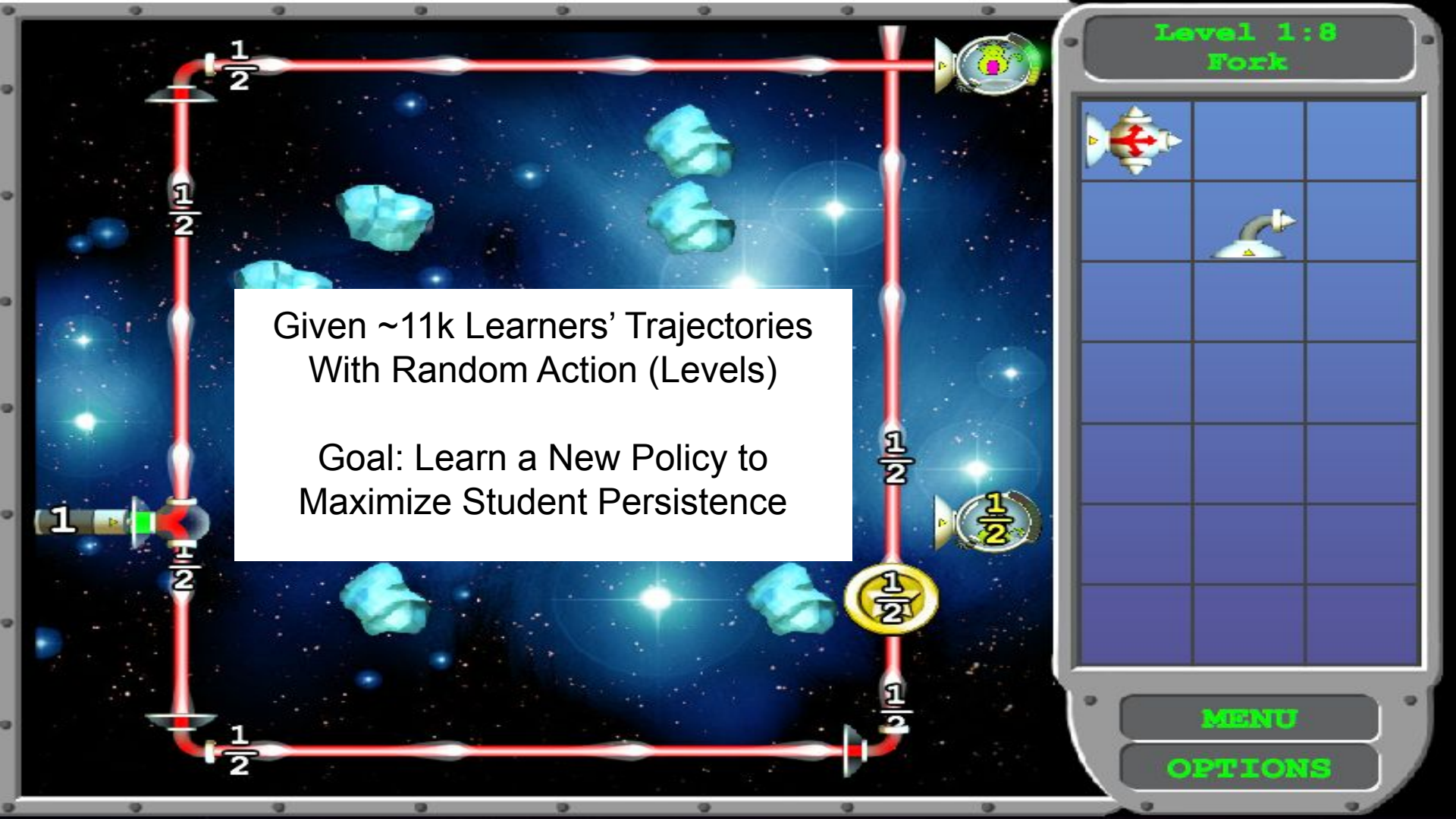
1. Exploration: how to quickly gather information to learn to make good decisions
2. **Counterfactual / batch RL: reasoning to best leverage existing data**


Level 1:8
Fork
MENU
OPTIONS



Took > 30s

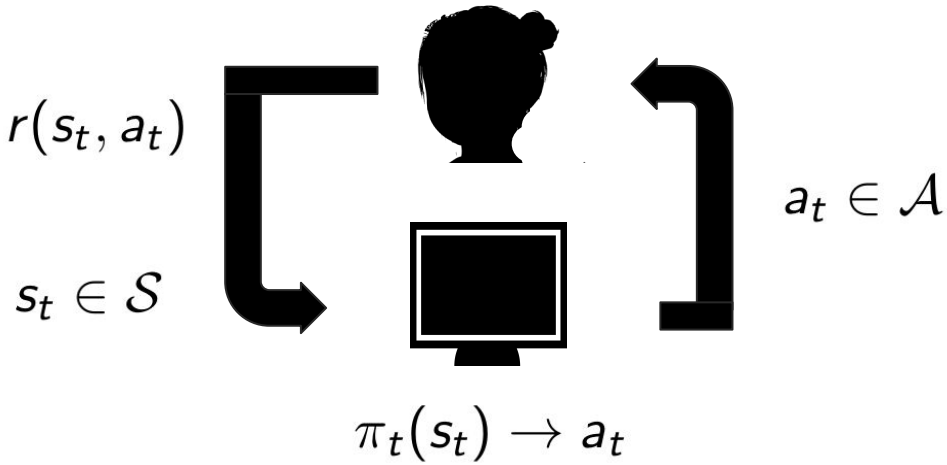
Took <= 30s



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

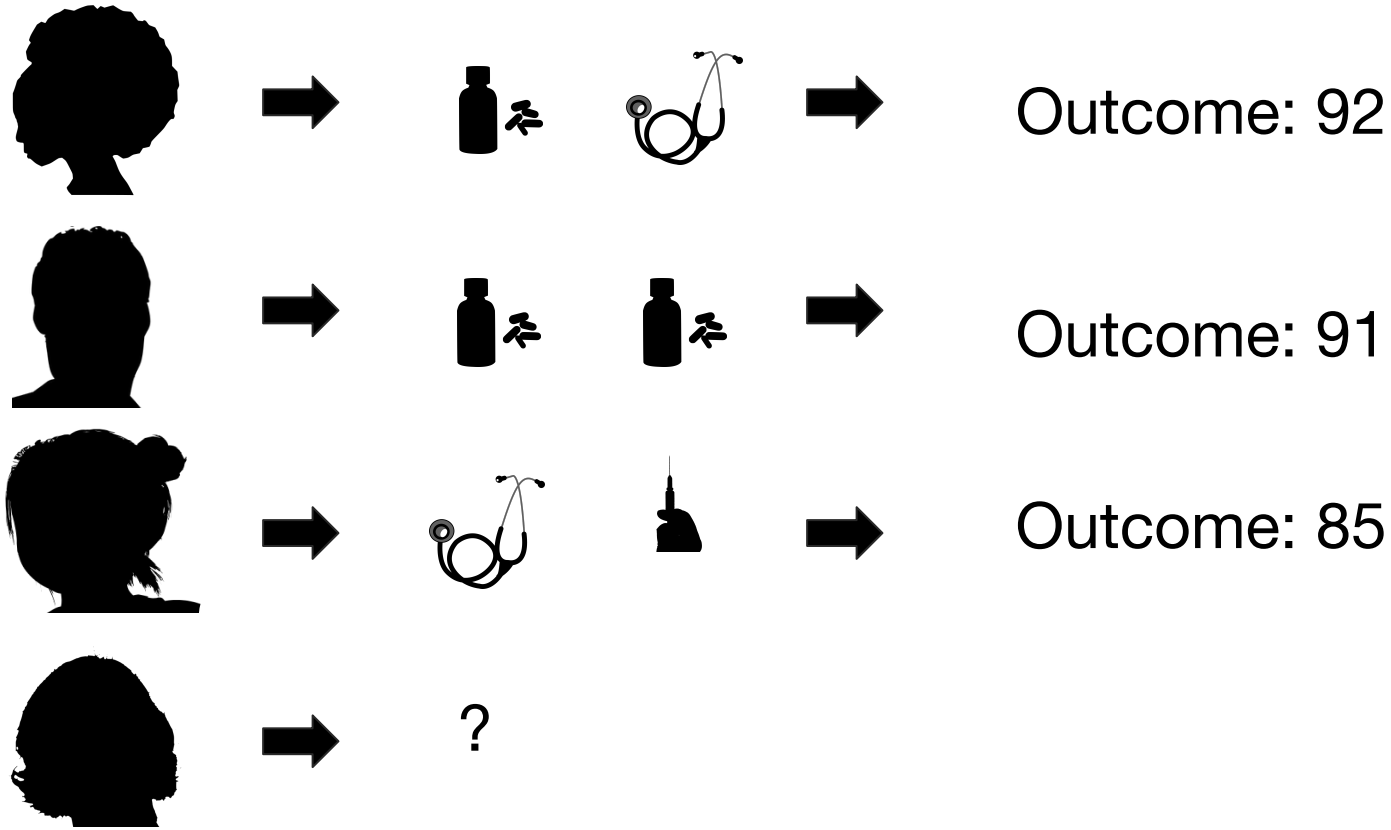
Goal: Learn a New Policy to
Maximize Student Persistence

