Simons Institute, Dec 2020

Statistical Efficiency in Offline Reinforcement Learning

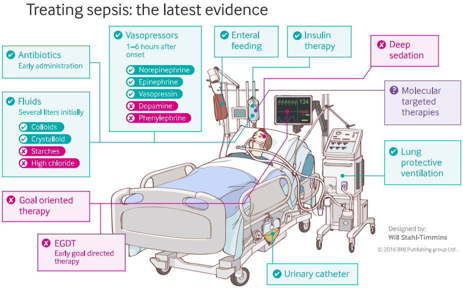
Nathan Kallus
Cornell University

Joint work with Masatoshi Uehara

Based on – "Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes" Kallus & Uehara,
– "Efficiently Breaking the Curse of Horizon: Double Reinforcement Learning in Infinite-Horizon Processes" Kallus & Uehara
– "Statistically Efficient Off-Policy Policy Gradients" Kallus & Uehara

Reinforcement Learning in Medicine

- Sepsis (extreme bodily reaction to infection) is 3rd leading cause of death worldwide! ☠️
  - Best treatment strategy unclear 😕
    - Lots of subtle symptoms, many levers, effect heterogeneity
  - Opportunity for reinforcement learning! 🤖💉



Off-Policy RL and the Curse of Horizon

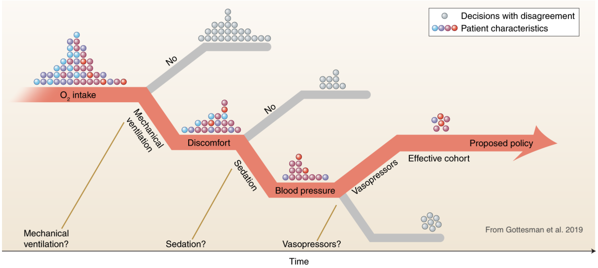
- In medicine and other high-stakes domains, exploration is limited and simulation unreliable
  - Must rely on existing data like EHRs 🏥📊📈

Off-Policy RL and the Curse of Horizon

- In medicine and other high-stakes domains, exploration is limited and simulation unreliable
  - Must rely on existing data like EHRs 🏥📊📈
- *E.g.*, Komorowski et al. ’18 proposed the “AI Clinician” for sepsis treatment by applying RL to observational ICU data

Off-Policy RL and the Curse of Horizon

- In medicine and other high-stakes domains, exploration is limited and simulation unreliable
  - Must rely on existing data like EHRs 🏥📊📈
- *E.g.*, Komorowski et al. '18 proposed the "AI Clinician" for sepsis treatment by applying RL to observational ICU data
  - Scrutiny, skepticism from RL and medical communities
  - Biggest gripe: unreliable due to *curse of horizon* 👻



From Gottesman et al. 2019

"Fig. 2: effective sample size" from Gottesman et al. 19

Statistically Efficient Offline RL

- **Aim**: Overcome fundamental limits in offline RL by leveraging Markovian, time-invariant, and ergodic structure
  - **Theme**: given limited data try to use it *efficiently*, and what's efficient depends on *structure*
- Contributions
  - Study efficiency limits in offline RL in MDPs for first time
    - Insight into when the curse of horizon bites
    - Problem-dependent phenomenon; not estimator-dependent
  - First efficient estimators for policy value/gradient in MDP in both finite- and infinite-horizon settings
    - Efficient even when nuisances estimated at slow rates by blackbox ML

This Talk

1 Introduction

2 Problem Setup

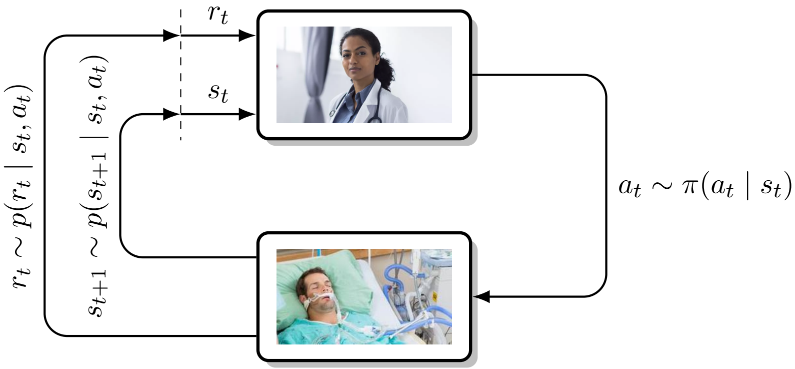
3 Efficiency Bounds

4 Efficient OPE via Double RL

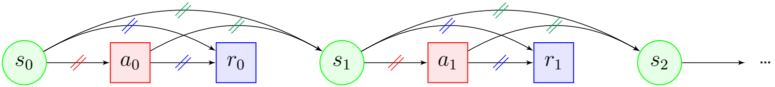
5 Efficient OPG & Policy Learning

6 Experimental Results

Markov Decision Processes


r_t
s_t
$a_t \sim \pi(a_t \mid s_t)$
$r_t \sim p(r_t \mid s_t, a_t)$
$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$

Off-Policy Evaluation and Gradients

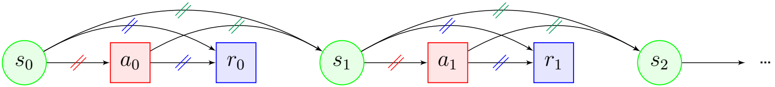


- MDP (state and reward probabilities)

  $+$ policy (action probabilities)

  $=$ joint distribution p_π over $(s_0, a_0, r_0, s_1, a_1, \ldots)$
  - Policy value: $J_T(\pi) = \frac{1}{\sum_{t=0}^{T}\gamma^t}\mathbb{E}_{p_\pi}\left[\sum_{t=0}^{T}\gamma^t r_t\right]$
  - $J_\infty(\pi) = \lim_{T\to\infty} J_T(\pi)$
- **Off-policy evaluation**: given π, estimate $J_T(\pi)$ from N observations of $(s_0, a_0, r_0, \ldots, a_T, r_T)$ from p_{π^b}
  - Behavior policy p_{π^b} may be known or unknown

