# State-Action Balancing in Causal Inference


Qiming Du

joint work with Gérard Biau, François Petit and Raphël Porcher

Laboratoire de Probabilités, Statistique et Modélisation - UMR 8001
Sorbonne Université


(Please visit https://mgimm.github.io/presentation for additional information.)

January 13, 2022 Online

# What we will talk about?

- Basic concepts of Causal Inference
  - Policy evaluation.
  - Unconfoundedness and Markov structure.
- Static Causal Inference
  - Inverse Probability Weighting (IPW) estimator (classical estimators).
  - Static balancing (our methods).
  - Doubly robust estimators.
- Dynamic Causal Inference
  - Measure flows in the change of policy.
  - Dynamical recursive balancing.

# Notation and conventions

- $\xi(\mathrm{d}x)$ (measure), $f$ (test function), $\xi(f) = \int f(x)\xi(\mathrm{d}x)$.
- $M(x, \mathrm{d}y)$ (transition kernel), $M(f)(\cdot) = \int M(x, \mathrm{d}y)f(y)$, and $\mu M(\mathrm{d}y) = \int \mu(\mathrm{d}x)M(x, \mathrm{d}y)$.
- A finite measure is given identified by the values tested on all the bounded measurable test function $\xi(f)$.
- A transition kernel is a state-indexed family of measure.
- For any random variables $X$ and $Y$, there exists a Markov transition kernel $M$ that connects their distributions.

# Conceptual example: Covid and Vaccination

**Step 1: Static Causal Inference without personalized treatment**

Question: To vaccinate or not?

Model: Potential Outcomes framework (Rubin, 1974):

- $X$: Population covariate (age, sex, blood-type, etc).
- $A \in \{0, 1\}$: Action assignment indicator (1:vaccinated vs. 0:not-vaccinated)
- $(Y(0), Y(1))$: Effects associated to vaccination (e.g., Infection, side-effects, etc).

Goal: Estimate the average potential outcomes $\mathbb{E}[Y(1)]$ and/or $\mathbb{E}[Y(0)]$.
$\iff$ Policy evaluation (Everyone vaccinates (or not)).

# How to collect data?

Dataset: $\mathcal{D}_n = \{(X^{(i)}, A^{(i)}, Y^{(i)}(A^{(i)})) : 1 \le i \le n\}$.

- Randomized study: Assume $A \perp\!\!\!\perp (Y(0), Y(1))$.
- Observational study (unconfoundedness): Assume $A \perp\!\!\!\perp (Y(0), Y(1))$ given $X$.

In both cases, we have

$$\mathbb{E}\left[Y(A)\frac{\mathbf{1}_{\{A=1\}}}{\mathbb{P}(A=1 \mid X)}\right] = \mathbb{E}\left[\mathbb{E}\left[Y(A)\frac{\mathbf{1}_{\{A=1\}}}{\mathbb{P}(A=1 \mid X)} \,\middle|\, X\right]\right] = \mathbb{E}\left[Y(1)\right].$$

Inverse Probability Weghting (IPW) estimator:

$$\text{IPWE} = \frac{1}{n}\sum_{i=1}^{n} Y^{(i)}(A^{(i)})\frac{\mathbf{1}_{\{A^{(i)}=1\}}}{\hat{\mathbb{P}}\left(A=1 \,\middle|\, X=X^{(i)}\right)}.$$

The nuisance estimator $\hat{\mathbb{P}}\left(A=1 \,\middle|\, X=X^{(i)}\right)$ is estimated through a separated supervised learning (classification) problem.

# Alternative approach: G-computation

Denote by $\mu_1(\cdot) = \mathbb{E}\left[Y(1) \mid X = \cdot\right]$, then

$$\mathbb{E}\left[\mu_1(X)\right] = \mathbb{E}\left[Y(1)\right].$$

G-computation estimator

$$\mathrm{GE} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(X^{(i)}),$$

where the nuisance estimator $\hat{\mu}_1$ can also be obtained by a separate supervised learning problem.

Can we combine these two ideas? yes!
Doubly robust estimator:

$$\mathrm{DRE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y^{(i)}(A^{(i)}) - \hat{\mu}_1(X^{(i)})\right)\frac{\mathbf{1}_{\{A^{(i)}=1\}}}{\hat{\mathbb{P}}\left(A = 1 \,\middle|\, X = X^{(i)}\right)} + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X^{(i)}).$$

Why interesting?

- Faster convergence rate (product of the two nuisance estimators).
- Semiparametric efficiency (optimal asymptotic variance over all parametric models) at optimal rate ($\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$).

