# Sinkhorn Treatment Effects: A Causal Optimal Transport Measure

**Alex Luedtke**

Harvard Medical School

# Joint work with

**Medha Agarwal**

University of Washington, Dept. of Statistics
Seattle, USA

# Early work on distributional treatment effects

Bitler et al. (2006) studied the effect of welfare reforms on earnings

What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments

*By* Marianne P. Bitler, Jonah B. Gelbach, and Hilary W. Hoynes*

# Early work on distributional treatment effects

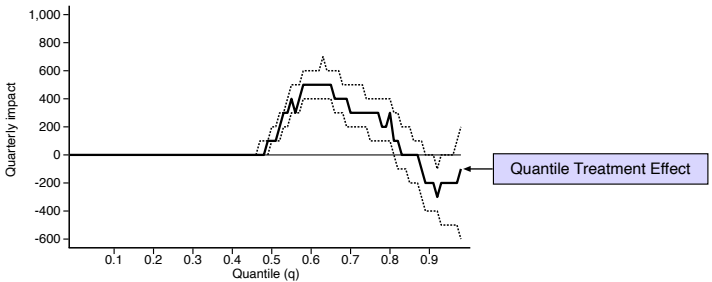
Average treatment effect $\approx 0$, so instead considered

$$\text{QTE}(q) = \Big(q\text{-Quantile under Treatment}\Big) - \Big(q\text{-Quantile under Control}\Big)$$

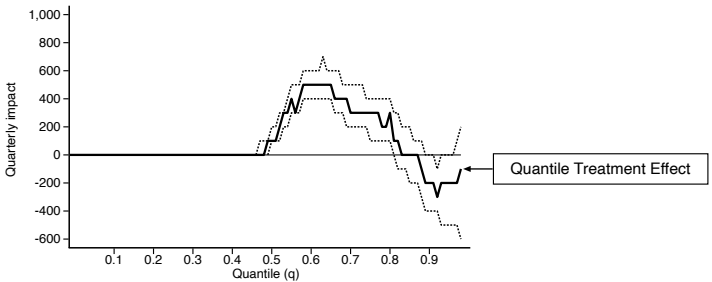
# Early work on distributional treatment effects

Average treatment effect $\approx 0$, so instead considered

$$\text{QTE}(q) = \Big(q\text{-Quantile under Treatment}\Big) - \Big(q\text{-Quantile under Control}\Big)$$

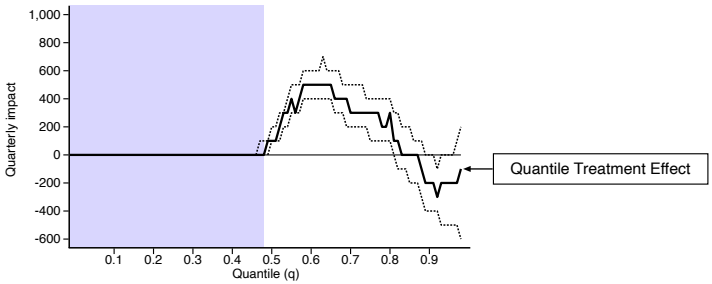


# Early work on distributional treatment effects



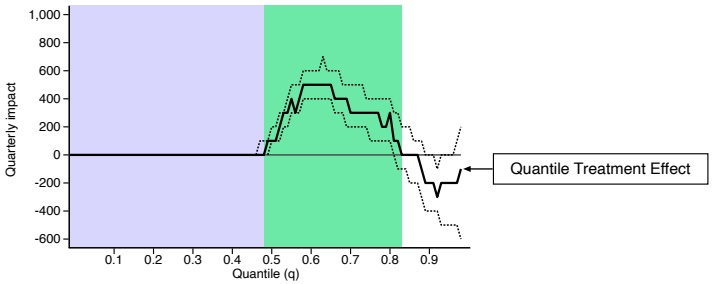

Different regions exhibited different behavior:

# Early work on distributional treatment effects



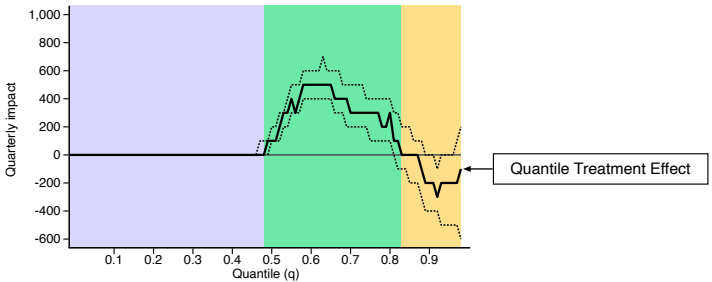

Different regions exhibited different behavior: **no effect for low earnings**

# Early work on distributional treatment effects



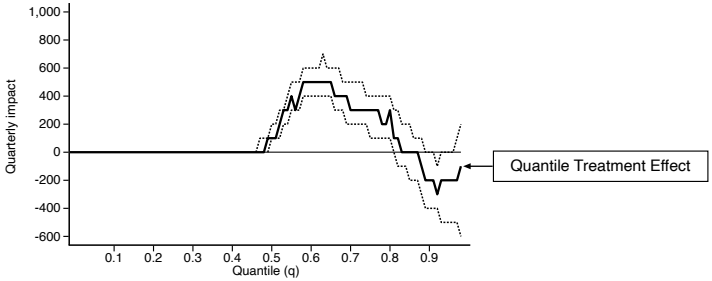

Different regions exhibited different behavior: **positive effect for mid earnings**

# Early work on distributional treatment effects

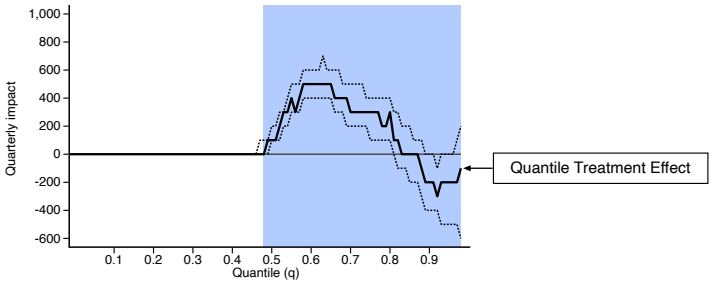



Different regions exhibited different behavior: **negative effect for high earnings.**