Off-Policy Evaluation and Gradients



- MDP (state and reward probabilities)
  $+$ policy (action probabilities)
  $=$ joint distribution p_π over $(s_0, a_0, r_0, s_1, a_1, \ldots)$
  - Policy value: $J_T(\pi) = \frac{1}{\sum_{t=0}^{T} \gamma^t} \mathbb{E}_{p_\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$
  - $J_\infty(\pi) = \lim_{T\to\infty} J_T(\pi)$
- **Off-policy evaluation**: given π, estimate $J_T(\pi)$ from N observations of $(s_0, a_0, r_0, \ldots, a_T, r_T)$ from p_{π^b}
  - Behavior policy p_{π^b} may be known or unknown

Off-Policy Evaluation and Gradients

- For learning, suppose given policy class $\Pi = \{\pi^\theta : \theta \in \Theta\}$
    - Let $J_T(\theta) = J_T(\pi^\theta)$
    - **Off-policy gradient**: given N observations of $(s_0, a_0, r_0, \ldots, a_T, r_T)$ from p_{π^b}, estimate

$$Z_T(\theta) = \nabla_\theta J_T(\theta)$$

Off-Policy Evaluation and Gradients

- For learning, suppose given policy class $\Pi = \{\pi^\theta : \theta \in \Theta\}$
  - Let $J_T(\theta) = J_T(\pi^\theta)$
  - **Off-policy gradient**: given N observations of $(s_0, a_0, r_0, \ldots, a_T, r_T)$ from p_{π^b}, estimate

$$Z_T(\theta) = \nabla_\theta J_T(\theta) = \tfrac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_{p_{\pi^\theta}} \left[\sum_{t=0}^T \gamma^t r_t \sum_{k=0}^t g_k \right]$$

$$g_t = \nabla_\theta \log \pi^\theta(a_t \mid s_t) \quad \text{(policy score)}$$

Off-Policy Evaluation and Gradients

- For learning, suppose given policy class $\Pi = \{\pi^\theta : \theta \in \Theta\}$
  - Let $J_T(\theta) = J_T(\pi^\theta)$
  - **Off-policy gradient**: given N observations of $(s_0, a_0, r_0, \ldots, a_T, r_T)$ from p_{π^b}, estimate

$$Z_T(\theta) = \nabla_\theta J_T(\theta) = \frac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_{p_{\pi^\theta}} \left[\sum_{t=0}^T \gamma^t r_t \sum_{k=0}^t g_k\right]$$

$$g_t = \nabla_\theta \log \pi^\theta(a_t \mid s_t) \quad \text{(policy score)}$$

Off-Policy Evaluation and Gradients

- For learning, suppose given policy class $\Pi = \{\pi^\theta : \theta \in \Theta\}$
  - Let $J_T(\theta) = J_T(\pi^\theta)$
  - **Off-policy gradient**: given N observations of $(s_0, a_0, r_0, \ldots, a_T, r_T)$ from p_{π^b}, estimate

$$Z_T(\theta) = \nabla_\theta J_T(\theta) = \tfrac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_{p_{\pi^\theta}} \left[\textstyle\sum_{t=0}^T \gamma^t r_t \sum_{k=0}^t g_k \right]$$

$$g_t = \nabla_\theta \log \pi^\theta(a_t \mid s_t) \quad \text{(policy score)}$$

  - Can be used for off-policy learning via gradient ascent
    - (Policy gradient methods have driven a lot of recent RL successes in *online* settings with experimentation/simulation)

Existing Approaches (very abridged version 😅)

- (OPE) SIS estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^{T} \gamma^t} \mathbb{E}_N[\sum_{t=0}^{T} \gamma^t \lambda_t r_t]$
  - $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^{t} \rho_k$ is the *cumulative density ratio*
  - Changes measure from $\mathbb{E}_{p_{\pi^b}}$ to $\mathbb{E}_{p_{\pi^\theta}}$

Existing Approaches (very abridged version 😅)

- (OPE) SIS estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_N[\sum_{t=0}^T \gamma^t \lambda_t r_t]$
  - $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^t \rho_k$ is the *cumulative density ratio*
  - Changes measure from $\mathbb{E}_{p_{\pi^b}}$ to $\mathbb{E}_{p_{\pi^\theta}}$
- (OPE) "Doubly Robust" ("DR") estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_N[\sum_{t=0}^T \gamma^t \left(\lambda_t(r_t - \hat{q}_t) + \lambda_{t-1}\mathbb{E}_{a_t \sim \pi^\theta}[\hat{q}_t \mid s_t]\right)]$
  - $q_t = \mathbb{E}_{p_{\pi^\theta}}[\sum_{k=t}^T \gamma^{k-t} r_k \mid s_t, a_t]$ is *q-function*; $\hat{q}_t$ an estimator
    - Notice $q_t = q$ is independent of t for $T = \infty$
  - *Lots* of variants: Jiang & Li '16, Thomas & Brunskill '16, K '18, Farajtabar et al. '18, K & Uehara '19, ...