Counterfactual / Batch Off Policy Reinforcement Learning

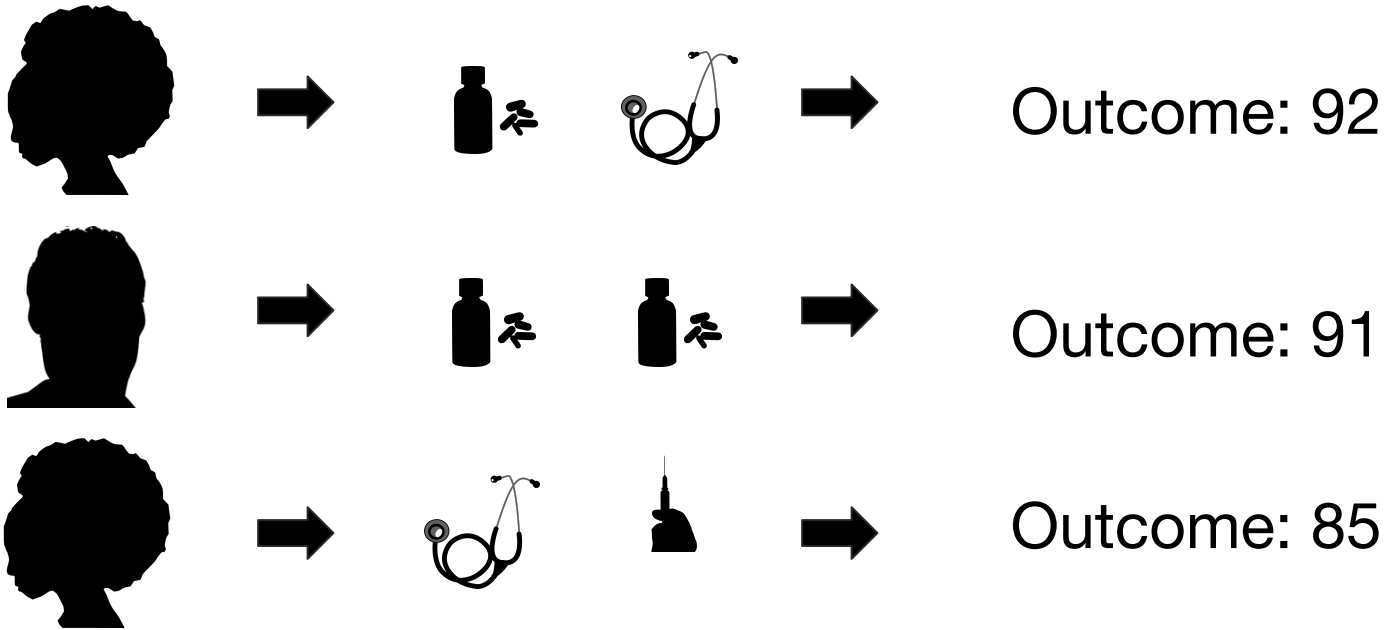


$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$

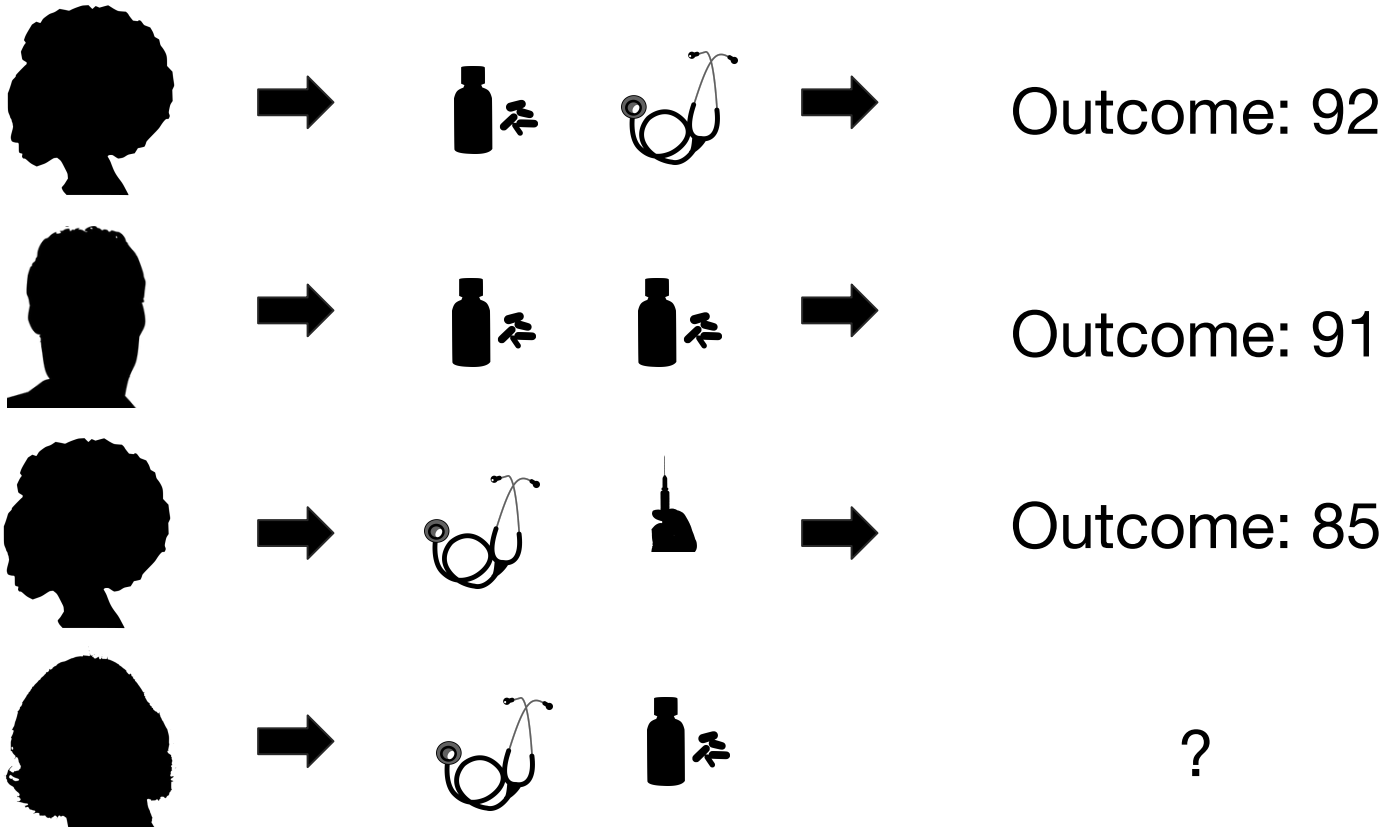
“What If?” Reasoning Given Past Data



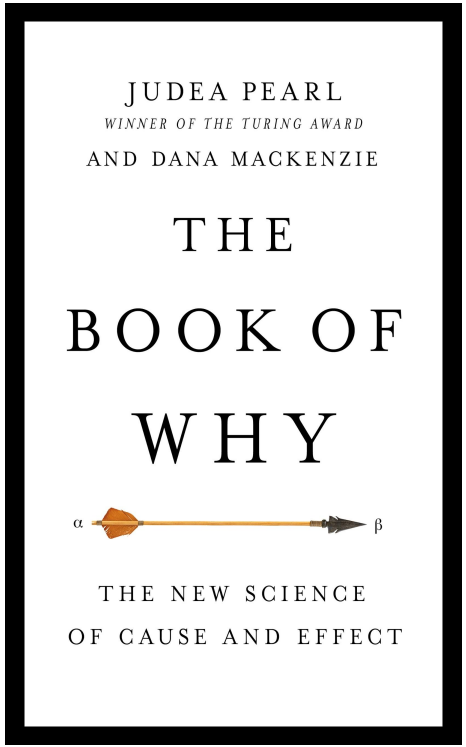
Data Is Censored



Need for Generalization



Growing Interest in Causal Inference & ML



Journal of Educational Psychology
1974, Vol. 66, No. 5, 688-701

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES[1]

DONALD B. RUBIN[2]

Educational Testing Service, Princeton, New Jersey

A discussion of matching, randomization, random sampling, and other methods of controlling extraneous variation is presented. The objective is to specify the benefits of randomization in estimating causal effects of treatments. The basic conclusion is that randomization should be employed whenever possible but that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.

Recent psychological and educational literature has included extensive criticism of the use of nonrandomized studies to estimate causal effects of treatments (e.g., Campbell & Erlebacher, 1970). The implication in much of this literature is that only properly randomized experiments can lead to useful estimates of causal effects. If taken as applying to all fields of study, this position is untenable. Since the extensive use of randomized experiments is limited to the last half century,[3] and in fact is *not* used in much scientific investigation today,[4] one is led to the conclusion that most scientific "truths" have been established without using randomized experiments. In addition, most of us successfully determine the causal effects of many of our everyday actions, even interpersonal behaviors, without the benefit of randomization.

Even if the position that causal effects of treatments can only be well established from randomized experiments is taken as applying only to the social sciences in which there are currently few well-established causal relationships, its implication—to ignore existing observational data—may be counter-productive. Often the only immediately available data are observational (nonrandomized) and either (*a*) the cost of performing the equivalent randomized experiment to test all treatments is prohibitive (e.g., 100 reading programs under study); (*b*) there are ethical reasons why the treatments cannot be randomly assigned (e.g., estimating the effects of heroin addiction on intellectual functioning); or (*c*) estimates based on results of experiments would be delayed many years (e.g., effect of childhood intake of cholesterol on longevity). In cases such as these, it seems more reasonable to try to estimate the effects of the treatments from nonrandomized studies than to ignore these data and dream of the ideal experiment or make "armchair" decisions without the benefit of data analysis. Using the indications from nonrandomized studies, one can, if necessary, initiate randomized experiments for those treatments that require better estimates or that look most promising.

The position here is *not* that randomization is overused. On the contrary, given a choice between the data from a randomized experiment and an equivalent nonrandomized study, one should choose the data from the experiment, especially in the social sciences where much of the variability is often unassigned to particular causes. However, we will develop the position that nonrandomized studies as well as randomized

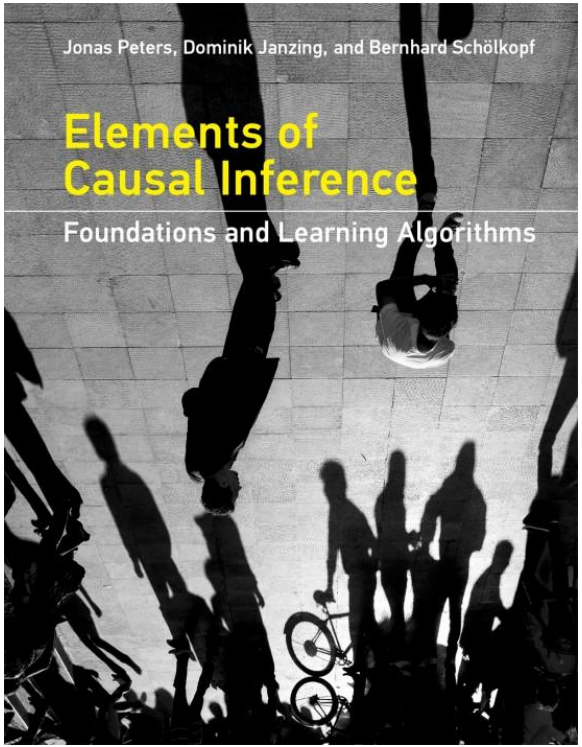
[1] I would like to thank E. J. Anastasio, A. E. Beaton, W. G. Cochran, K. M. Kazarow, and R. L. Linn for helpful comments on earlier versions of this paper. I would also like to thank the U. S. Office of Education for supporting work on this paper under contract OEC-0-71-3715.

