Causal Inference, Reinforcement Learning, and Estimation under (Markovian) Interference

Dominic DiSanto

Junwei Lu Reading Group - Spring 2024

April 15, 2024

Key Notes

From "Markovian Interference in Experiments" (Farias2022):

- Estimating potential-outcomes/causal inference estimands via solving off-policy evaluation problems
- Cramer-Rao Lower Bound on variance of unbiased, off-policy evaluation estimators
- Construct a MDP-motivated Taylor Expansion of ATE

1. Set-Up
2. ATE/Policy Estimation with Q-functions
   - "Markovian Interference" (Randomization with Interference)

Outline

1. Set-Up

2. ATE/Policy Estimation with Q-functions
   - "Markovian Interference" (Randomization with Interference)

Causal Set-Up

Goal: Estimate ATE $= \mathbb{E}[Y(1) - Y(0)]$

- In "statistical language", we observe $(Y, A, X) \sim \mathbb{P}$
  - Say $A, Y \in \{0, 1\}, X \in \mathbb{R}^d$
  - Some assignment mechanism $A \sim p(\cdot|X)$, or randomization rule $A \sim Ber(p) \perp\!\!\!\perp X$
  - Treatment assignment static A_i or temporal/dynamic/sequential A_{it}

Casting Treatment as an MDP

- Consider spaces of states $s \in S$, actions $a \in A$ under policies $\pi \in \Pi$, and rewards $r(s_t, a_t)$ or losses $\ell(s_t, a_t)$ over time $t \in [T]$
- **Goal:** Infer about an optimal policy π^* (via ATE)

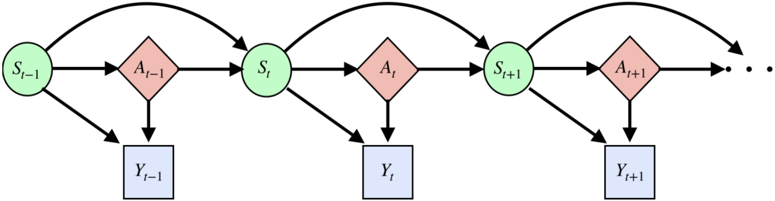

Figure 1: Causal diagram for MDP under settings where treatments depend on current states only. (S_t, A_t, Y_t) represents the state-treatment-outcome triplet. Solid lines represent causal relationships.

Figure: Figure 1 from Shi 2022

RL; V- and Q-Functions

- Transition matrix P^π with stationary distribution ρ_π
- Policy $\pi : S \mapsto A$
  - Control $\pi_0(s) = 0$ and treatment $\pi_1(s) = 1$
- $R^\pi = \lim_{t\to\infty} \frac{1}{T} \sum_{t\in[T]} r(s_t, \pi(s_t))$
  - State-average Reward under policy π
- $V_\pi(s) := \mathbb{E}[\sum_{t=0}^{\infty} r(s_t, a_t) - \lambda^\pi | s_0 = s]$
- $Q_\pi(s, a) := \mathbb{E}[\sum_{t=0}^{\infty} r(s_t, a_t) - \lambda^\pi | s_0 = s, a_0 = a]$

Typical Causal Assumptions

Back to our "statistical" framework:
Under a set of "standard" (untestable) assumptions, we have tools for efficient, DR estimation (double-ML/AIPW, TMLE, etc.)

- $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$ - "Ignorability/Unconfoundedness"
- $p(A|X) \in [\varphi, 1-\varphi]$ - "Overlap/Positivity"
- $Y_i = Y_i(A_i) = A_i Y_i(1) + (1-A_i) Y_i(0)$ - Consistency & "SUTVA" / "Non-interference"

Outline

1. Set-Up
2. ATE/Policy Estimation with Q-functions
   - "Markovian Interference" (Randomization with Interference)

Motivation

"Markovian Interference in Experiments"

"Treatment corresponds to an action which may interfere with state transitions. This form of interference, which we refer to as Markovian"

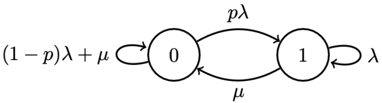

Figure 2: The discrete Markov chain analogous to the continuous-time chain depicted in Fig. 1, for the case $N = 1$. Arrows indicate transition *probabilities*, rather than rates. Without loss of generality, the parameters are normalized so that $\lambda + \mu = 1$.

Figure: Fig. 1 from Farias 2022

Set-Up

- **Goal:** Estimate ATE $= R^{\pi_1} - R^{\pi_0} = \rho_1^T r_1 - \rho_0^T r_0$
  - For stationary distrubions ρ and rewards r
- We observe $\{(a_t, s_t, r(a_t, s_t))\}_{t \in [T]}$ under $\pi_{1/2}$, a simple randomization policy
- We have some treatment π_1 policy that changes the transition probability to $\lambda(p + \delta)$
- Goal is to infer effect of treatment π_1 policy compared to control π_0
  - Phrasing as an off-policy evaluation problem

Why $\hat{\text{ATE}}_{DQ}$?