Existing Approaches (very abridged version 😅)

- (OPE) SIS estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_N[\sum_{t=0}^T \gamma^t \lambda_t r_t]$
  - $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^t \rho_k$ is the *cumulative density ratio*
  - Changes measure from $\mathbb{E}_{p_{\pi^b}}$ to $\mathbb{E}_{p_{\pi^\theta}}$
- (OPE) "Doubly Robust" ("DR") estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_N[\sum_{t=0}^T \gamma^t \left(\lambda_t(r_t - \hat{q}_t) + \lambda_{t-1}\mathbb{E}_{a_t \sim \pi^\theta}[\hat{q}_t \mid s_t]\right)]$
  - $q_t = \mathbb{E}_{p_{\pi^\theta}}[\sum_{k=t}^T \gamma^{k-t} r_k \mid s_t, a_t]$ is *q-function*; $\hat{q}_t$ an estimator
    - Notice $q_t = q$ is independent of t for $T = \infty$
  - *Lots* of variants: Jiang & Li '16, Thomas & Brunskill '16, K '18, Farajtabar et al. '18, K & Uehara '19, ...
- (OPG) REINFORCE: $\hat{Z}_T(\theta) = \frac{1}{\sum_{t=0}^T \gamma^t} \mathbb{E}_N[\sum_{t=0}^T \gamma^t \lambda_t r_t \sum_{k=0}^t g_k]$

Existing Approaches (very abridged version 😅)

- (OPE) SIS estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^{T} \gamma^t} \mathbb{E}_N[\sum_{t=0}^{T} \gamma^t \lambda_t r_t]$
  - $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^{t} \rho_k$ is the *cumulative density ratio*
  - Changes measure from $\mathbb{E}_{p_{\pi^b}}$ to $\mathbb{E}_{p_{\pi^\theta}}$
- (OPE) “Doubly Robust” (“DR”) estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^{T} \gamma^t} \mathbb{E}_N[\sum_{t=0}^{T} \gamma^t \left(\lambda_t (r_t - \hat{q}_t) + \lambda_{t-1} \mathbb{E}_{a_t \sim \pi^\theta}[\hat{q}_t \mid s_t]\right)]$
  - $q_t = \mathbb{E}_{p_{\pi^\theta}}[\sum_{k=t}^{T} \gamma^{k-t} r_k \mid s_t, a_t]$ is *q-function*; $\hat{q}_t$ an estimator
    - Notice $q_t = q$ is independent of t for $T = \infty$
  - *Lots* of variants: Jiang & Li ’16, Thomas & Brunskill ’16, K ’18, Farajtabar et al. ’18, K & Uehara ’19, ...
- (OPG) REINFORCE: $\hat{Z}_T(\theta) = \frac{1}{\sum_{t=0}^{T} \gamma^t} \mathbb{E}_N[\sum_{t=0}^{T} \gamma^t \lambda_t r_t \sum_{k=0}^{t} g_k]$
- (OPG) Off-PAC: $\hat{Z}_T(\theta) = \frac{1}{\sum_{t=0}^{T} \gamma^t} \mathbb{E}_N[\sum_{t=0}^{T} \gamma^t \lambda_t g_t \hat{q}_t]$

Existing Approaches (very abridged version 😅)

- (OPE) SIS estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^{T}\gamma^t}\mathbb{E}_N[\sum_{t=0}^{T}\gamma^t\lambda_t r_t]$
  - $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^{t}\rho_k$ is the *cumulative density ratio*
  - Changes measure from $\mathbb{E}_{p_{\pi^b}}$ to $\mathbb{E}_{p_{\pi^\theta}}$
- (OPE) "Doubly Robust" ("DR") estimator: $\hat{J}_T(\theta) = \frac{1}{\sum_{t=0}^{T}\gamma^t}\mathbb{E}_N[\sum_{t=0}^{T}\gamma^t\left(\lambda_t(r_t - \hat{q}_t) + \lambda_{t-1}\mathbb{E}_{a_t\sim\pi^\theta}[\hat{q}_t \mid s_t]\right)]$
  - $q_t = \mathbb{E}_{p_{\pi^\theta}}[\sum_{k=t}^{T}\gamma^{k-t}r_k \mid s_t, a_t]$ is *q-function*; $\hat{q}_t$ an estimator
    - Notice $q_t = q$ is independent of t for $T = \infty$
  - *Lots* of variants: Jiang & Li '16, Thomas & Brunskill '16, K '18, Farajtabar et al. '18, K & Uehara '19, ...
- (OPG) REINFORCE: $\hat{Z}_T(\theta) = \frac{1}{\sum_{t=0}^{T}\gamma^t}\mathbb{E}_N[\sum_{t=0}^{T}\gamma^t\lambda_t r_t\sum_{k=0}^{t}g_k]$
- (OPG) Off-PAC: $\hat{Z}_T(\theta) = \frac{1}{\sum_{t=0}^{T}\gamma^t}\mathbb{E}_N[\sum_{t=0}^{T}\gamma^t\lambda_t g_t\hat{q}_t]$
- Naïve view of curse of horizon: if $\frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)} \approx C > \gamma^{-1}$, we get $\lambda_t \approx C^t$, and all of the above explode exponentially 😱

Existing Approaches (very abridged version 😅)

- Direct method: $\mathbb{E}_N[\mathbb{E}_{a_0 \sim \pi^e}[\hat{q}(s_0, a_0) \mid s_0]]$
  - Can directly bake-in MDP structure into q-model
- Liu et al. (2018): importance sampling using stationary density ratios in infinite horizons
  Xie et al. (2019): importance sampling using marginalized density ratios in time-varying MDPs and finite state spaces
- All of the above leverage MDP structure! 😎
  - Motivates our current study
  - But still not efficient 😕
  - Will generally have *suboptimal* leading constant
  - In non-tabular settings, will generally even have *slow* rate ($\omega((NT)^{-1/2})$) 😫

Efficiency (very abridged version 😅)

- Consider model $\mathcal{M}$ and parameter of interest $\tau : \mathcal{M} \to \mathbb{R}$
  - Given iid data $X_i \sim \mathbb{P} \in \mathcal{M}$, want a good estimator $\hat{\tau}_n(X_1, \ldots, X_n)$ for $\tau(\mathbb{P})$ that uses data to the mostest

Efficiency (very abridged version 😅)