What can we improve? The IPW part!

# A closer look at IPW

Q: Why it works?

A: It balances two subpopulations, i.e.,

$$X \quad \text{and} \quad X(1) \text{ (people vaccinated)},$$

through re-weighting. The associated weight function is

$$\eta(\cdot) = \frac{\mathbb{E}[A]}{\mathbb{P}\left(A = 1 \mid X = \cdot\right)}.$$

Re-weighting transformation:

$$\Psi_\eta(\xi) : \xi \mapsto \xi(\eta \times \cdot)$$

We have

$$\Psi_\eta(\xi_1) = \xi,$$

where $\xi_1$ (resp. $\xi$) is the probability measure of $X(1)$ (resp. $X$).

# Reformulation of IPWE

We have

$$
\begin{aligned}
\mathrm{IPWE} =& \frac{1}{n}\sum_{i=1}^{n} Y^{(i)}(A^{(i)}) \frac{\mathbf{1}_{\{A^{(i)}=1\}}}{\hat{\mathbb{P}}\left(A=1 \,\middle|\, X = X^{(i)}\right)} \\
=& \frac{1}{N_1}\sum_{i=1}^{N_1} \hat{\eta}(X(1)^{(i)}) Y^{(i)}(1),
\end{aligned}
$$

where $\hat{\eta}(X(1)^{(i)}) = \frac{N_1/n}{\hat{\mathbb{P}}(A=1 \mid X=\cdot)}$, which is an empirical version of $\eta$.

Natural idea to improve IPW/get rid of the inverse manipulation:
**Directly estimate the weight function that corrects the difference between two measures!**

Are we able to generalize this simple idea to more general cases (e.g., with more complex policy/action assignment)? Yes!

# General policy evaluation

**Step 2: Policy evaluation with personalized treatment.**
Different people receive different treatment.

Policy is now modeled by a transition kernel $\pi(x, \mathrm{d}a)$ from state space to action space.

- Sampling policy $\pi(x, \mathrm{d}a)$: policy that generates the data set $\mathcal{D}_n$.
- Target policy $\mathring{\pi}(x, \mathrm{d}a)$: The policy to be evaluated.

Case 1: Finite-valued action (e.g., whether to vaccinate, Moderna or Pfizer?).

No need to change the framework. For example, one may replace all the $A = 1$ by $A = \mathring{A}^{(i)}$ where $\mathring{A}^{(i)} \sim \mathring{\pi}(X^{(i)}, \cdot)$.

Case 2: General cases (e.g., continuous-valued policy when considering dosage, expenses, etc.)

State-Action Markov reformulation.

# Markov structure of causal dynamics

- State space: $\mathcal{X}$
- Action space: $\mathcal{A}$
- State-action space: $\mathcal{X}^\natural = \mathcal{X} \times \mathcal{A}$.
- $\pi^\natural = \mathrm{id}_\mathcal{X} \times \pi$.
- State-action variable: $X^\natural = (X, A)$.
- Goal: estimate $\mathcal{V}^{\mathring{\pi}} = \mathbb{E}[\mathring{Y}] = \mathbb{E}[r(\mathring{Z})]$ .

$$\begin{tikzcd}
 & A & \\
X \arrow[ur, "\pi"] \arrow[r, "\pi^\natural"] & X^\natural \arrow[u, "\mathrm{Proj}(\cdot|\mathcal{A})"'] \arrow[r, "\textcolor{red}{M}"] & Z
\end{tikzcd}
\qquad
\begin{tikzcd}
 & \mathring{A} & \\
\mathring{X} \arrow[ur, "\mathring{\pi}"] \arrow[r, "\mathring{\pi}^\natural"] & \mathring{X}^\natural \arrow[u, "\mathrm{Proj}(\cdot|\mathcal{A})"'] \arrow[r, "\textcolor{red}{M}"] & \mathring{Z}
\end{tikzcd}$$

Figure: Markov structure of static causal model

When $\#\mathcal{A} < \infty$, the dynamic is equivalent to the potential outcomes framework with unconfoundedness assumption.

# Covariate shifts?

Denote by $\xi$ the population distribution ($X$) of sampling policy. Denote by $\mathring{\xi}$ the population distribution ($\mathring{X}$) of target policy. It is possible that $\xi \neq \mathring{\xi}$.
(Lacking of external validation (Pearl et al., 2014).)

What are we balancing?

