Off-Policy Evaluation in Partially Observed Markov Decision Processes

Stefan Wager
Stanford University

Workshop on Synthetic Control Methods
Princeton Center for Statistics and Machine Learning
3 June 2022

Joint work with Yuchen Hu
Synthetic control methods (Abadie & al.) provide a **flexible suite of solutions** that generalize difference-in-differences type analyses.
Synthetic Control Methods

Synthetic control methods (Abadie & al.) provide a **flexible suite of solutions** that generalize difference-in-differences type analyses. They help address failures in **parallel trends**:

- By balancing out **interactive fixed effects** (Abadie, Diamond & Hainmueller, 2010; Arkhangelsky, Athey, Hirshberg, Imbens & Wager, 2021).
- By **controlling for observed history** (Ben-Michael, Feller & Rothstein, 2021).
- By **reducing variance** in randomized trials (Bottmer, Imbens, Spiess & Warnick, 2021).

They represent **long-term treatment dynamics** via the SC plot.
Synthetic Control Methods

Synthetic control methods (Abadie & al.) provide a flexible suite of solutions that generalize difference-in-differences type analyses. They help address failures in parallel trends:

- By controlling for observed history (Ben-Michael, Feller & Rothstein, 2021).
- By reducing variance in randomized trials (Bottmer, Imbens, Spiess & Warnick, 2021).

They represent long-term treatment dynamics via the SC plot.

**Question:** What are good ways to understand long-term treatment dynamics in generic panels?
Off-Policy Evaluation

We follow $i = 1, \ldots, n$ people for $t = 1, \ldots, T$ time periods. For each $(i, t)$ pairs, we observe:

- A state variable $X_{it} \in \mathcal{X},$
- An action $W_{it} \in \{1, \ldots, A\},$
- A reward $Y_{it} \in \mathbb{R}.$

We’re interested in a target policy $\pi : \mathcal{X} \rightarrow [0, 1]^A$ that, given state $X_{it}$, picks action $a$ with probability $\pi_a(X_{it}).$

For identification, we make a sequential ignorability assumption:

- There is a behavior policy $e : \mathcal{X} \rightarrow [0, 1]^A$ such that, given state $X_{it},$ we observe action $a$ with probability $e_a(X_{it})$ (i.e., we have a micro-randomized trial).

**Question:** What would the expected average outcome have been had we chosen actions according to $\pi$?
In a micro-randomized trial, how do different models of treatment dynamics affect the difficulty of off-policy evaluation?

Currently, there are two dominant models considered in the literature on off-policy evaluation:


- **Option 2: Assume an MDP.** Past actions don’t matter conditionally on current state. [Antos et al., 2008, Kallus and Uehara, 2020, Liao et al., 2021, Luckett et al., 2019].

How do these assumptions affect the difficulty of off-policy learning? Can we consider an interesting class of models in between?
The value of the **Markov Decision Process assumption** for off-policy evaluation is discussed at length by Kallus and Uehara (2020). The short story is:

- Without modeling assumptions, we have a **curse of dimension**. As the horizon $T$ gets large, the difficulty of the problem blows up.

- In an MDP, **long trajectories help**. Under stationary, we can estimate long-run average rewards at rate $1/\sqrt{nT}$ [Liao et al., 2021], i.e., we get a **parametric rate** in the number of observations.

Can we consider an interesting class of models in between?
Assume a **POMDP** (Partially Observed MDP): Nature is Markovian, but we don’t observe what we’d need to fit an MDP.
Hidden States

In many applications, POMDP may be a much more credible assumption than MDP.

- Consider, e.g., mobile health applications where a patient’s mood affects treatment response.

POMDPs are widely used for planning under uncertainty [Monahan, 1982, Smallwood and Sondik, 1973], but not as models for off-policy evaluation.

One recent exception: Work in POMDPs where the hidden state is an unobserved confounder! This approach is pursued in Tennenholtz et al. [2020], and more recently Nair and Jiang [2021], and Bennett and Kallus [2021].
Main Results

**Upper bounds.** The POMDP assumption implies **mixing**. We use **partial-history importance weighting** to derive upper bounds that depend on the mixing time.

**Lower bounds.** We show there exist simple instances where our upper bounds based on mixing are **nearly sharp**.

Take-home points:

- Off-policy evaluation in POMDP is strictly easier than in the model-free setting, in that **longer trajectories help**.
- **Minimax rates** of convergence are of the form $1/\text{poly}(nT)$.
- But cannot achieve **parametric rates** $1/\sqrt{nT}$ like in an MDP.

More in paper: CLTs, and adaptivity via Lepski’s method, etc.
Upper Bounds

We rely on two main assumptions:

- **Overlap:** We have $\pi_a(x) \leq e^{\xi \pi} e_a(x)$ for all $a, x$.
- **Mixing:** The POMDP mixes in time $t_\pi$.