- Consider model $\mathcal{M}$ and parameter of interest $\tau : \mathcal{M} \to \mathbb{R}$
  - Given iid data $X_i \sim \mathbb{P} \in \mathcal{M}$, want a good estimator $\hat{\tau}_n(X_1, \ldots, X_n)$ for $\tau(\mathbb{P})$ that uses data to the mostest
- Semiparametric efficiency says: any estimator that works for all instances $\mathbb{P} \in \mathcal{M}$ (is regular) must satisfy for each $\mathbb{P} \in \mathcal{M}$:

$$\liminf n \cdot \mathbb{E}[(\hat{\tau}_n(X_{1:n}) - \tau(\mathbb{P}))^2] \geq \underbrace{\mathbb{E}[\psi^2(X; \mathbb{P})]}_{\text{Efficiency bound}},$$

Efficient influence function ψ is the least-norm derivative of τ

Efficiency (very abridged version 😅)

- Consider model $\mathcal{M}$ and parameter of interest $\tau : \mathcal{M} \to \mathbb{R}$
  - Given iid data $X_i \sim \mathbb{P} \in \mathcal{M}$, want a good estimator $\hat{\tau}_n(X_1, \ldots, X_n)$ for $\tau(\mathbb{P})$ that uses data to the mostest
- Semiparametric efficiency says: any estimator that works for all instances $\mathbb{P} \in \mathcal{M}$ (is regular) must satisfy for each $\mathbb{P} \in \mathcal{M}$:

$$\liminf n \cdot \mathbb{E}[(\hat{\tau}_n(X_{1:n}) - \tau(\mathbb{P}))^2] \geq \underbrace{\mathbb{E}[\psi^2(X;\mathbb{P})]}_{\text{Efficiency bound}},$$

  Efficient influence function ψ is the least-norm derivative of τ
- For us: $\tau = J_T(\theta), Z_T(\theta)$, $\mathcal{M}$ = set of *all* p_{π^b} for *all* MDPs

Efficiency (very abridged version 😅)

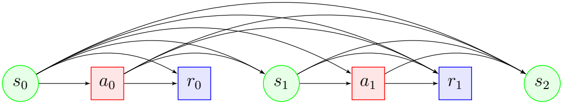
- Consider model $\mathcal{M}$ and parameter of interest $\tau : \mathcal{M} \to \mathbb{R}$
  - Given iid data $X_i \sim \mathbb{P} \in \mathcal{M}$, want a good estimator $\hat{\tau}_n(X_1, \dots, X_n)$ for $\tau(\mathbb{P})$ that uses data to the mostest
- Semiparametric efficiency says: any estimator that works for all instances $\mathbb{P} \in \mathcal{M}$ (is regular) must satisfy for each $\mathbb{P} \in \mathcal{M}$:

$$\liminf n \cdot \mathbb{E}[(\hat{\tau}_n(X_{1:n}) - \tau(\mathbb{P}))^2] \geq \underbrace{\mathbb{E}[\psi^2(X; \mathbb{P})]}_{\text{Efficiency bound}},$$

  Efficient influence function ψ is the least-norm derivative of τ
- For us: $\tau = J_T(\theta), Z_T(\theta)$, $\mathcal{M} =$ set of *all* p_{π^b} for *all* MDPs
  - Will actually also be insightful to consider other models ...

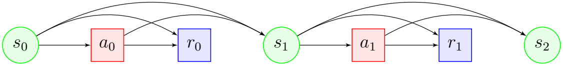
Three Nested Models: MDP $\subseteq$ TMDP $\subseteq$ NMDP

- $\mathcal{M}_1$: Non-Markov Decision Process (NMDP)



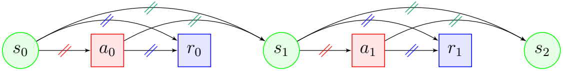
$$\mathcal{H}_{a_t} = (s_0, a_0, \ldots, s_t, a_t)$$
$$s_t \sim p_t(s_t \mid \mathcal{H}_{a_t})$$
$$r_t \sim p_t(r_t \mid \mathcal{H}_{a_t})$$
$$a_t \sim \pi_t(a_t \mid \mathcal{H}_{s_t})$$

- $\mathcal{M}_2$: Time-Varying Markov Decision Process (TMDP)



$$p_t(s_t \mid \mathcal{H}_{a_t}) = p_t(s_t \mid s_t, a_t)$$
$$p_t(r_t \mid \mathcal{H}_{a_t}) = p_t(r_t \mid s_t, a_t)$$
$$\pi_t(a_t \mid \mathcal{H}_{s_t}) = \pi_t(a_t \mid s_t)$$

- $\mathcal{M}_3$: Time-Invariant Markov Decision Process (MDP)



$$p_t(s' \mid s, a) = p(s' \mid s, a)$$
$$p_t(r \mid s, a) = p(r \mid s, a)$$
$$\pi_t(a \mid s) = \pi(a \mid s)$$

This Talk

1. Introduction
2. Problem Setup
3. Efficiency Bounds
4. Efficient OPE via Double RL
5. Efficient OPG & Policy Learning
6. Experimental Results

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE	Assumptions
NMDP		
TMDP		
MDP		

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE	Assumptions
NMDP	∞ 👻	$\exp(\mathbb{E}[\log(\rho_t)]) \geq 1/\gamma$
TMDP		
MDP		

- Recall $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^t \rho_k$

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE		Assumptions
NMDP	∞	👻	$\exp(\mathbb{E}[\log(\rho_t)]) \geq 1/\gamma$
	$\mathcal{O}(1/N)$	👾	$\lambda_t = o(\gamma^{-t}), \quad \begin{array}{c} N \to \infty, \\ T = \omega(\log^{1/2}(N)) \end{array}$
TMDP			
MDP			

- Recall $\rho_t = \frac{\pi^\theta(a_t \mid s_t)}{\pi^b(a_t \mid s_t)}, \quad \lambda_t = \prod_{k=0}^{t} \rho_k$