[2] Requests for reprints should be sent to Donald B. Rubin, Division of Data Analysis Research, Educational Testing Service, Princeton, New Jersey 08540.

[3] Essentially since Fisher (1925).

[4] For example, in Davies (1954), a well-known textbook on experimental design in industrial work, randomization is not emphasized.

688



Batch / Counterfactual Policy Optimization:
Pick Policy w/Best Estimated Expected Sum of Rewards

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

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

Challenge: Using Unbiased Value Estimators Can Fail

Doroudi, Thomas, B UAI 2017 best paper

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

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

Quest: Batch Policy Optimization w/ Generalization Bounds

$$\arg\max_{\pi \in \mathcal{H}_i} \quad \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \quad \int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds \quad - \quad \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}}$$

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

Quest: Batch Policy Optimization w/ Generalization Bounds

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

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

Quest: Batch Policy Optimization w/ Generalization Bounds

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

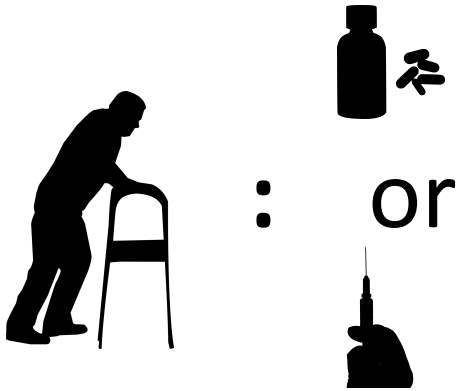
$$\mathcal{H} = \mathcal{M}, \mathcal{V}, \Pi \ ?$$

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

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

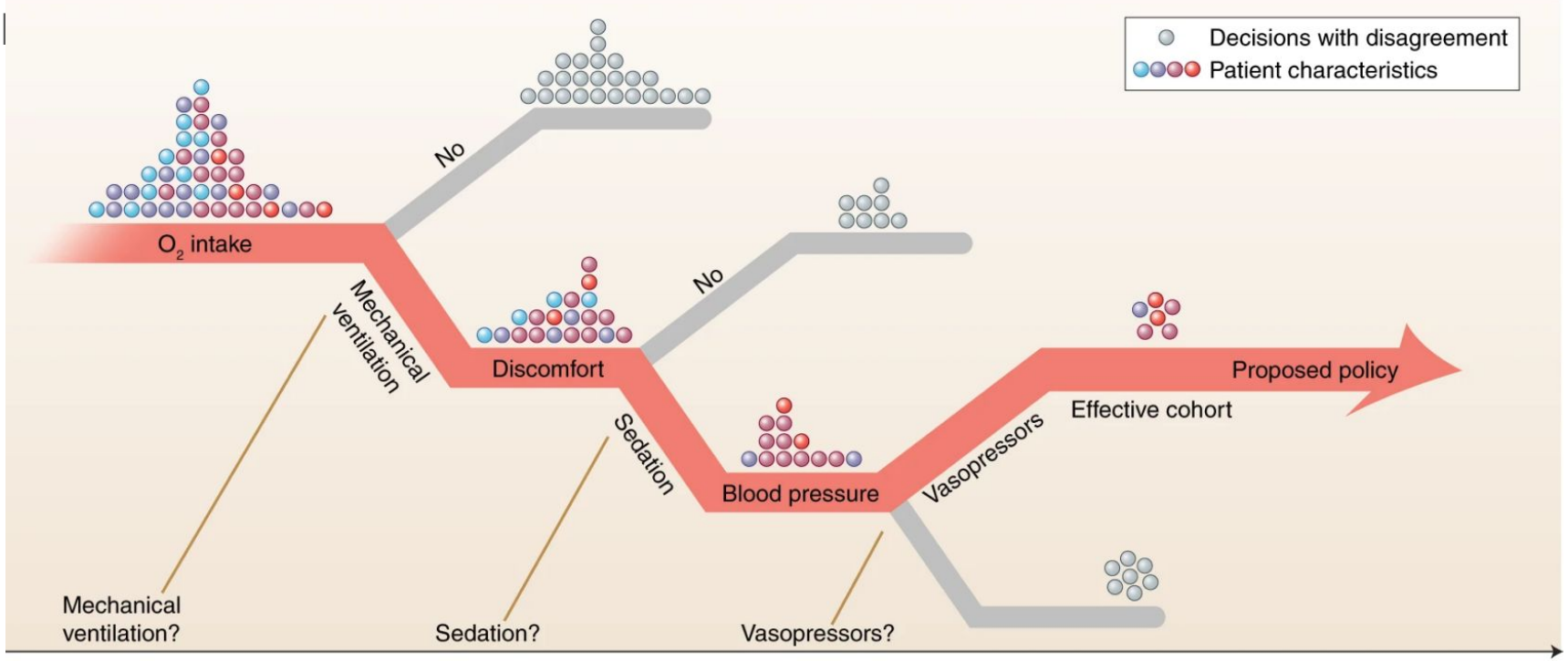
$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \to a$
S_0: Set of initial states
$\hat{V}^{\pi}(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$

Substantial Literature Focuses on 1 Binary Decision: Treatment Effect Estimation from Old Data



Challenge: Covariate Shift

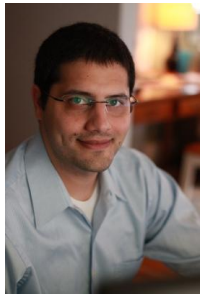
Different Policies → Different Actions → Different State Distributions



Gottesman et a. Guidelines for reinforcement learning in healthcare. Nature Medicine 2019. Figure by Debbie Maizels/Springer Nature

More Data Efficient Policy Evaluation with Biased Estimators (Thomas and B, ICML 2016)

$$\underbrace{\arg\max_{\pi \in H_i} \quad \max_{H_i \in \{H_1, H_2, \ldots\}}}_{\text{Policy Optimization}} \quad \underbrace{\int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds}_{\text{Policy Evaluation}} - \underbrace{\sqrt{\frac{f(VC(h), \ldots)}{n}}}_{\text{Error Bound}}$$



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

Counterfactual Reasoning for Policy Evaluation*

Parametric Models
of dynamics, rewards
or values fit to data

+ Low variance
- Bias (unless realizable)

Counterfactual Reasoning for 1-Step Decision Making

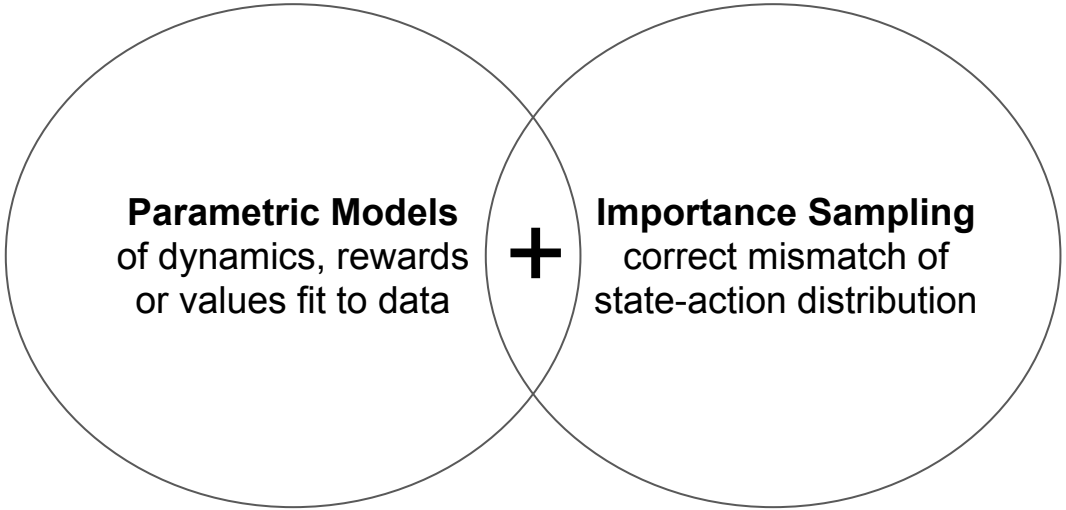
Parametric Models
of dynamics, rewards
or values fit to data

Importance Sampling
correct mismatch of
state-action distribution

+ Low variance
- Bias (unless realizable)