The (empirical) state-action distribution of $X^\natural$ and $\mathring{X}^\natural$, denoted respectively by

$$\xi^{\pi^\natural} \ (\xi_n^{\pi^\natural}) \quad \text{and} \quad \mathring{\xi}^{\mathring{\pi}^\natural} \ (\mathring{\xi}_m^{\mathring{\pi}^\natural}).$$

What are we collecting?

- State-action variables under sampling policy: $\{(X^{(i)}, A^{(i)}) : 1 \leq i \leq n\}$
- State variables of target policy: $\{\mathring{X}^{(i)} : 1 \leq i \leq m\}$.
- Causal effects under sampling policy: $\{r(Z^{(i)}) : 1 \leq i \leq n\}$.

# Estimators

Direct estimator:

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\eta}^{\natural}(X^{\natural(i)}) r(Z^{(i)}),$$

Doubly robust estimator:

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\eta}^{\natural}(X^{\natural(i)})(r(Z^{(i)}) - \hat{r}^{\natural(i)}(X^{\natural(i)})) + \frac{1}{m}\sum_{j=1}^{m} \mathring{\pi}^{\natural}(\hat{r}^{\natural(j)})(\mathring{X}^{(j)}),$$

where $\hat{\eta}^{\natural}$ is the estimated weight function $\mathring{\eta}^{\natural}$ that corrects the difference between the two state-action distributions, i.e., $\Psi_{\mathring{\eta}^{\natural}}(\xi^{\pi^{\natural}}) = \mathring{\xi}^{\mathring{\pi}^{\natural}}$; $\hat{r}^{\natural(j)}$ is the estimated conditional expectation function $\mathbb{E}\left[r(Z) \,\middle|\, X^{\natural} = \cdot\right]$ (or simply $r^{\natural} = M(r)$) by a separated regression.

How?

Generalized IPW (through density ratio estimation, see, e.g., Sugiyama et al. (2012)).
Idea: Constructing an artificial 0-1 classification problem and solve it to construct weight function estimator.
or
Balancing!

# Teaser of balancing

**Direct estimator and worst-case-error interpretation**

Idea: $r(Z^{(i)})$ can be regarded as a noisy version of $r^\natural(X^{\natural(i)})$. Hence, one considers to minimize the worst-case-error, i.e.,

$$\hat{H} = \operatorname*{arg\,min}_{\eta^\natural \in H} \sup_{\gamma \in \Gamma} \left| \Psi_{\eta^\natural}(\xi_n^{\pi^\natural})(\gamma) - \mathring{\xi}_{mN}^{\dot{\pi}^\natural}(\gamma) \right|,$$

It is well-known that that sup-term is called Integral Probability Metric.

$$\hat{H} = \operatorname*{arg\,min}_{\eta^\natural \in H} \left( \mathrm{IPM}_\Gamma \left( \Psi_{\eta^\natural}(\xi_n^{\pi^\natural}), \mathring{\xi}_{mN}^{\dot{\pi}^\natural} \right)^2 + \lambda \left\| \eta^\natural \right\|_{L^2(\xi_n^{\pi^\natural})}^2 \right),$$

(conditional Bias-Variance decomposition)

In this case, we may choose $H = L^2(\xi^{\pi^\natural})$ or $\lambda U$ with $U$ the unit ball of $L^2(\xi^{\pi^\natural})$.

Well-specification: $r^\natural \in \Gamma$.

# Balancing for the DE

- OT balancing without $L^2$ penalty: (Reygner and Touboul, 2020).
- MMD balancing with/without $L^2$ penalty: (Kallus, 2020).
- OT balancing with $L^2$ penalty: new.
- Neural Network balancing with/without $L^2$ penalty: new.
- ...

**Theorem (Informal)**

*When well-specified, the error of the DE is controlled by the sampling complexity (i.e., $\mathrm{IPM}_\Gamma(\Psi_{\mathring{\eta}^\natural}(\xi_n^{\pi^\natural}), \Psi_{\mathring{\eta}^\natural}(\xi^{\pi^\natural})) + \mathrm{IPM}_\Gamma(\mathring{\xi}_{mN}^{\mathring{\pi}^\natural}, \mathring{\xi}^{\mathring{\pi}^\natural})$) of the chosen IPM.*

However, there is in general no reason that $\hat{\eta}^\natural$ will converge to the ideal weight function $\mathring{\eta}^\natural$ in an $L^2$ sense, which is required by the DRE.