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE	Assumptions
NMDP	∞ 👻	$\exp(\mathbb{E}[\log(\rho_t)]) \geq 1/\gamma$
	$\mathcal{O}(1/N)$ 👾	$\lambda_t = o(\gamma^{-t})$, $\begin{matrix} N \to \infty, \\ T = \omega(\log^{1/2}(N)) \end{matrix}$
TMDP	$\mathcal{O}(1/N)$ 👾	$\mathbb{E}[\lambda_t \mid s_t, a_t] = o(\gamma^{-t})$, $\begin{matrix} N \to \infty, \\ T = \omega(\log^{1/2}(N)) \end{matrix}$
MDP		

- Recall $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^{t} \rho_k$

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE		Assumptions
NMDP	∞	👻	$\exp(\mathbb{E}[\log(\rho_t)]) \geq 1/\gamma$
	$\mathcal{O}(1/N)$	👾	$\lambda_t = o(\gamma^{-t}),\quad \begin{matrix} N \to \infty, \\ T = \omega(\log^{1/2}(N)) \end{matrix}$
TMDP	$\mathcal{O}(1/N)$	💩	$\mathbb{E}[\lambda_t \mid s_t, a_t] = o(\gamma^{-t}),\quad \begin{matrix} N \to \infty, \\ T = \omega(\log^{1/2}(N)) \end{matrix}$
MDP			

- Recall $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}, \quad \lambda_t = \prod_{k=0}^{t} \rho_k$

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE		Assumptions
NMDP	∞	👻	$\exp(\mathbb{E}[\log(\rho_t)]) \geq 1/\gamma$
	$\mathcal{O}(1/N)$	👾	$\lambda_t = o(\gamma^{-t})$, $N \to \infty$, $T = \omega(\log^{1/2}(N))$
TMDP	$\mathcal{O}(1/N)$	💩	$\mathbb{E}[\lambda_t \mid s_t, a_t] = o(\gamma^{-t})$, $N \to \infty$, $T = \omega(\log^{1/2}(N))$
MDP	$\mathcal{O}(1/(NT))$	🤑	$T \to \infty$, $N \geq 1$, Ergodic

- Recall $\rho_t = \frac{\pi^\theta(a_t|s_t)}{\pi^b(a_t|s_t)}$, $\lambda_t = \prod_{k=0}^t \rho_k$

Efficiency Bounds for Infinite-Horizon OPE/OPG

- Setting: observe N trajectories of length T, estimate $J_\infty(\pi)$

Model	Efficient MSE	
NMDP	∞	👻
TMDP	$\mathcal{O}(1/N)$	💩
MDP	$\mathcal{O}(1/(NT))$	🤑

This Talk

1. Introduction

2. Problem Setup

3. Efficiency Bounds

4. **Efficient OPE via Double RL**

5. Efficient OPG & Policy Learning

6. Experimental Results

Overview

- **Derive** the efficient influence function (EIF) ψ for each case: {MDP,TMDP,NMDP} $\times$ {OPE,OPG} $\times$ $\{T < \infty, T = \infty\}$
  - EIFs involve some unknown *nuisances*: $\psi = \phi_\eta - \tau$
    - *E.g.*, the q-function is a nuisance in all of the cases
  - If knew η, $\tilde{\tau} = \mathbb{E}_N[\phi_\eta]$ would be an efficient estimator

Overview

- **Derive** the efficient influence function (EIF) ψ for each case: $\{\text{MDP},\text{TMDP},\text{NMDP}\} \times \{\text{OPE},\text{OPG}\} \times \{T < \infty, T = \infty\}$
  - EIFs involve some unknown *nuisances*: $\psi = \phi_\eta - \tau$
    - *E.g.*, the q-function is a nuisance in all of the cases
  - If knew η, $\tilde{\tau} = \mathbb{E}_N[\phi_\eta]$ would be an efficient estimator
- Idea: estimate $\hat{\eta}$ and use $\hat{\tau} = \mathbb{E}_N[\phi_{\hat{\eta}}]$
  - But need to make sure this works

Overview

- **Derive** the efficient influence function (EIF) ψ for each case: $\{$MDP,TMDP,NMDP$\} \times \{$OPE,OPG$\} \times \{T < \infty, T = \infty\}$
  - EIFs involve some unknown *nuisances*: $\psi = \phi_\eta - \tau$
    - *E.g.*, the q-function is a nuisance in all of the cases
  - If knew η, $\tilde{\tau} = \mathbb{E}_N[\phi_\eta]$ would be an efficient estimator
- Idea: estimate $\hat{\eta}$ and use $\hat{\tau} = \mathbb{E}_N[\phi_{\hat{\eta}}]$
  - But need to make sure this works
- **Prove** that the EIFs satisfy **double robustness**
  - For OPE: $\tau = \mathbb{E}[\phi_{(\eta_1,\star)}] = \mathbb{E}[\phi_{(\star,\eta_2)}]$ (Special case for OPG)
  - $\implies \partial_{\eta'}\mathbb{E}[\phi_{\eta'}] \mid_{\eta'=\eta} = 0$ so $\hat{\tau}$ is insensitive to errors in $\hat{\eta}$