+ Unbiased under certain assumptions
- High variance

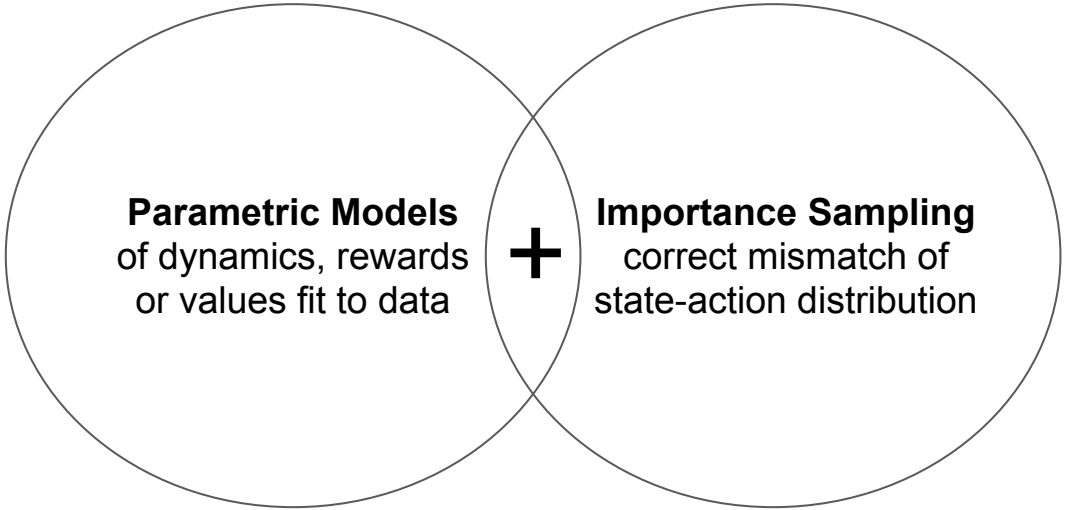
Doubly Robust for 1 Step Decision Making



+ Smaller variance than importance sampling
+ Unbiased if either model realizable or behavior policy known

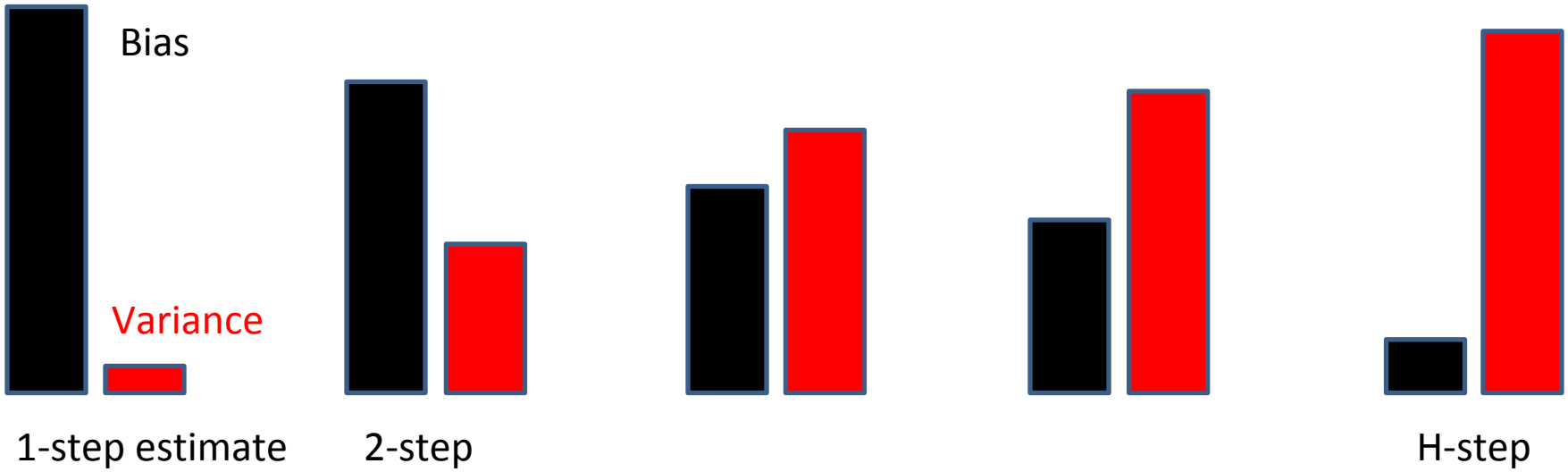
- Brought to multi-armed bandits (Dudik et al. 2011); and RL (Jiang & Li 2016)