# Why $L^2$ convergence matters?

Denote by $\mathcal{V}^{\mathring{\pi}}_{\text{ODRE}}$ the oracle version of the DRE, namely, by replacing the nuisance estimatros $\hat{\eta}^{\natural}$ and $\hat{r}^{\natural}$ by their oracle/ideal counterparts $\mathring{\eta}^{\natural}$ and $r^{\natural}$.

## Theorem (Informal)

*Under mild assumptions, we have*

$$\left|\mathcal{V}^{\mathring{\pi}}_{\text{ODRE}} - \hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}}\right| \leq C \left\|\hat{\eta}^{\natural} - \mathring{\eta}^{\natural}\right\|_{L^2(\xi_n^{\pi^{\natural}})} \left\|\hat{r}^{\natural} - r^{\natural}\right\|_{L^2(\xi_n^{\pi^{\natural}})} + o_{\mathbb{P}}\left(\sqrt{\frac{n+m}{nm}}\right).$$

*In addition, when no covariate shifts are involved,* $\mathcal{V}^{\mathring{\pi}}_{\text{ODRE}}$ *achieves semiparametric efficiency.*

To understand the $L^2$ behavior of the weight function estimation, we need a little bit maths...

# Riesz representable measure space

- Source measure: $\xi^{\pi^\natural}$.
- Target measure: $\mathring{\xi}^{\mathring{\pi}^\natural}$.
- Riesz representable measure space: $\Xi(\xi^{\pi^\natural}) := \{\Psi_{\eta^\natural}(\xi^{\pi^\natural}) : \eta^\natural \in L^2(\xi^{\pi^\natural})\}$

## Proposition

*Let $\xi$ be a positive finite measure on $\mathcal{X}$. Denote by $U$ the unit ball in $L^2(\xi)$. We have the following isometric isomorphism between the metric spaces $L^2(\xi)$ and $\Xi(\xi)$:*

$$(L^2(\xi), \|\cdot - \cdot\|_{L^2(\xi)}) \underset{\frac{\mathrm{d}\cdot}{\mathrm{d}\xi}}{\overset{\Psi_\cdot(\xi)}{\rightleftarrows}} (\Xi(\xi), \mathrm{IPM}_U(\cdot,\cdot))$$

Heuristic:
On a compact state-action space, what if we use a RELU network to approximate the $L^2$ unit ball?

# One step further

Q: Do we really need an isometry in order to have an $L^2$ convergence?
A: Not really!
Notation: $\partial H = \{\eta - \eta' : \eta, \eta' \in H\}$.

**Lemma**

*If $\mathring{\eta}^\natural \in H$ and there exists $\alpha > 0$ and $\beta > 0$ such that $\alpha(\partial H) \cap \beta U \subset \Gamma$, then we have almost surely*

$$\forall \hat{\eta}^\natural \in \hat{H}, \quad \left\| \hat{\eta}^\natural - \mathring{\eta}^\natural \right\|_{L^2(\xi_n^{\pi^\natural})} \leq \frac{2}{\min(\alpha, \beta)} \left( \mathrm{IPM}_\Gamma \left( \Psi_{\mathring{\eta}^\natural}(\xi^{\pi^\natural}), \Psi_{\mathring{\eta}^\natural}(\xi_n^{\pi^\natural}) \right) + \mathrm{IPM}_\Gamma \left( \mathring{\xi}^{\mathring{\pi}^\natural}, \mathring{\xi}_{mN}^{\mathring{\pi}^\natural} \right) \right).$$

The construction can be regarded as a dual version (in a Fenchel sense) of Chernozhukov et al. (2020).

Take home message:
when $\Gamma$ is rich enough (that contains at least $\alpha(\partial H) \cap \beta U$), the $L^2$-error of weight function estimation is controlled by the sampling complexity of the chosen IPM. The error analysis of the DRE is therefore transparent.

# Practical considerations

**Pipeline:**

1. Fix $H$ to ensure that $\mathring{\eta}^{\natural} \in H$.
2. Construct $\Gamma$ such that $\alpha(\partial H) \cap \beta U$ (when $H$ is rich enough, it is in general only to ensure that $\Gamma = H \cap \beta U$, i.e., to implement an $L^2$-regularization)
3. Solve the adversarial optimization, i.e., arg min max-optimization.

- $H$ : Intersection of RKHS ball and $L^2$-ball; $\Gamma$: RKHS ball. (quadratic programming: explicitely solvable/ gradient descend based method)
- $H$: RELU/Groupsort network + $L^2$ regularization; $\Gamma$: RELU/Groupsort network (with nodes number doubled at each layer) + $L^2$ regularization.
- (possible) $H$ : Intersection of Lipschitz ball, $L^2$-ball, and relative entropy reguralization (Sinkhorn); $\Gamma$: Lipschitz ball with relative entropy reguralization . (gradient descend based method)
- ...