# QTE curve summarization via 2-Wasserstein distance

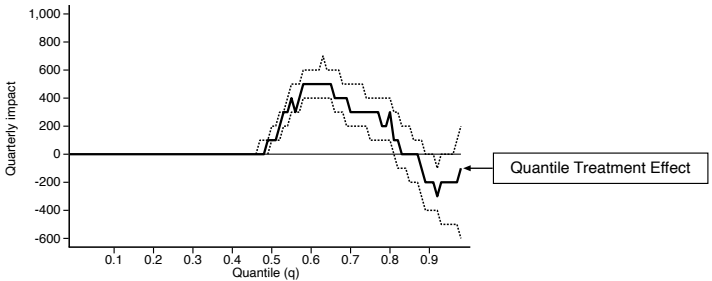


# QTE curve summarization via 2-Wasserstein distance



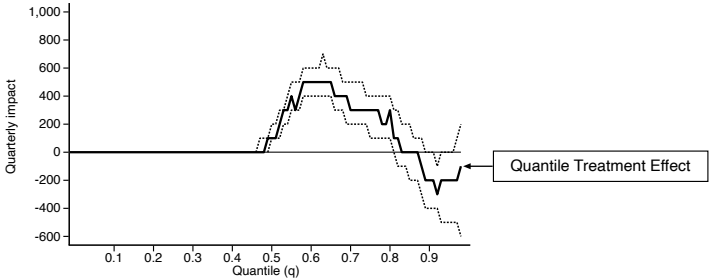

**Goal:** Summarize extent to which the QTE curve is **different from 0.**

# QTE curve summarization via 2-Wasserstein distance




**Goal:** Summarize extent to which the QTE curve is different from 0.

# QTE curve summarization via 2-Wasserstein distance




**Goal:** Summarize extent to which the QTE curve is different from 0.

**Approach:** For $P_1, P_0$ potential outcome distributions, use

$$W_2^2(P_1, P_0) = \int_0^1 \text{QTE}^2(q)\, dq,$$

# Confidence intervals for 2-Wasserstein distance

Balakrishnan et al. (2025) show how to derive $n^{-1/2}$-rate confidence intervals for

$$W_2^2(P_1, P_0) = \int_0^1 \mathrm{QTE}^2(q)\, dq.$$

# Confidence intervals for 2-Wasserstein distance

Balakrishnan et al. (2025) show how to derive $n^{-1/2}$-rate confidence intervals for

$$W_2^2(P_1, P_0) = \int_0^1 \mathrm{QTE}^2(q)\, dq.$$

**Identification strategy:** no-unmeasured confounders (G-formula):

$$P_a(Y \in \mathcal{B}) = \int P(Y \in \mathcal{B} \mid A = a, X = x)\, dP_X(x)$$

# Confidence intervals for 2-Wasserstein distance

Balakrishnan et al. (2025) show how to derive $n^{-1/2}$-rate confidence intervals for

$$W_2^2(P_1, P_0) = \int_0^1 \mathrm{QTE}^2(q)\, dq.$$

**Identification strategy:** no-unmeasured confounders (G-formula):

$$P_a(Y \in \mathcal{B}) = \int P(Y \in \mathcal{B} \mid A = a, X = x)\, dP_X(x)$$

Same strategy I'll use today.

# Beyond dimension 1

For higher dimensional outcomes $Y_1, Y_0 \in \mathbb{R}^d$, 2-Wasserstein definition generalizes as

$$W_2^2(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \; \mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] \quad \text{subject to} \;\; Y_a' \overset{d}{=} Y_a.$$

# Beyond dimension 1

For higher dimensional outcomes $Y_1, Y_0 \in \mathbb{R}^d$, 2-Wasserstein definition generalizes as

$$W_2^2(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \; \mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] \quad \text{subject to} \;\; Y_a' \overset{d}{=} Y_a.$$

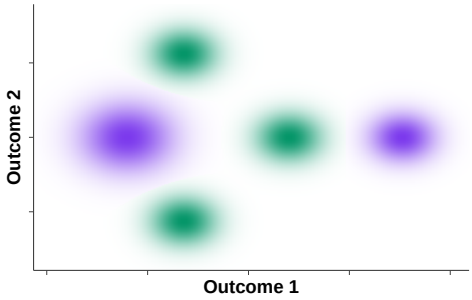
Also expressible via (randomized) transport maps from control to treatment outcomes

# Beyond dimension 1

For higher dimensional outcomes $Y_1, Y_0 \in \mathbb{R}^d$, 2-Wasserstein definition generalizes as

$$W_2^2(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \; \mathbb{E}\left[\|Y_1' - Y_0'\|_2^2\right] \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

Also expressible via (randomized) transport maps from **control** to **treatment** outcomes

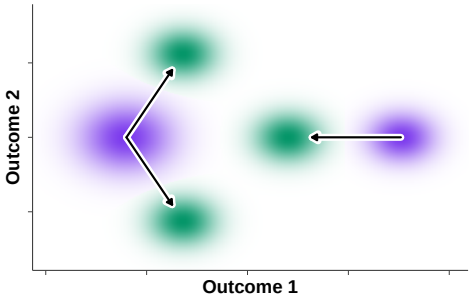


# Beyond dimension 1

For higher dimensional outcomes $Y_1, Y_0 \in \mathbb{R}^d$, 2-Wasserstein definition generalizes as

$$W_2^2(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \; \mathbb{E}\left[\|Y_1' - Y_0'\|_2^2\right] \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