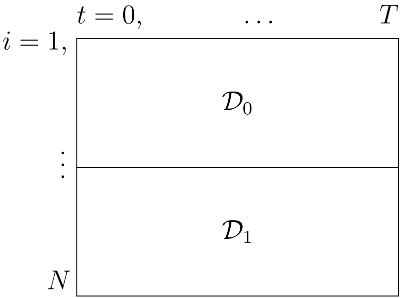
Overview

- **Derive** the efficient influence function (EIF) ψ for each case: $\{\text{MDP,TMDP,NMDP}\} \times \{\text{OPE,OPG}\} \times \{T < \infty, T = \infty\}$
    - EIFs involve some unknown *nuisances*: $\psi = \phi_\eta - \tau$
        - *E.g.*, the q-function is a nuisance in all of the cases
    - If knew η, $\tilde{\tau} = \mathbb{E}_N[\phi_\eta]$ would be an efficient estimator
- Idea: estimate $\hat{\eta}$ and use $\hat{\tau} = \mathbb{E}_N[\phi_{\hat{\eta}}]$
    - But need to make sure this works
- **Prove** that the EIFs satisfy **double robustness**
    - For OPE: $\tau = \mathbb{E}[\phi_{(\eta_1,\star)}] = \mathbb{E}[\phi_{(\star,\eta_2)}]$ (Special case for OPG)
    - $\implies \partial_{\eta'}\mathbb{E}[\phi_{\eta'}] \mid_{\eta'=\eta} = 0$ so $\hat{\tau}$ is insensitive to errors in $\hat{\eta}$
- To enable flexible ML estimators for $\hat{\eta}$, use cross-fitting (Double ML; Chernozhukov et al., 2018)
    - (Special case for infinite horizon due to dependent data)
- Result: Efficient Estimation via Double RL

DRL for OPE in MDP

- Step 1: Split the data into folds

Two folds over many trajectories



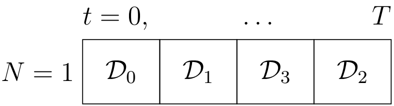
$t = 0, \quad \dots \quad T$

$i = 1, \ \vdots \ N$

$\mathcal{D}_0$

$\mathcal{D}_1$

Four folds over one trajectory



$t = 0, \quad \dots \quad T$

$N = 1$

$\mathcal{D}_0$ $\mathcal{D}_1$ $\mathcal{D}_3$ $\mathcal{D}_2$

DRL for OPE in MDP

- Step 1: Split the data into folds

	$t = 0,$		$\dots$	T
$N = 1$	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_3$	$\mathcal{D}_2$

DRL for OPE in MDP

- Step 1: Split the data into folds

	$t = 0,$		$\dots$	T
$N = 1$	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_3$	$\mathcal{D}_2$

- Let $w(s)$ be the ratio of the γ-discounted average visitation distribution at s under π^θ and the *undiscounted* stationary distribution at s under π^b
  - (This is slightly different than the ratio in Liu et al. 2018)
- For each fold j, construct* estimators $\hat{w}^{(j)}$ and $\hat{q}^{(j)}$ for w and q based only on the training data $\mathcal{D}_j$

DRL for OPE in MDP

- Step 1: Split the data into folds

	$t = 0,$		$\dots$	T
$N = 1$	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_3$	$\mathcal{D}_2$

- Set $\hat{J}_{\mathrm{DRL(MDP)}}(\theta)$ to

$$\frac{1}{(T+1)} \sum_{j=0}^{3} \sum_{t \in \mathcal{D}_j} \phi(s_t, a_t, r_t, s_{t+1}; \hat{w}^{(3-j)}, \hat{q}^{(3-j)})$$

where $\phi(s, a, r, s'; w, q) = (1-\gamma)\mathbb{E}_{p_0}[\mathbb{E}_{a_0 \sim \pi^\theta}[q(s_0, a_0) \mid s_0]]$
$$+\, w(s)\rho(a, s)\left(r + \gamma \mathbb{E}_{a' \sim \pi^\theta}[q(s', a') \mid s'] - q(s, a)\right)$$

Efficiency of DRL in MDP

Assumption

$p_{\pi^b}, p_{\pi^\theta}$ induce Haris ergodic chains, corresponding w is a bounded r.v., and $\hat{w}^{(j)}, \hat{q}^{(j)}$ are bounded

Theorem

Assume $\|\hat{q}^{(j)} - q\|_2 = \mathrm{o}_p((NT)^{-\alpha_1})$, $\|\hat{w}^{(j)} - w\|_2 = \mathrm{o}_p((NT)^{-\alpha_2})$, $\alpha_1 > 0,\ \alpha_2 > 0, \alpha_1 + \alpha_2 \geq 1/2$, *and* p_{π^b} *is a strongly* ρ*-mixing process. Then,* $\sqrt{NT}(\hat{J}_{\mathrm{DRL(MDP)}}(\theta) - J(\theta)) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi^2_{\mathrm{MDP}}])$.

Key feature: no assumptions on $\hat{q}, \hat{w}$, just a slow rate

$\implies$ can use black-box ML to fit nuisances

(Works without cross-fold if we impose Donsker conditions)

Double Robustness of DRL

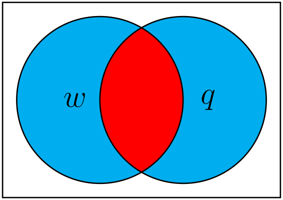
Theorem

Assume $\phi(\cdot,\cdot,\cdot,\cdot;\hat{w}^{(j)},\hat{q}^{(j)}) \in \mathcal{F}_\phi$ *almost surely where* $\mathcal{F}_\phi$ *is VC-major. Assume* $\|\hat{w}^{(j)} - w^\dagger\|_2 = \mathrm{o}_p(1)$*,* $\|\hat{q}^{(j)} - q^\dagger\|_2 = \mathrm{o}_p(1)$*, and either* $w^\dagger = w$ *or* $q^\dagger = q$*. Then,* $\hat{J}_{\mathrm{DRL(MDP)}}(\theta) \to J(\theta)$*.*

Double Robustness of DRL

Theorem

Assume $\phi(\cdot,\cdot,\cdot,\cdot;\hat{w}^{(j)},\hat{q}^{(j)}) \in \mathcal{F}_\phi$ *almost surely where* $\mathcal{F}_\phi$ *is VC-major. Assume* $\|\hat{w}^{(j)} - w^\dagger\|_2 = \mathrm{o}_p(1)$, $\|\hat{q}^{(j)} - q^\dagger\|_2 = \mathrm{o}_p(1)$, *and either* $w^\dagger = w$ *or* $q^\dagger = q$. *Then,* $\hat{J}_{\mathrm{DRL(MDP)}}(\theta) \to J(\theta)$.