No conservation of mass:
One may use it for tuning, or consider implementing an additional regularization.

# Comparison between the DE and the DRE

DE:

- Requires that $r^\natural \in \Gamma$.
- Allows to optimized in $H = L^2(\xi^{\pi^\natural})$, i.e., $n$-value optimization.
- The error of the DE is controlled by the sampling complexity of the chosen IPM.

DRE:

- Requires that $\mathring{\eta}^\natural \in H$.
- Requires to model properly the candidate space $H$.
- The $L^2$-error of the weight function estimation is controlled by the sampling complexity of the chosen IPM.

Q: When the optimal (parametric) rate is achieved, the DRE is always better than the DE?
A: No, their asymptotic variances are not comparable in general (see, e.g., Kallus and Uehara (2020)).

# Dynamical Causal Inference

**Step 3: Reinforcement Learning/Dynamical Treatment Regimes**

We consider the 3-vaccination and their causal effects estimation.

Causal dynamics:

$$
\begin{tikzcd}
 & A_0 & & & & A_t & & \\
X_0 \arrow[r, "\pi_0^\natural"] \arrow[ur, "\pi_0"] & X_0^\natural \arrow[r, "M_1"] \arrow[u, "\mathrm{Proj}(\cdot|\mathcal{A}_0)"'] & X_1 \arrow[r, "\pi_1^\natural"] & \cdots \arrow[r, "M_{t-1}"] & X_t \arrow[r, "\pi_t^\natural"] \arrow[ur, "\pi_t"] & X_t^\natural \arrow[r, "M_{t+1}"] \arrow[u, "\mathrm{Proj}(\cdot|\mathcal{A}_t)"'] & X_{t+1} \arrow[r] & \cdots
\end{tikzcd}
$$

Figure: Markov structure of causal dynamics

Goal: Estimate

$$\mathcal{V}^{\mathring{\pi}} = \mathbb{E}\left[\sum_{t=0}^{T} r_t(\mathring{X}_{t+1})\right],$$

# Models

What do we collect?

- State-Action trajectories under sampling policy $\pi$ (a sequence of Markov kernels): $\{(Z_t^{(i)}, A_t^{(i)}; 0 \leq t \leq T) : 1 \leq i \leq n\}$.
- Target initial state covariates: $\{\mathring{Z}_0^{(i)} : 1 \leq i \leq m\}$.
- Observed causal effects at each time step: $\{r_t(Z_{t+1}^{(i)}) : 0 \leq t \leq T, 1 \leq i \leq n\}$.

Identification of Markov Structure?

One may simply take $X_t^{(i)} = Z_t^{(i)}$.
or
One may consider $X_t^{(i)} = (Z_s^{(i)}; 0 \leq s \leq t)$.
or even $X_t^{(i)} = ((Z_s^{(i)}; 0 \leq s \leq t), (A_s^{(i)}; 0 \leq s \leq t-1))$.

Less fluctuation vs. higher dimension (harder to estimate).
For simplicity, we choose $X_t^{(i)} = Z_t^{(i)}$.

# Double semigroup structure

If we let respectively

$$M_t^{\pi} = \pi_{t-1}^{\natural} \circ M_t \quad \text{and} \quad M_t^{\pi^{\natural}} = M_t \circ \pi_t^{\natural},$$

one gets a double partial semigroup structure:

$$\begin{tikzcd}
X_0 \arrow[r, "\pi_0^{\natural}"] \arrow[rr, bend right, "M_1^{\pi}"'] & X_0^{\natural} \arrow[r, "M_1"] \arrow[rr, bend left, "M_1^{\pi^{\natural}}"] & X_1 \arrow[r, "\pi_1^{\natural}"] & X_1^{\natural} \arrow[r] & \cdots \arrow[r] & X_{t-1}^{\natural} \arrow[r, "M_t"] \arrow[rr, bend left, "M_t^{\pi^{\natural}}"] & X_t \arrow[r, "\pi_t^{\natural}"] \arrow[rr, bend right, "M_{t+1}^{\pi}"'] & X_t^{\natural} \arrow[r, "M_{t+1}"] & X_{t+1} \arrow[r] & \cdots
\end{tikzcd}$$