Also expressible via (randomized) transport maps from **control** to **treatment** outcomes



# Challenges with 2-Wasserstein distance beyond dimension 1

Even in non-causal settings, there are challenges with trying to characterize behavior of $W_2^2(P_n, Q_n)$, for $P_n, Q_n$ empirical distributions:

1) **Only implicitly defined:** no simple quantile representation.

2) **Nonsmooth:** generally only directionally differentiable, so asymptotics nonstandard.

3) **Curse of dimensionality:** under null that $P = Q$, $\boldsymbol{n^{-2/d}}$ decay rate if $d > 4$.

# Alternative approach: counterfactual maximum mean discrepancy

Muandet et al. (2021) introduced the distributional treatment effect:

$$\text{MMD}(P_1, P_0) = \sup_{f \in \text{unit ball of RKHS}} \int f \, d(P_1 - P_0)$$

# Alternative approach: counterfactual maximum mean discrepancy

Muandet et al. (2021) introduced the distributional treatment effect:

$$\text{MMD}(P_1, P_0) = \sup_{f \in \text{unit ball of RKHS}} \int f \, d(P_1 - P_0)$$

**Doubly robust estimators** with **weak convergence guarantees** are available
(Fawkes et al., 2022; Martinez Taboada et al., 2023; Luedtke and Chung, 2024)

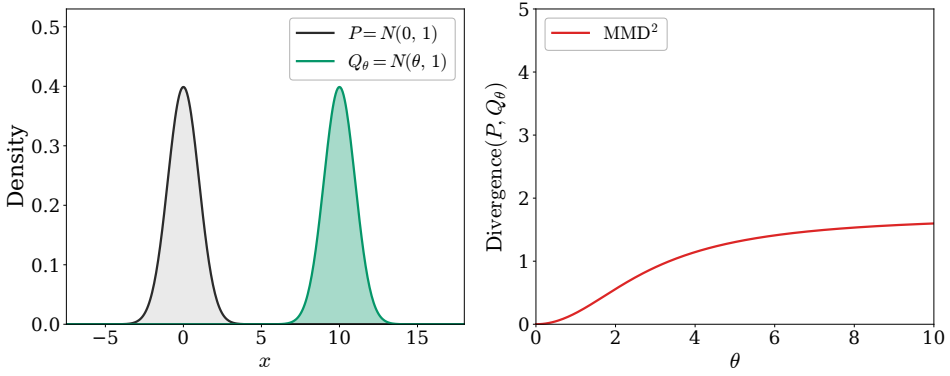
# ‘Flat’ geometry induced by MMDs (Feydy et al., 2019)

**Goal:** Distinguish between two Gaussians with different means.

# ‘Flat’ geometry induced by MMDs (Feydy et al., 2019)

**Goal:** Distinguish between two Gaussians with different means.

Gaussian-kernel **MMD saturates**, not distinguishing between ‘different’ and ‘very different’.