Estimator	Bias	Variance
Naive	$\Omega(\delta)$	$O(1)$
Off-Policy Evaluation	0	$e^{\Omega(N)}$
Differences-In-Q's (DQ)	$O(\delta^2)$	$O(N)$

Figure: Table 1 - Farias 2022

"Differences in Q's Estimator"

$$\widehat{\text{ATE}}_{DQ} := \frac{1}{|T_1|} \sum_{t \in T_1} \hat{Q}_{\pi_{1/2}}(s_t, a_t) - \frac{1}{|T_0|} \sum_{t \in T_0} \hat{Q}_{\pi_{1/2}}(s_t, a_t)$$

$$\hat{Q}_{\pi_{1/2}} = \min_{\hat{V}, \hat{\lambda}} \sum_{s \in \mathcal{S}} \left(\sum_{t, s_t = s} r(s_t, a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t) \right)^2$$

"Differences in Q's Estimator"

Theorem (Theorem 1 (Bias of DQ))

Assume that for any state $s \in \mathcal{S}, d_{\mathsf{TV}}(p(s,1,\cdot), p(s,0,\cdot)) \leq \delta$. *Then,*

$$\left|\mathrm{ATE} - \mathrm{E}_{\rho_{1/2}}\left[\mathrm{ATE}_{\mathrm{DQ}}\right]\right| \leq C' \left(\frac{1}{1-\lambda}\right)^2 r_{\max} \cdot \delta^2$$

where $r_{\max} := \max_{s,a} |r(s,a)|$ *and* C' *is a constant depending (polynomially) on* $\log(C)$.

"Differences in Q's Estimator"

Theorem (Theorem 2 (Variance and Asymptotic Normality of DQ))

$$\sqrt{T}\left(\hat{\mathrm{ATE}}_{\mathrm{DQ}} - \mathrm{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_{\mathrm{DQ}} \right]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{\mathrm{DQ}}^2\right)$$

$$\sigma_{\mathrm{DQ}} \leq C' \left(\frac{1}{1-\lambda}\right)^{5/2} \log\left(\frac{1}{\rho_{\min}}\right) r_{\max}$$

where $\rho_{\min} := \min_{s \in S} \rho_{1/2}(s)$ *and* C' *is a constant depending (polynomially) on* $\log(C)$.

Off-Policy Evaluation

Theorem (Theorem 3 from Farias 2022 (Variance Lower Bound for Unbiased Estimators))

Assume we are given a dataset $\{(s_t, a_t, r(s_t, a_t)) : t = 0, \dots, T\}$ *generated under the experimentation policy* $\pi_{1/2}$, *with* s_0 *distributed according to* $\rho_{1/2}$. *Then for any unbiased estimator* $\hat{\tau}$ *of ATE, we have that*

$$T \cdot \mathrm{Var}(\hat{\tau}) \geq$$

$$2 \sum_s \frac{\rho_1(s)^2}{\rho_{1/2}(s)} \sum_{s'} p\left(s, 1, s'\right) \left(V_{\pi_1}\left(s'\right) - V_{\pi_1}(s) + r(s, 1) - \lambda^{\pi_1}\right)^2$$

$$+ 2 \sum_s \frac{\rho_0(s)^2}{\rho_{1/2}(s)} \sum_{s'} p\left(s, 0, s'\right) \left(V_{\pi_0}\left(s'\right) - V_{\pi_0}(s) + r(s, 0) - \lambda^{\pi_0}\right)^2 \triangleq \sigma_{off}^2$$

Off-Policy Evaluation

Theorem (Theorem 4)

For any $0 < \delta \leq \frac{1}{5}$, there exists a class of MDPs parameterized by $n \in \mathbb{N}$, where n is the number of states, such that $\frac{\sigma_{DQ}}{\sigma_{off}} = O(\frac{n}{c^n}), c > 1$. Furthermore, $|(ATE - \mathbb{E}\hat{ATE}_{DQ})/ATE| \leq \delta$

Estimator	Bias	Variance
Naive	$\Omega(\delta)$	$O(1)$
Off-Policy Evaluation	0	$e^{\Omega(N)}$
Differences-In-Q's (DQ)	$O(\delta^2)$	$O(N)$

$\hat{\text{ATE}}_{DQ}$ as bias-correction

Observe Lemma 2 in Farias2022 (pg 14-5)

Questions

Questions of Confusion

- How does this estimator/model include/account for interference? μ term (feedback/relapse mechanism) is not alone to account for some interference
  - Simulations do so explicitly but nothing in the crafting of this estimator seems
- How general is Theorem 4, the comparison of σ_{DQ}/σ_{off} (and analytically, how does this $e^{|S|}$ term arrive?)

Questions of Opportunity

- How does $\hat{\text{ATE}}_{DQ}$ translate to (observed) policies without randomization?
- How does $\hat{\text{ATE}}_{DQ}$ scale wrt the action space $|A|$?

Compiling Resources

- https://crl.causalai.net/
- Junzhe Zhang - https://junzhez.com/
- Jiang & Li (2016) - "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning"
  - https://arxiv.org/pdf/1511.03722.pdf