Figure: Double semigroups in causal dynamics

We consider two partial semigroups defined respectively by

$$\forall s > t, \quad M_{t,s}^{\pi} = M_{t+1}^{\pi} \circ \cdots \circ M_s^{\pi}, \quad \text{with} \quad M_{t,t}^{\pi} = \mathrm{id}_{\mathcal{X}_t},$$

and

$$\forall s > t, \quad M_{t,s}^{\pi^{\natural}} = M_{t+1}^{\pi^{\natural}} \circ \cdots \circ M_s^{\pi^{\natural}}, \quad \text{with} \quad M_{t,t}^{\pi^{\natural}} = \mathrm{id}_{\mathcal{X}_t^{\natural}}.$$

# State/State-action terminal measures

Considering the initial distributions $\xi_0^{\pi} = \xi_0$ and $\xi_0^{\pi^\natural} = \xi_0 \circ \pi_0^\natural$, we define the terminal measures $\xi_t^{\pi}$ and $\xi_t^{\pi^\natural}$ respectively by

$$\xi_t^{\pi} = \xi_0^{\pi} M_{0,t}^{\pi} \quad \text{and} \quad \xi_t^{\pi^\natural} = \xi_0^{\pi^\natural} M_{0,t}^{\pi^\natural}.$$

The objective re-writes

$$\mathcal{V}^{\mathring{\pi}} = \sum_{t=0}^{T} \mathring{\xi}_{t+1}^{\mathring{\pi}}(r_t) = \sum_{t=0}^{T} \mathring{\xi}_t^{\mathring{\pi}^\natural}(r_t^\natural).$$

# Measure flows in the change of policy

$$
\begin{tikzcd}
\xi_0^{\pi} \arrow[r, "\pi_0^\natural"] & \xi_0^{\pi^\natural} \arrow[r, "M_1"] \arrow[d, "\Psi_{\mathring{\eta}_0^\natural}"] & \xi_1^{\pi} \arrow[r] \arrow[d, "\Psi_{\mathring{\eta}_1}"] & \cdots \arrow[r] & \xi_t^{\pi^\natural} \arrow[r, "M_{t+1}"] \arrow[d, "\Psi_{\mathring{\eta}_t^\natural}"] & \xi_{t+1}^{\pi} \arrow[r, "\pi_{t+1}^\natural"] \arrow[d, "\Psi_{\mathring{\eta}_{t+1}}"] & \xi_{t+1}^{\pi^\natural} \arrow[r] \arrow[d, "\Psi_{\mathring{\eta}_{t+1}^\natural}"] & \cdots \\
\mathring{\xi}_0^{\pi} \arrow[r, "\mathring{\pi}_0^\natural"] & \mathring{\xi}_0^{\mathring{\pi}^\natural} \arrow[r, "M_1"] & \mathring{\xi}_1^{\mathring{\pi}} \arrow[r] & \cdots \arrow[r] & \mathring{\xi}_t^{\mathring{\pi}^\natural} \arrow[r, "M_{t+1}"] & \mathring{\xi}_{t+1}^{\mathring{\pi}} \arrow[r, "\mathring{\pi}_{t+1}^\natural"] & \mathring{\xi}_{t+1}^{\mathring{\pi}^\natural} \arrow[r] & \cdots
\end{tikzcd}
$$

Figure: Measure flows in the change of policy

What are these weight functions?

$$\forall 1 \leq t \leq T,\ \forall x_t^\natural = (x_t, a_t) \in \mathcal{X}_t^\natural, \quad \mathring{e}_t^\natural(x_t^\natural) = \frac{\mathrm{d}\mathring{\pi}_t(x_t, \cdot)}{\mathrm{d}\pi_t(x_t, \cdot)}(a_t).$$

For $t = 0$, we let, taking into account the covariate shifts,

$$\forall x_0^\natural = (x_0, a_0) \in \mathcal{X}_0^\natural, \quad \mathring{e}_0^\natural(x_0^\natural) = \frac{\mathrm{d}\mathring{\xi}_0}{\mathrm{d}\xi_0}(x_0)\frac{\mathrm{d}\mathring{\pi}_0(x_0, \cdot)}{\mathrm{d}\pi_0(x_0, \cdot)}(a_0).$$