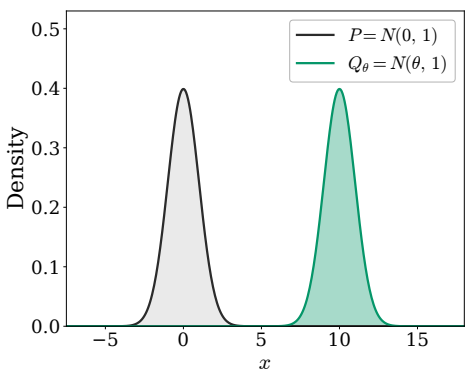


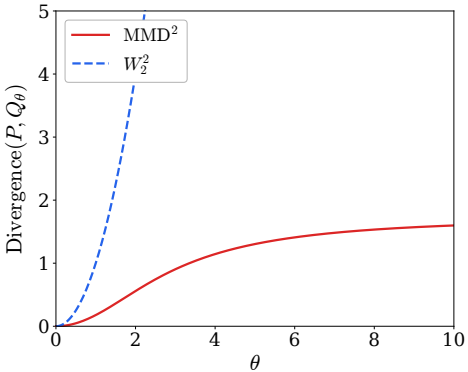

Unless stated otherwise, all bandwidths/regularization levels in this talk are chosen via the **median heuristic**.

# ‘Flat’ geometry induced by MMDs (Feydy et al., 2019)

**Goal:** Distinguish between two Gaussians with different means.

**2-Wasserstein** does not saturate.

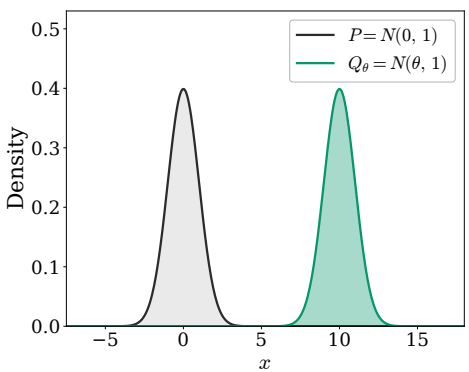


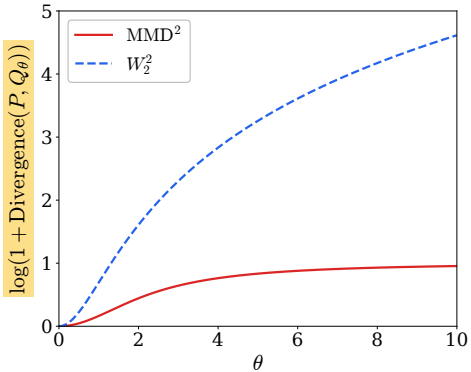



# ‘Flat’ geometry induced by MMDs (Feydy et al., 2019)

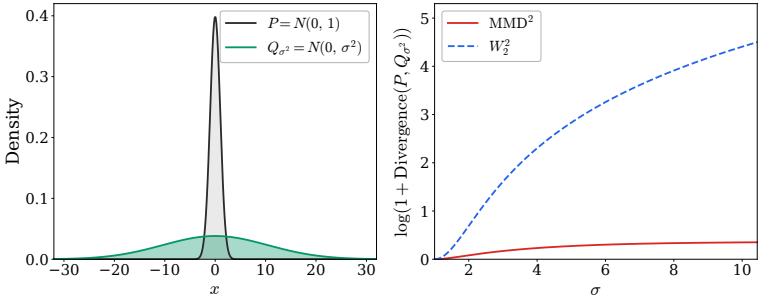
**Goal:** Distinguish between two Gaussians with different means.