Approach: Bias to Save Variance



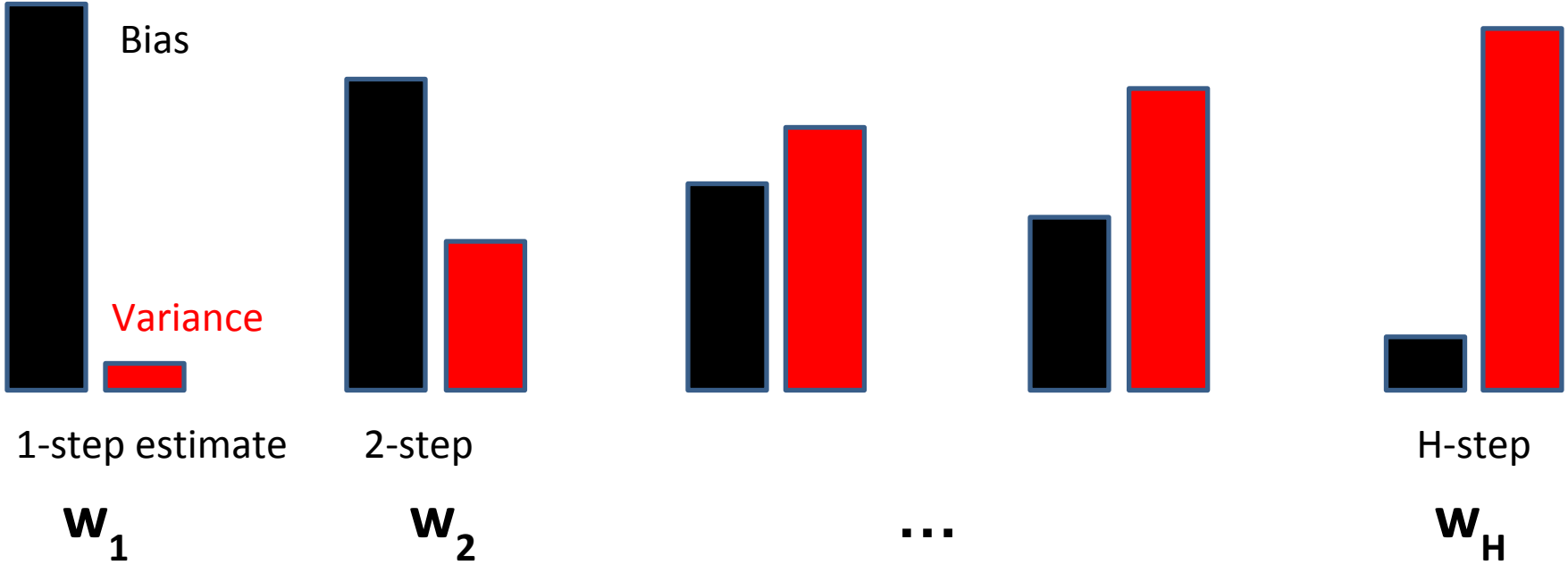
Thomas and B 2016

Blend IS-Based & Model Based Estimators to Directly Min Mean Squared Error

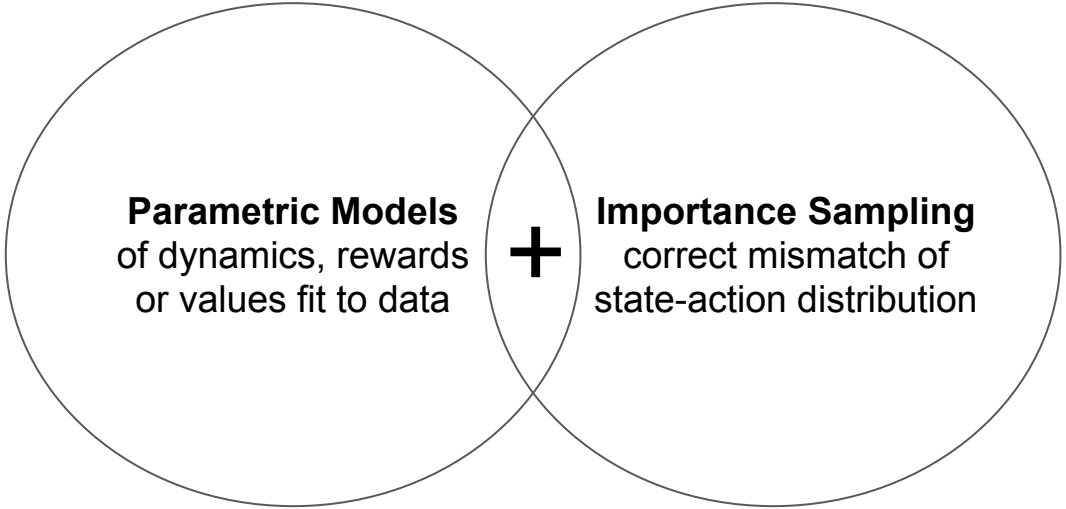


Blend IS-Based & Model Based Estimators to Directly Min Mean Squared Error

Solve quadratic program using **estimated bias and variance**



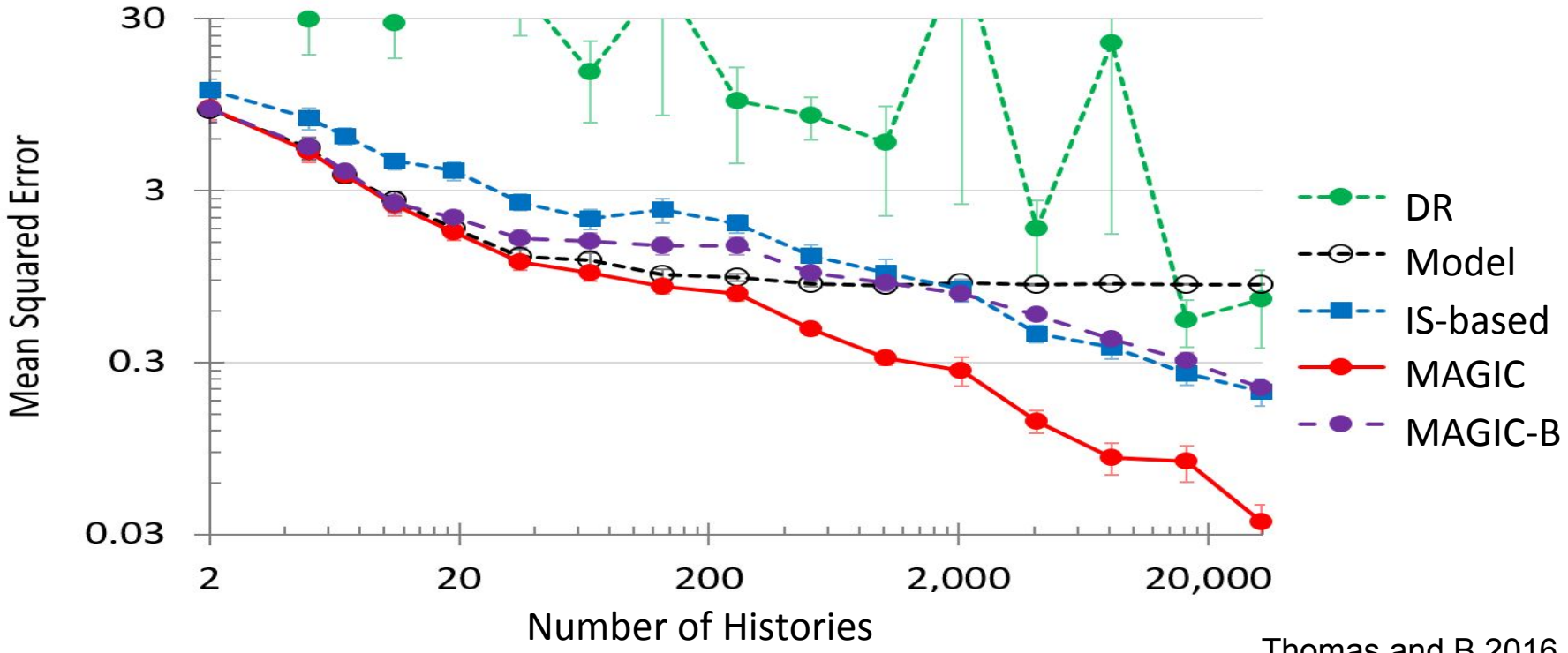
Weighted Doubly Robust and Model & Guided Importance Sampling Estimators



- **Under some standard assumptions, proved both estimators consistent**

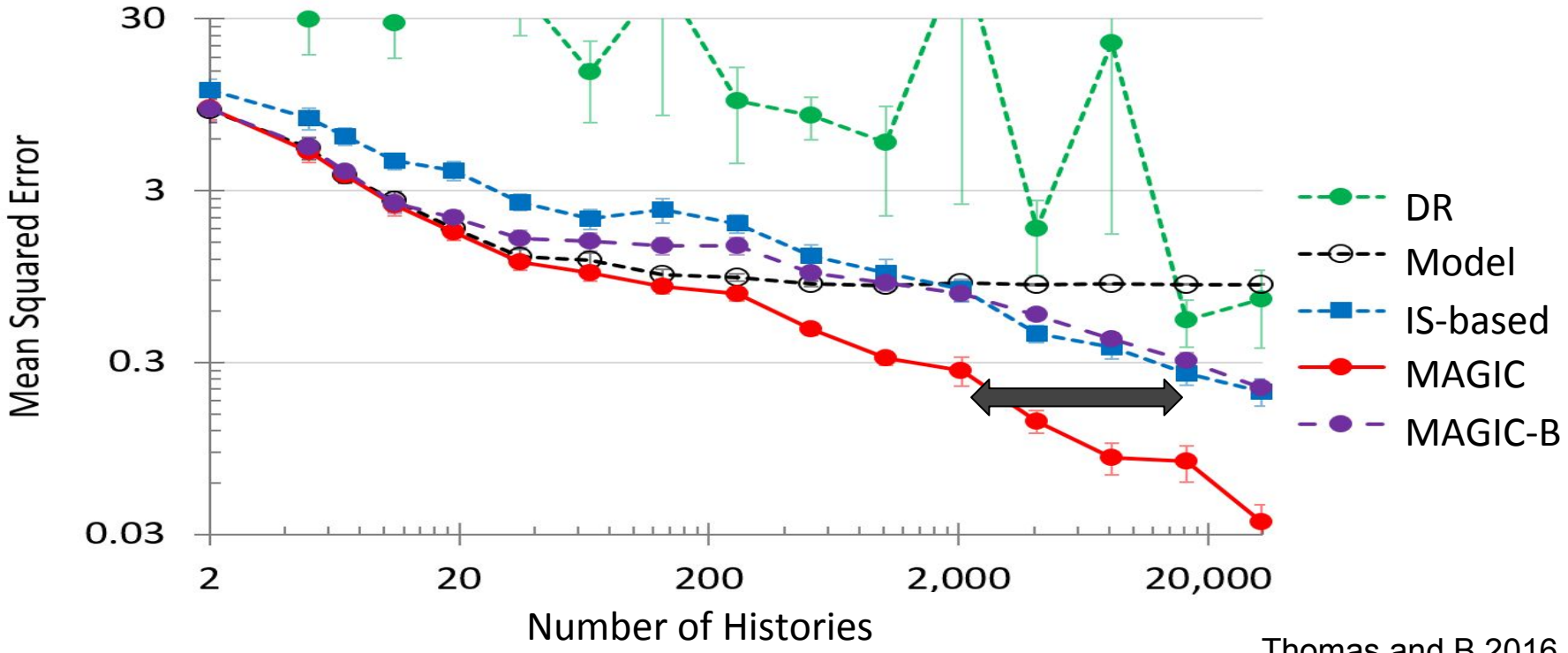
Thomas and B 2016

Gridworld Simulation



Thomas and B 2016

Gridworld Simulation: Needed **Only 10% of** the Data to Learn a Good Estimate of New Policy's Value



Thomas and B 2016

More Data Efficient Policy Evaluation with Biased Estimators (Thomas and B, ICML 2016)

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

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

Identifying a Good Policy for Future Use

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

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

Challenge: Good Error Bound Analysis

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

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

Challenge: Good Error Bound Analysis

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

- Importance sampling bounds (e.g. Thomas et al, 2015) ignore hypothesis class structure & are typically require very large n

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

Challenge: Good Error Bound Analysis

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

- Importance sampling bounds (e.g. Thomas et al, 2015) ignore hypothesis class structure & are typically require very large n
- Kernel function & averager approaches (e.g. Ormoneit & Sten 2002) can need # samples exponential in input state dimension

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

Challenge: Good Error Bound Analysis

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

- Importance sampling bounds (e.g. Thomas et al, 2015) ignore hypothesis class structure & are typically require very large n
- Kernel function & averager approaches (e.g. Ormoneit & Sten 2002) can need # samples exponential in input state dimension
- FQI bounds (e.g. Munos 2003; Munos & Szepesvári 2008; Antos et al., 2008; Lazaric et al., 2012; Farahmand et al., 2009; Maillard et al., 2010; Le, Voloshin, Yue 2019; Chen & Jiang 2019)
  - Require stronger assumptions (realizability and bounds on the inherent Bellman error)
  - If not realizable, these bounds depend on unknown quantities

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

Challenge: Good Error Bound Analysis

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

- Importance sampling bounds (e.g. Thomas et al, 2015)
- Kernel function (e.g. Ormoneit & Sten 2002)
- FQI bounds (e.g. Munos 2003; Munos & Szepesvári 2008; Antos et al., 2008; Lazaric et al., 2012; Farahmand et al., 2009; Maillard et al., 2010; Le, Voloshin, Yue 2019; Chen & Jiang 2019)
  - Require stronger assumptions (realizability and bounds on the inherent Bellman error)
  - If not realizable, FQI bounds depend on unknown quantities
- Primal dual approaches (e.g. Dai, Shaw, Li, Xiao, He, Liu, Chen, Song 2018) are promising and have similar dependencies

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

Aim: Strong Generalization Guarantees on Policy Performance, Alternative: Guarantee Find Best in Class Policy

$$\arg\max_{\pi \in \Pi} \int_{s \in S_0} V^{\pi}(s) ds$$

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states

Aim: Strong Generalization Guarantees on Policy Performance, Alternative: Guarantee Find Best in Class Policy

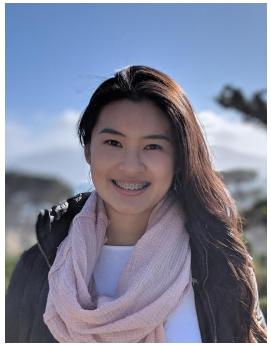
$$\underbrace{\max_{\pi \in \Pi} \int_{s \in S_0} V^{\pi}(s) ds}_{\text{Best policy } \in \Pi} - \underbrace{\int_{s \in S_0} V^{\hat{\pi} = f(\mathcal{D})}(s) ds}_{\text{Policy selected } \in \Pi} \leq \sqrt{\frac{f(\kappa(\Pi), \ldots)}{n}}$$

$$\kappa(\Pi) = \int_0^1 \sqrt{\log N_{dh}(\epsilon^2, \Pi)} d\epsilon$$

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \to a$
S_0: Set of initial states

1st Guarantees on Performance of Policy Choose Vs Best in Class for When to Treat Policies (w/Xinkun Nie & Stefan Wager, arxiv)

$$\underbrace{\max_{\pi \in \Pi} \int_{s \in S_0} V^{\pi}(s) ds}_{\text{Best policy } \in \Pi} - \underbrace{\int_{s \in S_0} V^{\hat{\pi}=f(\mathcal{D})}(s) ds}_{\text{Policy selected } \in \Pi} \leq \; o\left(\left(c_1 \kappa(\Pi) + c_2 + \sqrt{2 \log\left(\frac{1}{\delta}\right)}\right) \sqrt{\frac{V_{\max}}{n}}\right)$$