# Measure flows in the change of policy

$$\begin{tikzcd}
\xi_0^{\pi} \arrow[r, "\pi_0^{\natural}"] & \xi_0^{\pi^{\natural}} \arrow[r, "M_1"] \arrow[d, "\Psi_{\mathring{\eta}_0^{\natural}}"] & \xi_1^{\pi} \arrow[r] \arrow[d, "\Psi_{\mathring{\eta}_1}"] & \cdots \arrow[r] & \xi_t^{\pi^{\natural}} \arrow[r, "M_{t+1}"] \arrow[d, "\Psi_{\mathring{\eta}_t^{\natural}}"] & \xi_{t+1}^{\pi} \arrow[r, "\pi_{t+1}^{\natural}"] \arrow[d, "\Psi_{\mathring{\eta}_{t+1}}"] & \xi_{t+1}^{\pi^{\natural}} \arrow[r] \arrow[d, "\Psi_{\mathring{\eta}_{t+1}^{\natural}}"] & \cdots \\
\mathring{\xi}_0^{\pi} \arrow[r, "\mathring{\pi}_0^{\natural}"] & \mathring{\xi}_0^{\mathring{\pi}^{\natural}} \arrow[r, "M_1"] & \mathring{\xi}_1^{\mathring{\pi}} \arrow[r] & \cdots \arrow[r] & \mathring{\xi}_t^{\mathring{\pi}^{\natural}} \arrow[r, "M_{t+1}"] & \mathring{\xi}_{t+1}^{\mathring{\pi}} \arrow[r, "\mathring{\pi}_{t+1}^{\natural}"] & \mathring{\xi}_{t+1}^{\mathring{\pi}^{\natural}} \arrow[r] & \cdots
\end{tikzcd}$$

## Proposition

*Under mild assumptions, the weight functions $\mathring{\eta}_t$ and $\mathring{\eta}_t^{\natural}$ are well-defined respectively in $L^1(\mathcal{X}_t)$ and $L^1(\mathcal{X}_t^{\natural})$. In addition, for any $0 \leq t \leq T$, we have*

$$\mathring{\eta}_t^{\natural}(\cdot) = \mathbb{E}\left[\prod_{s=0}^{t} \mathring{e}_s^{\natural}(X_s^{\natural}) \,\middle|\, X_t^{\natural} = \cdot\right] \quad \textit{and} \quad \mathring{\eta}_{t+1}(\cdot) = \mathbb{E}\left[\mathring{\eta}_t^{\natural}(X_t^{\natural}) \,\middle|\, X_{t+1} = \cdot\right].$$

So, possible to implement balancing?
In a smart way, yes!

# Recursive balancing strategy

- Initial balancing: Compare $\xi_n^{\pi^\natural}$ with $\mathring{\xi}_{0,mN}^{\mathring{\pi}^\natural}$ to get the estimation of $\{\hat{\eta}_0^\natural(X_0^{(i)}) : 1 \leq i \leq n\}$.
- Weight smoothing: After getting $\{\hat{\eta}_t^\natural(X_t^{\natural(i)}) : 1 \leq i \leq n\}$ for $t \geq 0$, we run a separate regression (or do nothing, i.e., let $\hat{\eta}_{t+1}(X_{t+1}^{(i)}) = \hat{\eta}_t^\natural(X_t^{\natural(i)})$) to get $\{\hat{\eta}_{t+1}(X_{t+1}^{(i)}) : 1 \leq i \leq n\}$.
- Update balancing: Once we have $\{\hat{\eta}_{t+1}(X_{t+1}^{(i)}) : 1 \leq i \leq n\}$, we compare $\xi_{t+1,n}^{\pi^\natural}$ with $\mathring{\xi}_{t+1,mN}^{\mathring{\pi}^\natural}$ to get the estimation $\{\hat{\eta}_{t+1}^\natural(X_{t+1}^{\natural(i)}) : 1 \leq i \leq n\}$.

Now that we have estimated the weight functions, what about the actual estimators?

# Estimators

Direct estimator:

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}} = \sum_{t=0}^{T} \Psi_{\hat{\eta}_{t+1}}(\xi^{\pi}_{t+1,n})(r_t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \hat{\eta}^{(i)}_{t+1}(X^{(i)}_{t+1}) r_t(X^{(i)}_{t+1}).$$

Doubly robust estimator:

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \left( \hat{\eta}^{\natural}_t (X^{\natural(i)}_t) \left( r_t(X^{(i)}_{t+1}) - \hat{r}^{\natural(i)}_t (X^{\natural(i)}_t) \right) + \hat{\eta}^{\natural}_{t-1}(X^{\natural(i)}_{t-1}) \mathring{\pi}^{\natural}_t (\hat{r}^{\natural(i)}_t)(X^{(i)}_t) \right),$$