**2-Wasserstein** does not saturate.

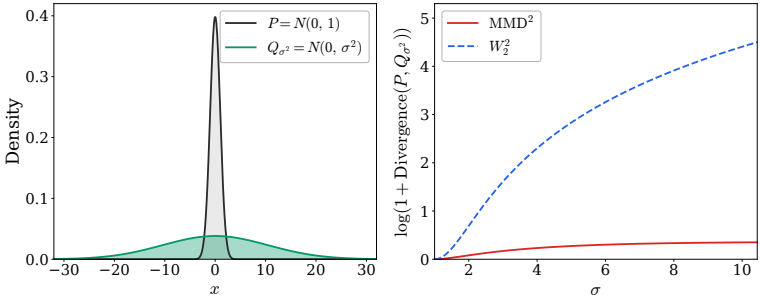


# Similar flat geometry in other settings



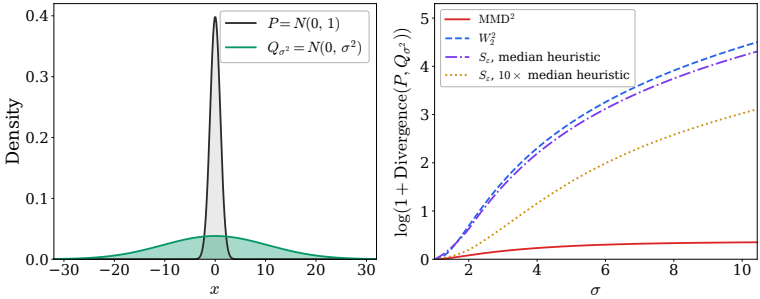
# Similar flat geometry in other settings




**Summary of what we've seen so far:**

- $W_2$ has **compelling geometric properties**
- MMD is **easy to work with statistically**

# Similar flat geometry in other settings

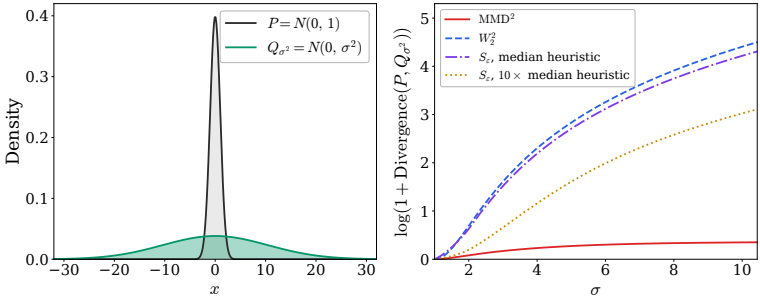



**Summary of what we've seen so far:**

- $W_2$ has **compelling geometric properties**
- MMD is **easy to work with statistically**

**Sinkhorn divergence** attains both properties (Ramdas et al., 2017; Feydy et al., 2019)

# Similar flat geometry in other settings




**Summary of what we've seen so far:**

- $W_2$ has **compelling geometric properties**
- MMD is **easy to work with statistically**

**Sinkhorn divergence** $\leftarrow$ **the focus of this talk**

# Our contributions

**New estimand:** the **Sinkhorn treatment effect**.

# Our contributions

**New estimand:** the **Sinkhorn treatment effect**.

**1st- and 2nd- (!) order influence functions** for this new estimand.

# Our contributions

**New estimand:** the **Sinkhorn treatment effect**.

**1st- and 2nd- (!) order influence functions** for this new estimand.

**Debiased estimators/tests** with theoretical guarantees.

Distributional treatment effects

Sinkhorn treatment effects

- Definition
- First-order estimation
- Second-order estimation and testing $H_0 : P_1 = P_0$

Numerical experiments

Conclusion and future work

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

Start with the **2-Wasserstein distance**.

$$W_2^2(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \quad \mathbb{E}\left[\|Y_1' - Y_0'\|_2^2\right] \qquad \text{subject to } Y_a' \overset{d}{=} Y_a.$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

**Regularize mutual information** to increase entropy of coupling.

$$\underset{(Y_1', Y_0')}{\text{minimize}} \quad \mathbb{E}\left[\|Y_1' - Y_0'\|_2^2\right] + \varepsilon I(Y_1'; Y_0') \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

> Insert a **constant** to avoid pesky '2's in downstream derivatives.

$$\underset{(Y_1', Y_0')}{\text{minimize}} \ \tfrac{1}{2}\mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] + \varepsilon\, I(Y_1';\, Y_0') \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

This gives our first **candidate divergence.**

$$\mathrm{OT}_\varepsilon(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \ \tfrac{1}{2}\,\mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] + \varepsilon\, I(Y_1';\, Y_0') \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

This gives our first **candidate divergence.**

⚠ But using it may be unintuitive: typically $\mathrm{OT}_\varepsilon(P_1, P_0) > 0$ even if $P_1 = P_0$.