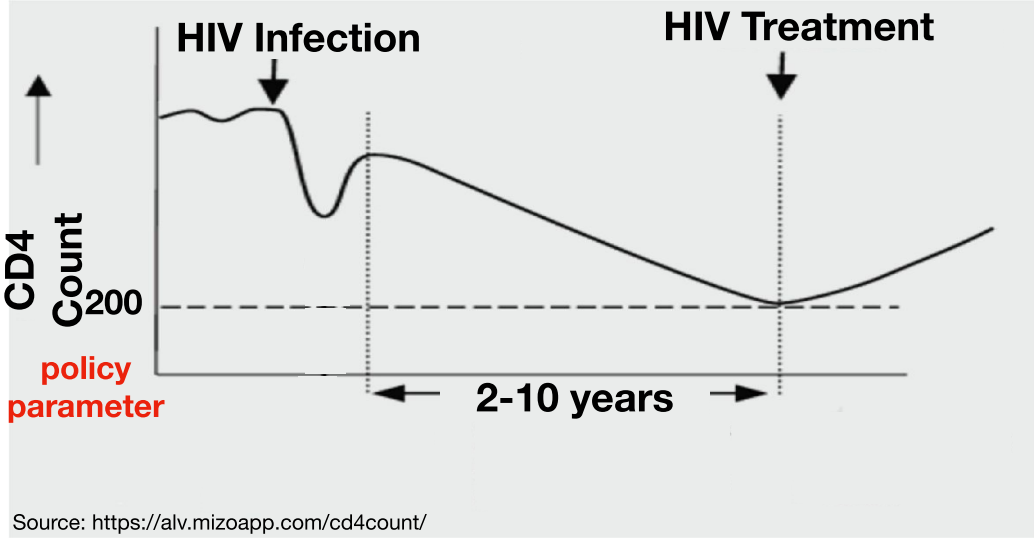


$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states

$$\kappa(\Pi) = \int_0^1 \sqrt{\log N_{dh}(\epsilon^2, \Pi)} d\epsilon$$

Example: Linear Thresholding Policies

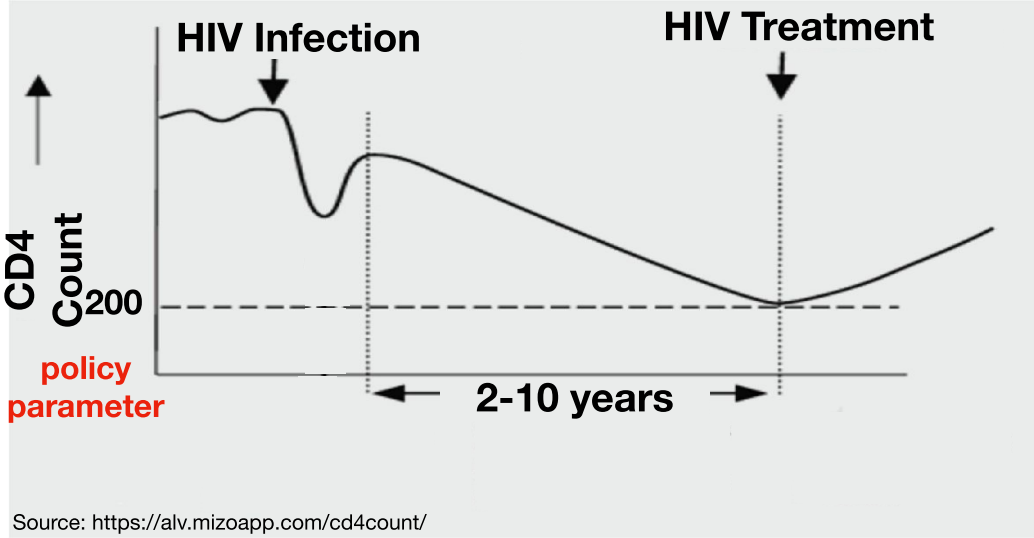
Starting HIV treatment as soon as CD4 count dips below 200



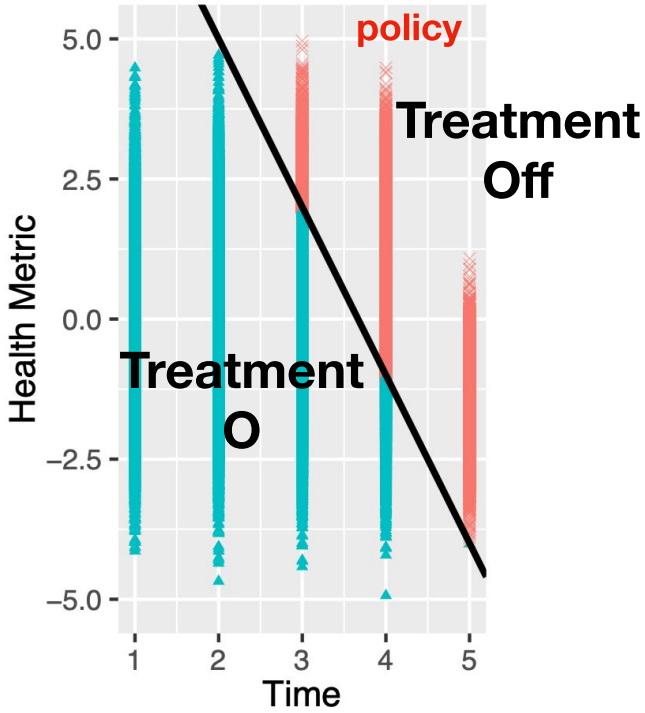
Source: https://alv.mizoapp.com/cd4count/

Example: Linear Thresholding Policies

Starting HIV treatment as soon as CD4 count dips below 200



Stopping treatment as soon as health metric above line



Selecting a When to Treat Policy

$$\underbrace{\max_{\pi \in \Pi} \int_{s \in S_0} V^{\pi}(s) ds}_{\text{Best policy } \in \Pi} - \underbrace{\int_{s \in S_0} V^{\boxed{\hat{\pi} = f(\mathcal{D})}}(s) ds}_{\text{Policy selected } \in \Pi} \leq \; o\left(\left(c_1 \kappa(\Pi) + c_2 + \sqrt{2 \log\left(\frac{1}{\delta}\right)}\right) \sqrt{\frac{V_{\max}}{n}}\right)$$

$\mathcal{D}$: Dataset of n traj.s τ, $\tau \sim \pi_b$
π: Policy mapping $s \rightarrow a$
S_0: Set of initial states

$$\kappa(\Pi) = \int_0^1 \sqrt{\log N_{dh}(\epsilon^2, \Pi)} d\epsilon$$

Use an Advantage Decomposition

$$\Delta_\pi := \underbrace{V_\pi - V_0}_{\text{Difference in expected value of following policy } \pi \text{ vs. } 0} = \mathbb{E}_0\left[\sum_{t=1}^{T} \mathbf{1}_{t \geq \tau_\pi} \underbrace{\left(\mu_{now,t}(S_t) - \mu_{next,t}(S_t)\right)}_{\text{“Advantage”: Difference in expected value if we start treating now vs. next time-step}}\right]$$

[Murphy 2005]

[Kakade 2003]

never acting

Use a Doubly Robust Advantage Decomposition

$$\Delta_\pi := \underbrace{V_\pi - V_0}_{\text{Difference in expected value of following policy } \pi \text{ vs. } 0} = \mathbb{E}_0\left[\sum_{t=1}^{T} \mathbf{1}_{t \geq \tau_\pi} \underbrace{\left(\mu_{now,t}(S_t) - \mu_{next,t}(S_t)\right)}_{\text{“Advantage”: Difference in expected value if we start treating now vs. next time-step}}\right]$$

[Murphy 2005]

[Kakade 2003]

never acting

This is a single-step treatment effect estimation problem!

- Estimate treatment effect with a doubly robust estimator given available dataset D
- Can learn “nuisance” parameters (propensity weights and value function estimates) at a slower rate and still get sqrt(n) regret bounds, under various assumptions

Keeping a health metric above 0

Evolves with brownian motion

Treatment nudges it up, but at a cost

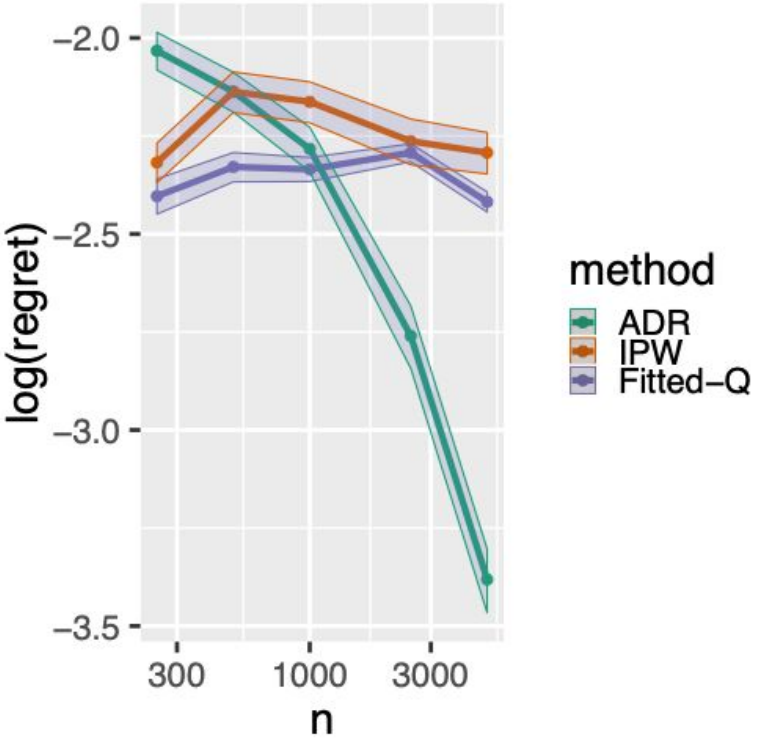
Always start with treatment ON

- *Optimal stopping time* of treatment?
- *Unknown* propensity
- Linear Decision Rules