Sampling complexity of IPM:

$$\forall t \geq 1, \quad \sigma^{\text{IPM}}_t(n) = \text{IPM}_{\Gamma_t} \left( \Psi_{\mathring{\eta}^{\natural}_t}(\xi^{\pi^{\natural}}_{t,n}), \Psi_{\mathring{\eta}^{\natural}_t}(\xi^{\pi^{\natural}}_t) \right).$$

With a slight abuse of notation, we omit $m$ and $N$ at time 0, i.e.,

$$\sigma^{\text{IPM}}_0(n) = \sigma^{\text{IPM}}_0(n, m, N) = \text{IPM}_{\Gamma_0} \left( \Psi_{\mathring{\eta}^{\natural}_0}(\xi^{\pi^{\natural}}_{0,n}), \Psi_{\mathring{\eta}^{\natural}_0}(\xi^{\pi^{\natural}}_0) \right) + \text{IPM}_{\Gamma_0} \left( \mathring{\xi}^{\mathring{\pi}^{\natural}}_0, \mathring{\xi}^{\mathring{\pi}^{\natural}}_{0,mN} \right).$$

# Direct estimator

Well-specifiedness:
We say that the causal dynamics are well-specified by a sequence of collections of test functions $\Gamma = (\Gamma_t; 0 \leq t \leq T)$ if

$$\forall 0 \leq t \leq T, \quad r_t^{\natural} = M_{t+1}(r_t) \in \Gamma_t,$$

and

$$\forall 1 \leq t \leq T, \ \forall \gamma_t \in \Gamma_t, \quad M_t^{\mathring{\pi}^{\natural}}(\gamma_t) \in \Gamma_{t-1}.$$

## Theorem (Informal)

*Under mild assumptions, if the causal dynamics are well-specified by* $\Gamma = (\Gamma_t; 0 \leq t \leq T)$*, then we have*

$$\left| \hat{\mathcal{V}}_{\mathrm{DE}}^{\mathring{\pi}} - \mathcal{V}^{\mathring{\pi}} \right| \leq C \left( \sum_{t=0}^{T} (T - t + 1) \left( \sigma_t^{\mathrm{IPM}}(n) + \frac{1}{\sqrt{N}} \right) \right),$$

*where* $C > 0$ *is a constant that is independent to* $n, m, N$ *and* $T$*.*

# Doubly robust estimator

For the error term given by weight smoothing, we denote

$$\sigma_t^{\text{WS}}(n) = \left\| \hat{\eta}_{t+1} - \mathring{\eta}_{t+1} \right\|_{L^2(\xi^{\pi}_{t+1,n})} .$$

we denote the error of the additional regression of the average reward function $r^{\natural}$ by

$$\forall 0 \leq t \leq T, \quad \sigma_t^{\text{REG}}(n) = \left\| \hat{r}_t^{\natural} - r_t^{\natural} \right\|_{L^2(\xi^{\pi\natural}_{t,n})} .$$

## Theorem (Informal)

*It the weight functions are well specified in $H_t$ and if the implemented balancing satisfies*

$$\forall 0 \leq t \leq T, \qquad \exists \alpha_t, \beta_t > 0, \quad \alpha_t(\partial H_t) \cap \beta_t U_t \subset \Gamma_t,$$

*then we have*

$$\left| \hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}} - \mathcal{V}^{\mathring{\pi}} \right| \leq C \left( \sum_{t=0}^{T} (T - t + 1) \left( \sigma_t^{\text{IPM}}(n) + \frac{1}{\sqrt{N}} + \sigma_t^{\text{WS}}(n) \right) \sigma_t^{\text{REG}}(n) \right)$$

*where $C > 0$ is a constant that is independent from $n, m, N$, and $T$.*

# Conclusion

Our contributions:

- New state-action Markov reformulation of causal dynamics, that is capable of dealing with general action spaces and covariate shifts.
- New theoretical framework for balancing method through Riesz representable measure space arguments (connection between existing methods and inspiration of new methods).
- Recursive balancing strategy, with transparent error analysis for the DE and the DRE.

Perspectives:

- Sinkhorn balancing?
- Connection with the Feynman-Kac formalism/Sequential Monte Carlo (e.g., how to efficiently interact with the environment with the collected offline data).

Thank you for your attention!

# References


Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. (2020). Adversarial estimation of riesz representers. *arXiv e-prints*.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

Pearl, J., Bareinboim, E., et al. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

Reygner, J. and Touboul, A. (2020). Reweighting samples under covariate shift using a wasserstein distance criterion.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.