**Assumption 2.** Let $\pi$ be any policy that maps current observed state to action probabilities such that, under Model 3,

$$\pi : \mathcal{X} \to [0, 1]^A, \quad \mathbb{P}_\pi [W_{i,t} = a \mid X_{i,1}, H_{i,t}, Y_{i,1}, \ldots, X_{i,t}, H_{i,t}] = \pi(X_{i,t}). \quad (2)$$

Let $P^\pi_i$ denote the state transition operator on $(X_{i,t}, H_{i,t})$ associated with $\pi$. We assume that, for all considered policies $\pi$, there is a mixing time $t^\pi_i$ such that

$$\|f'P^\pi_i - fP^\pi_i\|_{TV} \leq \exp(-1/t^\pi_i) \|f' - f\|_{TV}, \quad (3)$$

for any pair of distributions $f$ and $f'$ on $(X_{i,t}, H_{i,t})$. 
Upper Bounds

We use **partial-history importance weighting** for estimation,

\[
\hat{V}(\pi; k) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T-k} \sum_{t=k+1}^{T} \left( \prod_{s=0}^{k} \frac{\pi W_{i,t-s}(X_{i,t-s})}{\pi W_{i,t-s}(X_{i,t-s})} \right) Y_{i,t},
\]

where \( k \) is a tuning parameter (the relevant history length).

**Theorem.** Under our assumptions, for a well chosen sequence \( k \) and supposing that \( T \) is not too short and that the conditional first and second moments of \( Y_{i,t} \) are uniformly bounded,

\[
\mathbb{E} \left[ \left( \hat{V}(\pi; k) - V(\pi) \right)^2 \right]^{\frac{1}{2}} \leq O \left( (nT)^{-\frac{1}{t_0 \zeta \pi + 2}} \right),
\]

where \( V(\pi) \) is the average stationary reward under \( \pi \).
Theorem. There exists a set of instances satisfying the above mixing and overlap condition and with bounded conditional first and second moments for which:

$$\inf_{\hat{V}} \max_{\text{instance}} \mathbb{E} \left[ \left( \hat{V} - V(\pi) \right)^2 \right]^{1/2} = \Omega \left( \max \left\{ T^{-\frac{1}{t_0 \zeta \pi + 1}}, T^{-1/2} \right\} \right).$$

\[\implies\] with $n = 1$, the minimax error for off-policy evaluation under our mixing and overlap conditions satisfies:

$$\max \left\{ T^{-\frac{1}{t_0 \zeta \pi + 1}}, T^{-1/2} \right\} \lesssim R_{\text{minimax}} \lesssim T^{-\frac{1}{t_0 \zeta \pi + 2}}.$$
Lower Bounds

Proof via **LeCam’s two-point method**:

- Pick **two candidate instances**, $I_1$ and $I_2$, that satisfy the assumptions for our upper bound.
- The instances $I_1$ and $I_2$ should be **hard to tell apart** under the behavior (or observation) policy.
- The instances $I_1$ and $I_2$ should yield **meaningfully different rewards** under the target policy $\pi$.

Main idea: If you can’t tell $I_1$ and $I_2$ apart, you can’t uniformly beat random guessing (and incur meaningful errors).
Lower Bounds

We consider the following instance (with no observed state $X_{it}$):

- The rewards depend on hidden state:
  \[ Y_{it} \mid H_{it} \sim \mathcal{N} \left( \pm \Delta 1(\{H_{it} = h_Q\}), \sigma^2 \right) . \]

- The target policy is deterministic \[ \mathbb{P} \left[ W_{it} = 1 \right] = 1 . \]

- The behavior policy is random \[ \mathbb{P} \left[ W_{it} = 1 \right] = e^{-\zeta \pi} . \]

The behavior policy rarely visits $h_Q$ so it’s hard to learn the sign of $\Delta$, but $\pi$ spends non-trivial time there, so the sign matters.
Numerical Example

Numerical result motivated by a mobile health app that monitors the blood glucose level of type 1 diabetic patients.

- Task: Evaluate policies for timing of insulin injections based on patient blood glucose, physical activity, and dietary intake with the goal of controlling future blood glucose as close as possible to the optimal range.

- Data: Simulator from Luckett et al. [2019], based on data from Maahs et al. [2012].

- The original simulator is an MDP, but we hide some states.
Numerical Example

Figure 6: MSE as a function of $k$ under different horizon length $T$. The lightest orange corresponds to the case with $T = 200$, with a gradient to dark red representing $T$ increases from $T = 200$ to $T = 1000$ gradually.
Off-policy evaluation is of central importance to many applications in medicine, economics, etc.:

- In general, off-policy evaluation is essentially impossible without assumptions (at least with long horizons).
- Interest in applications where precise simulators are not available, and parametric models may not be credible.
- What are useful yet credible modeling assumptions that can help?
Thanks!