  #covariates = 2

- Observe states + noise

Horizon T=10



Keeping a health metric above 0

Evolves with brownian motion

Treatment nudges it up, but at a cost

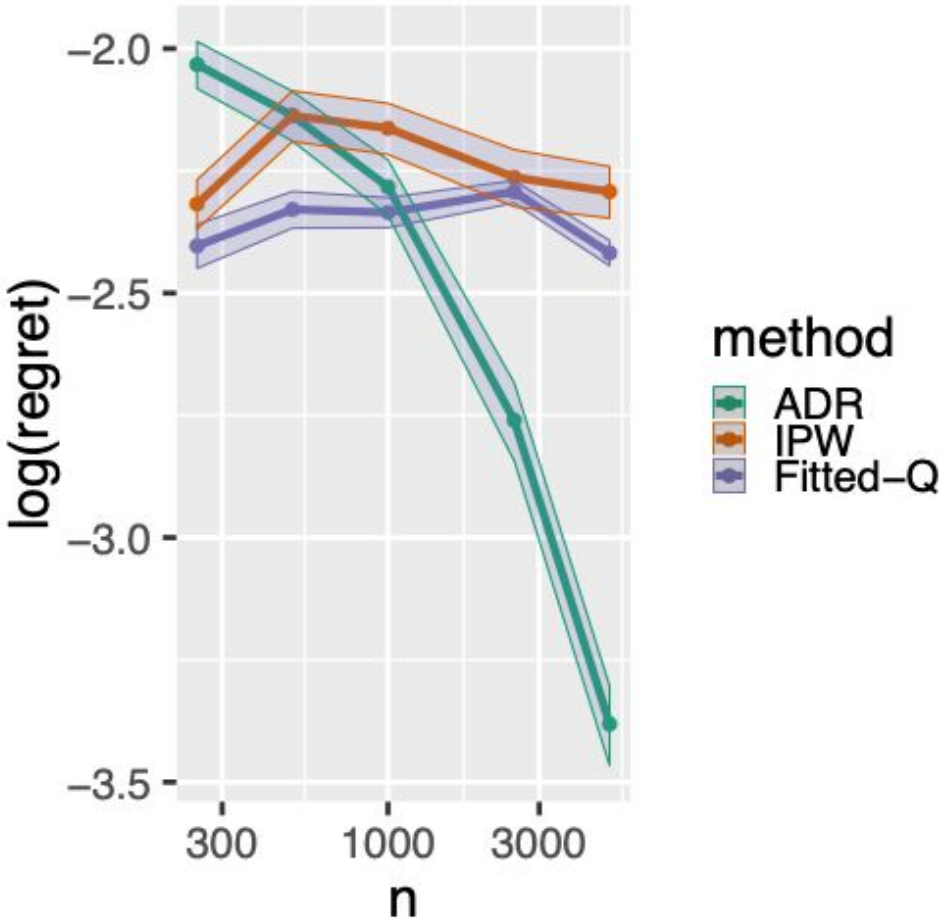
Always start with treatment ON

- *Optimal stopping time* of treatment?
- *Unknown* propensity
- Linear Decision Rules

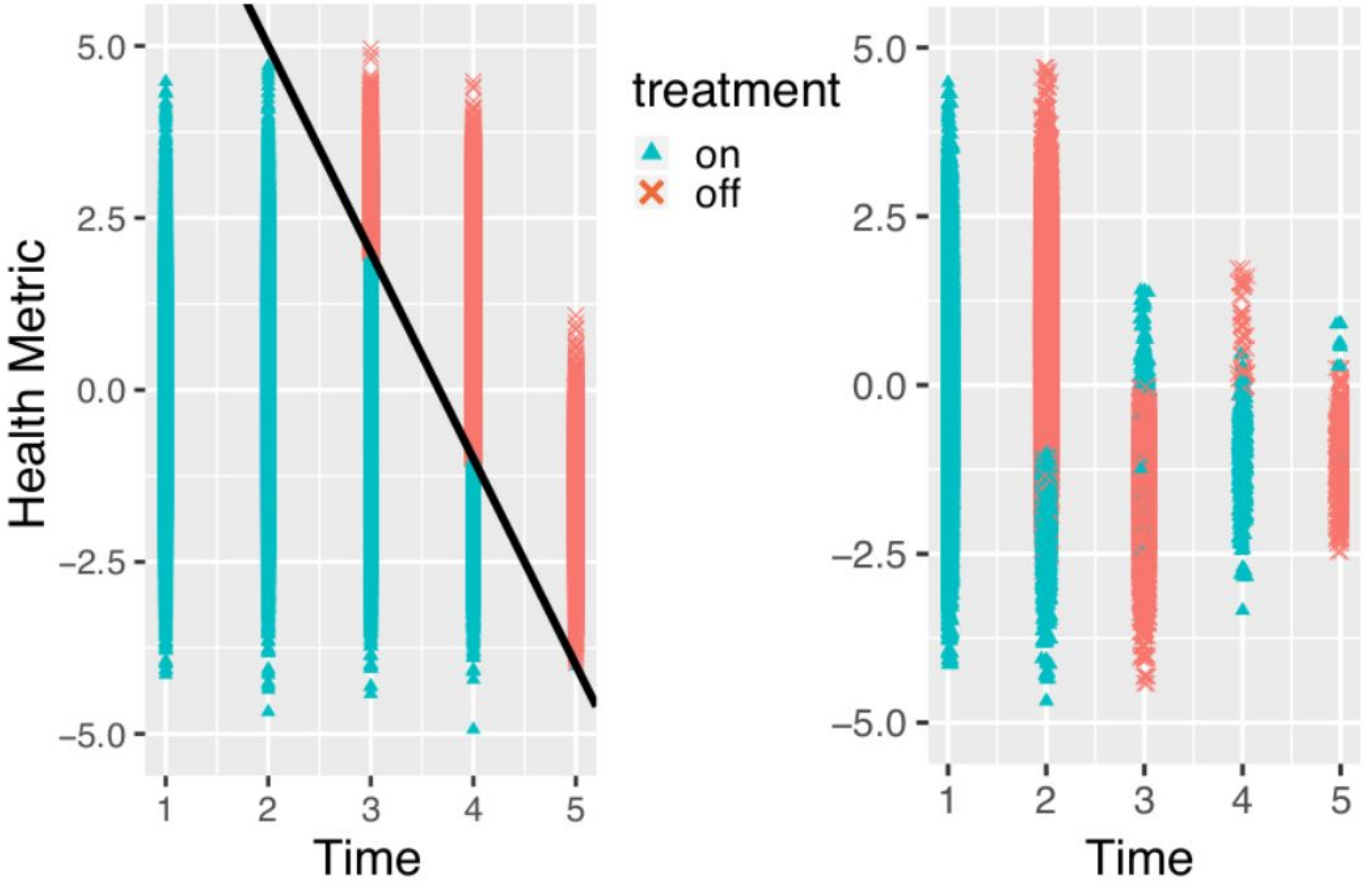
  #covariates = 2

- Observe states + noise

Horizon T=10



Fitted Q Iteration Policy Less Interpretable



Quest for Batch Policy Optimization with Generalization Guarantees

$$\underbrace{\arg\max_{\pi \in H_i} \quad \max_{H_i \in \{H_1, H_2, \ldots\}}}_{\text{Policy Optimization}} \quad \underbrace{\int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds}_{\text{Policy Evaluation}} \quad - \quad \underbrace{\sqrt{\frac{f(VC(h), \ldots)}{n}}}_{\text{Error Bound}}$$

AAAI 2015, AAAI 2016, L@S 2017, UAI 2017 (best paper), UAI 2019, Nie, B, Wagner arxiv

AAMAS 2014, AAAI 2015, AAAI 2016, ICML 2016, IJCAI 2016, AAAI 2017, NeurIPS 2017, NeurIPS 2018, ICML 2019

Nie, B, Wagner arxiv; Thomas, da Silva, Barto, B, arxiv

Quest for Batch Policy Optimization with Generalization Guarantees

$$\underbrace{\arg\max_{\pi \in H_i} \quad \max_{H_i \in \{H_1, H_2, \ldots\}}}_{\text{Policy Optimization}} \quad \underbrace{\int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds}_{\text{Policy Evaluation}} \quad - \quad \underbrace{\sqrt{\frac{f(VC(h), \ldots)}{n}}}_{\text{Error Bound}}$$

AAAI 2015, AAAI 2016, L@S 2017, UAI 2017 (best paper), UAI 2019, Nie, B, Wagner arxiv

AAMAS 2014, AAAI 2015, AAAI 2016, ICML 2016, IJCAI 2016, AAAI 2017, NeurIPS 2017, NeurIPS 2018, ICML 2019

Nie, B, Wagner arxiv; Thomas, da Silva, Barto, B, arxiv

Many others also purusing: Farahmand, Ghavamzadeh, Jiang, Lazaric, Li, Mannor, Murphy, Munos, Murphy, Pineau, Szepesvari, Yue, ...

→ Much to be done, including to relax common assumptions

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

Learn a Policy that Increases
Student Persistence

(Mandel, Liu, B, Popovic 2014)

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

Learned a Policy that Increased Student Persistence by +30%

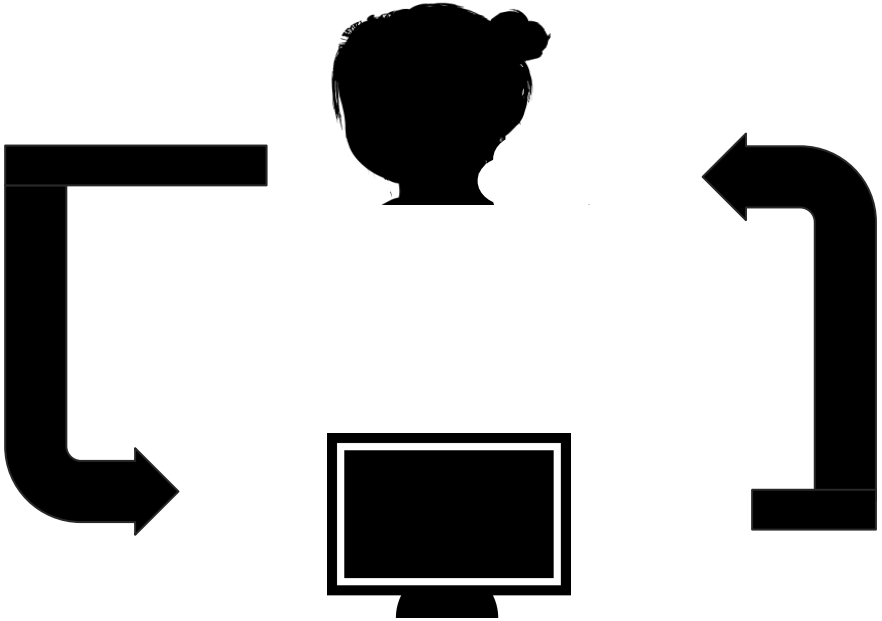
(Mandel, Liu, B, Popovic 2014)

Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

1. Understanding how to gather data to quickly learn to make good decisions
   - Matching upper/lower PAC bounds for tabular RL & instance dependent bounds

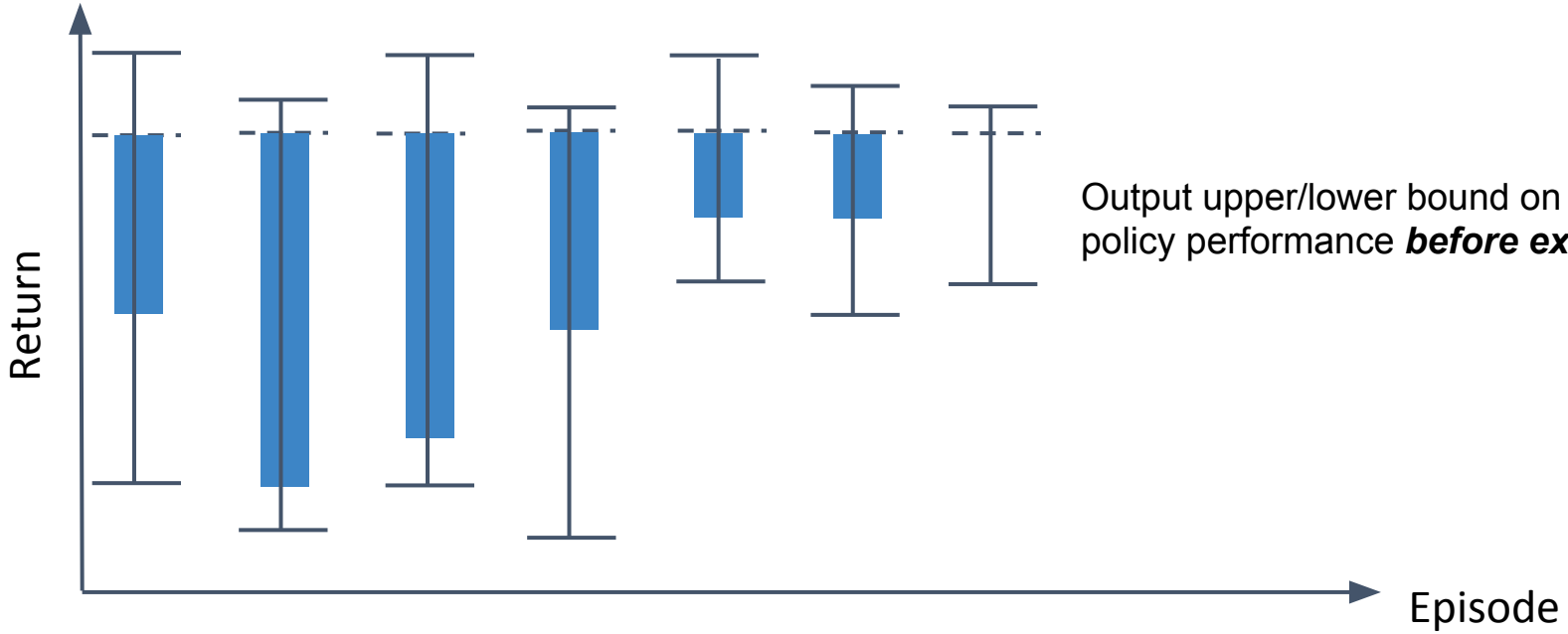
2. Inferring the outcome of alternate futures
   - More accurate estimators and steps towards strong generalization guarantees

Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions



⇒ If can learn to make good decisions faster, benefit more people

Better Together? Exploration & Counterfactuals/ Off Policy Evaluation (Dann, Wei, Li, B ICML 2019)



Output upper/lower bound on policy performance ***before execute it***

- Get tighter confidence bounds!
- Insight: policy deployed is approximation to optimal (static) policy

Ongoing Work and Open Questions

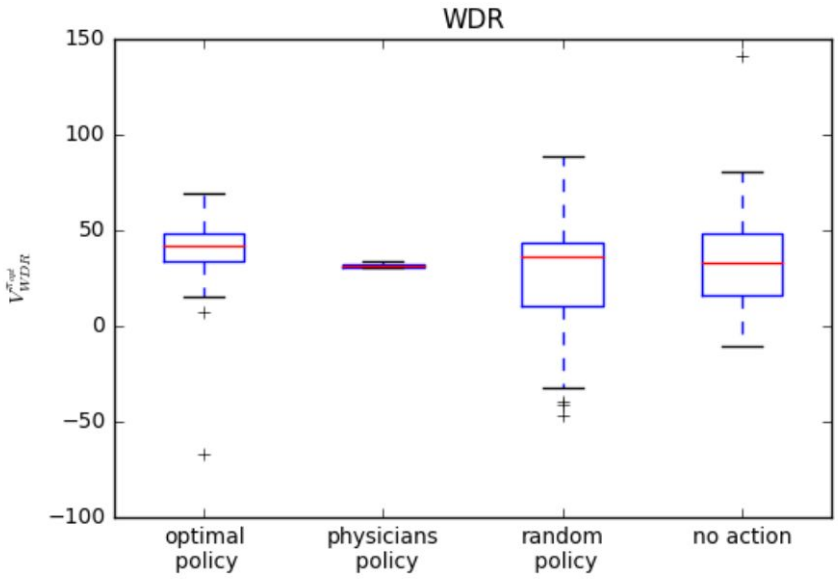
- Regret analysis using semiparamtrics for general Q functions
- Stronger generalization guarantees under weaker assumptions for high dimensional settings
- Nonstationarity
- Standard assumptions -- overlap, no confounding, no interference -- limit these approaches from impacting all the cases want to impact

WDR in Health Example

Sepsis treatment example

(Gottesman et al. arxiv 2018)

- Actions: IV fluids & vasopressors
- Reward: +100 survival, -100 death
- State space: 750 (discretized)
- 19,275 ICU patients

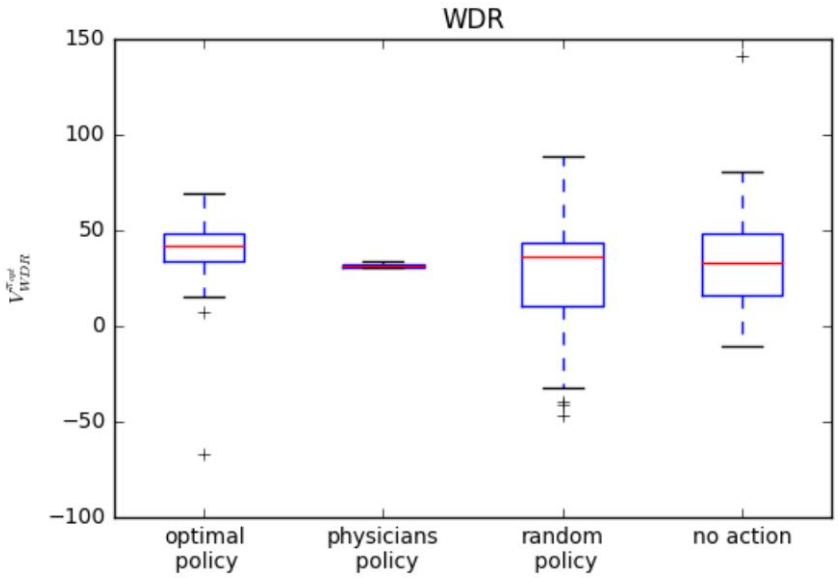


WDR in Health Example

Sepsis treatment example

(Gottesman et al. arxiv 2018)

- Actions: IV fluids & vasopressors
- Reward: +100 survival, -100 death
- State space: 750 (discretized)
- 19,275 ICU patients



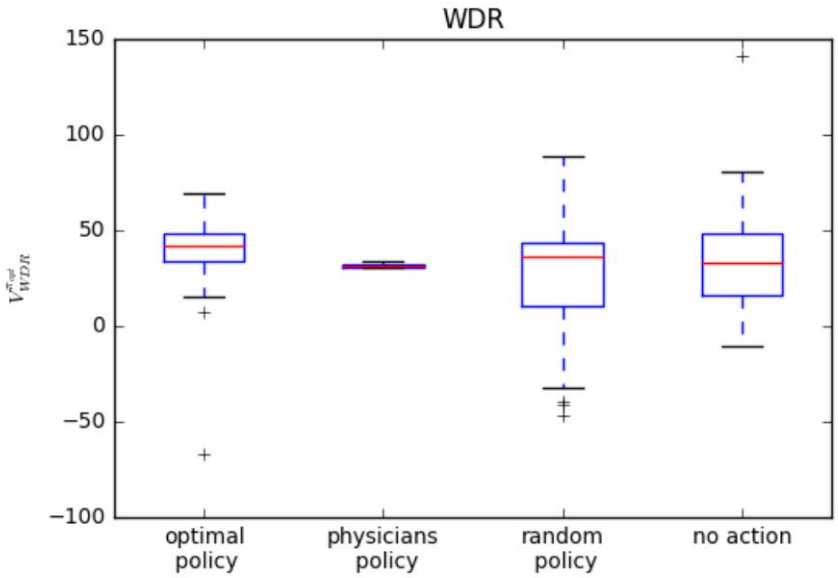
Our weighted DR (WDR) was only consistent off policy estimator tried (PDDR, PDIS, WPDIS, WDR) that could find an optimal policy which estimated would improve over prior

WDR in Health Example

Sepsis treatment example

(Gottesman et al. arxiv 2018)

- Actions: IV fluids & vasopressors
- Reward: +100 survival, -100 death
- State space: 750 (discretized)
- 19,275 ICU patients



Our weighted DR (WDR) was only consistent off policy estimator tried (PDDR, PDIS, WPDIS, WDR) that could find an optimal policy which estimated would improve over prior

<u>Under (common) assumption of no confounding, that is not likely to hold in practice</u>