Guarantees for DRL

- Examples cases:
  - Tabular case in (T)MDP: If state and action spaces finite, can obtain $O_p(n^{-1/2})$ rate for nuisances and get efficient estimates (don't even need cross-fold)
  - Finite state space, known behavior policy in TMDP: Xie et al. (2019) provide $O_p(n^{-1/2})$ rate for marginalized density ratio, so only need $o_p(1)$ for q-estimate (no rate)
    - Boundedness is enough – can use kernel regression estimates
  - General non-parametric case: can use flexible ML estimates; *e.g.*, *DICE; more generally: w, q defined by conditional moment restrictions so can use Newey (1990), Ai and Chen (2003), Bennett, K, Schnabel (2019).

Guarantees for DRL

- More results in papers...
  - Efficiency in $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$
  - Efficiency under various conditions on plug-in estimators
  - Finite-sample guarantees (PAC-style)
  - Finite horizon
  - Inefficiency of other estimators
    - IS, Marginalized IS, Stationary IS
    - “DR” in $\mathcal{M}_2, \mathcal{M}_3$

This Talk

1. Introduction
2. Problem Setup
3. Efficiency Bounds
4. Efficient OPE via Double RL
5. Efficient OPG & Policy Learning
6. Experimental Results

Efficient Off-Policy Policy Gradients

- Need additional nuisances:
  - q, w as before; Also $d^q = \nabla_\theta q$, $d^w = \nabla_\theta w$
- Estimation technique similar to before:
  - Cross-fold estimate q, w, d^q, d^w
  - Plug into EIF that we derived

Theorem (Efficiency)

$$\|\hat{w}^{(j)} - w\| = \mathrm{o}_p((NT)^{-\alpha_w}),\ \|\hat{d}^{w,(j)} - d^w\| = \mathrm{o}_p((NT)^{-\beta_w}),$$
$$\|\hat{q}^{(j)} - q\| = \mathrm{o}_p((NT)^{-\alpha_q}),\ \|\hat{d}^{q,(j)} - d^q\| = \mathrm{o}_p((NT)^{-\beta_q}).$$

If $\min(\alpha_w, \beta_w) + \min(\alpha_q, \beta_q) \geq 1/2$ *and* $\alpha_w, \beta_w, \alpha_q, \beta_q > 0$. *Then,*

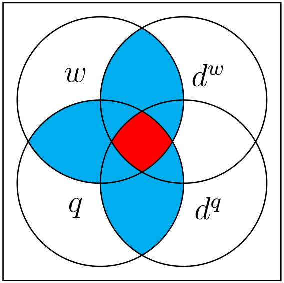
$$\sqrt{NT}(\hat{Z}(\theta) - Z(\theta)) \rightarrow_d \mathcal{N}(0, \mathbb{E}[\psi^2_{\mathit{MDP}}])$$

Robustness Guarantees

Theorem (3-way Double Robustness)

$$\hat{w}^{(j)} \to w^{\dagger}, \quad \hat{d}^{w,(j)} \to d^{w,\dagger}, \quad \hat{q}^{(j)} \to q^{\dagger}, \quad \hat{d}^{q,(j)} \to d^{q,\dagger}$$

Then, $\hat{Z}(\theta) \to_p Z(\theta)$ as long as one of the of following hold: $w^{\dagger} = w, d^{w,\dagger} = d^{w}$*;* $q^{\dagger} = q, d^{q,\dagger} = d^{q}$*; or* $w^{\dagger} = w, q^{\dagger} = q$*.*

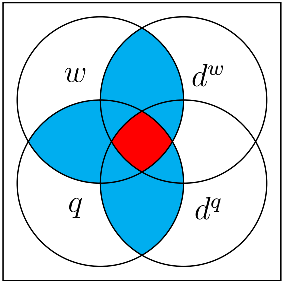


Robustness Guarantees

Theorem (3-way Double Robustness)

$$\hat{w}^{(j)} \to w^\dagger, \quad \hat{d}^{w,(j)} \to d^{w,\dagger}, \quad \hat{q}^{(j)} \to q^\dagger, \quad \hat{d}^{q,(j)} \to d^{q,\dagger}$$

Then, $\hat{Z}(\theta) \to_p Z(\theta)$ as long as one of the of following hold: $w^\dagger = w, d^{w,\dagger} = d^w$; $q^\dagger = q, d^{q,\dagger} = d^q$; *or* $w^\dagger = w, q^\dagger = q$.



Also suggests three new (inefficient) policy gradient methods given by using any good (blue) combination of only *two* nuisances

Efficient Off-Policy Gradient Ascent

- Consider the efficiently-estimated-gradient ascent algorithm:

$$\theta_{i+1} = \mathrm{Proj}_{\Theta}(\theta_i + \alpha_i \hat{Z}(\theta_i))$$

  - Run for K steps and return $\hat{\theta} = \theta_i$ with probability $\propto \alpha_i$

Theorem

Suppose $J(\theta)$ is differentiable and M-smooth, $M < 1/(4\alpha_i)$, ψ is a.s. differentiable with bounded gradient, Θ compact. Then, with probability at least $1 - \delta$:

$$\|Z(\hat{\theta})\|^2 \leq \frac{4(\max_\theta J(\theta) - J(\theta_1))}{K} + \frac{c \log(1/\delta)}{KNT}$$

- If $J(\theta)$ concave: $\mathrm{Regret}(\hat{\theta}) = O_p(\sqrt{\log(NT)/(NT)})$
  - More generally: global optimality of policy gradient ascent (Agarwal et al., 2019; Bhandari and Russo, 2019)

This Talk

1 Introduction

2 Problem Setup

3 Efficiency Bounds

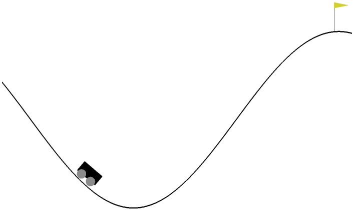
4 Efficient OPE via Double RL

5 Efficient OPG & Policy Learning

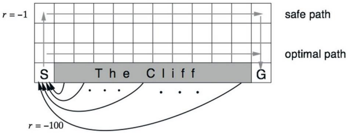
6 Experimental Results

Experiments: OpenAI Gym, Finite Horizon

- Two OpenAI Gym Environments
- Mountain Car



- Cliff Walk



Experiments: OpenAI Gym, Finite Horizon

- Cliff Walking: RMSE (and std errs)

Size	$\hat{\rho}_{\mathrm{IS}}$	$\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$	$\hat{\rho}_{\mathrm{DM}}$	$\hat{\rho}_{\mathrm{MIS}}$	$\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$
500	18.8 (7.67)	3.78(1.14)	2.63 (0.01)	12.8 (4.96)	1.44 (0.29)
1000	7.99 (0.89)	0.28 (0.026)	1.27 (0.002)	5.92 (0.78)	0.22 (0.34)
1500	7.64 (1.63)	0.098 (0.013)	1.01 (0.001)	5.55 (1.10)	0.075 (0.008)

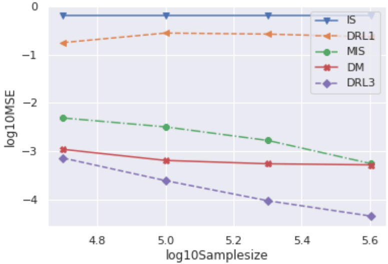
- Mountain Car: RMSE (and std errs)

n	$\hat{\rho}_{\mathrm{IS}}$	$\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_1)}$	$\hat{\rho}_{\mathrm{DM}}$	$\hat{\rho}_{\mathrm{MIS}}$	$\hat{\rho}_{\mathrm{DRL}(\mathcal{M}_2)}$
500	6.85 (0.13)	3.72 (0.08)	4.30 (0.05)	6.82 (0.12)	3.53 (0.12)
1000	4.73 (0.07)	2.12 (0.04)	3.40 (0.008)	4.83 (0.06)	2.07 (0.04)
1500	3.41 (0.04)	1.82 (0.02)	3.30 (0.008)	3.40 (0.05)	1.69 (0.03)

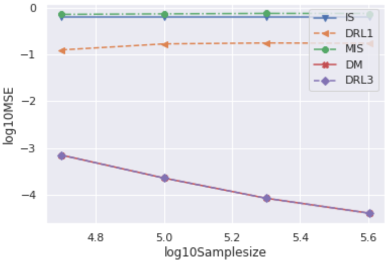
Simulation: Infinite Horizon OPE

- $N = 1$, T varies

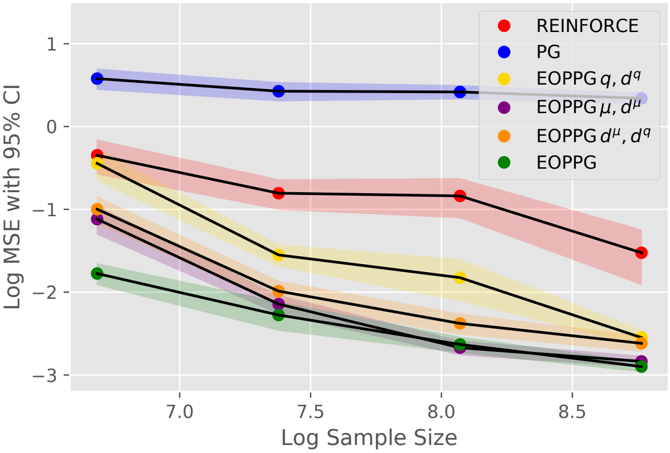
q-model wrong



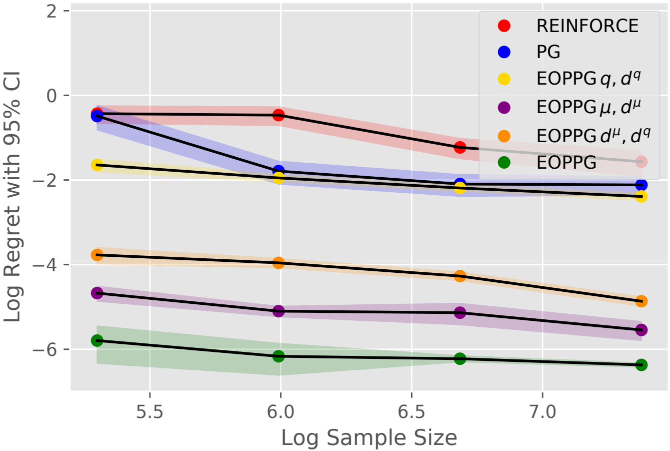
w-model wrong



Simulation: Infinite Horizon OPG (MSE)



Simulation: Infinite Horizon Learning (Regret)



Statistically Efficient Offline Reinforcement Learning

- **Aim**: Overcome fundamental limits in offline RL by leveraging Markovian, time-invariant, and ergodic structure
    - **Theme**: What's *efficient* depends on *structure*
- Contributions
    - Study efficiency limits of OPE/OPG in MDPs for first time
        - Insight into when the curse of horizon bites
        - Problem-dependent phenomenon; not estimator-dependent
    - Provide the *first* efficient OPE/OPG estimator in MDPs
        - Remains efficient even when nuisances estimated at slow rates by blackbox ML
        - Enjoys double robustness guarantees
        - Efficient OPG + gradient ascent leads to learning guarantees

Thank you! 🙏