$$\mathrm{OT}_\varepsilon(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \ \tfrac{1}{2}\mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] + \varepsilon\, I(Y_1';\, Y_0') \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

**Re-center** to get a semimetric.

↳ Satisfies all properties of a metric except the triangle inequality.

$$\mathrm{OT}_\varepsilon(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \ \tfrac{1}{2}\mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] + \varepsilon\, I(Y_1';\, Y_0') \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

$$\mathrm{OT}_\varepsilon(P_1, P_0) - \tfrac{1}{2}\textstyle\sum_{a=0}^{1} \mathrm{OT}_\varepsilon(P_a, P_a)$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

This gives the **Sinkhorn divergence.**

$$\mathrm{OT}_\varepsilon(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \ \tfrac{1}{2}\,\mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] + \varepsilon\, I(Y_1';\, Y_0') \quad \text{subject to} \quad Y_a' \overset{d}{=} Y_a.$$

$$\mathrm{Sinkhorn}_\varepsilon(P_1, P_0) = \mathrm{OT}_\varepsilon(P_1, P_0) - \tfrac{1}{2}\sum_{a=0}^{1} \mathrm{OT}_\varepsilon(P_a, P_a)$$

# Sinkhorn treatment effect: formal definition

To define the Sinkhorn treatment effect:

This gives the **Sinkhorn divergence.**

$$\mathrm{OT}_\varepsilon(P_1, P_0) = \underset{(Y_1', Y_0')}{\text{minimize}} \ \tfrac{1}{2}\mathbb{E}\big[\|Y_1' - Y_0'\|_2^2\big] + \varepsilon\, I(Y_1';\, Y_0') \ \text{ subject to } \ Y_a' \overset{d}{=} Y_a.$$

$$\mathrm{Sinkhorn}_\varepsilon(P_1, P_0) = \mathrm{OT}_\varepsilon(P_1, P_0) - \tfrac{1}{2}\textstyle\sum_{a=0}^{1} \mathrm{OT}_\varepsilon(P_a, P_a)$$

To simplify notation, I'll suppress the dependence of $\mathrm{Sinkhorn}_\varepsilon$ on $\varepsilon$ hereafter.

Distributional treatment effects

Sinkhorn treatment effects
- Definition
- **First-order estimation**
- Second-order estimation and testing $H_0 : P_1 = P_0$

Numerical experiments

Conclusion and future work

# First-order estimation

With $P_1, P_0$ identified through the G-formula, define the **observed data parameter**

$$\mathcal{S}(P) = \text{Sinkhorn}(P_1, P_0).$$

# First-order estimation

With $P_1, P_0$ identified through the G-formula, define the **observed data parameter**

$$\mathbb{S}(P) = \text{Sinkhorn}(P_1, P_0).$$

We derive the **efficient influence function** of $\mathbb{S}$, yielding the von Mises approximation

$$\mathbb{S}(P) \approx \mathbb{S}(\widehat{P}) + \int \dot{\mathbb{S}}_{\widehat{P}} \, dP,$$

with $\widehat{P}$ an initial estimator of $P$.

# First-order estimation

With $P_1, P_0$ identified through the G-formula, define the **observed data parameter**

$$\mathbb{S}(P) = \text{Sinkhorn}(P_1, P_0).$$

We derive the **efficient influence function** of $\mathbb{S}$, yielding the von Mises approximation

$$\mathbb{S}(P) \approx \mathbb{S}(\widehat{P}) + \int \dot{\mathbb{S}}_{\widehat{P}} \, dP,$$

with $\widehat{P}$ an initial estimator of $P$.

This suggests using the **one-step estimator**

$$\mathbb{S}(P) \approx \mathbb{S}(\widehat{P}) + \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \dot{\mathbb{S}}_{\widehat{P}}(Z_i).$$

# Confidence intervals

$$\widehat{\mathbb{S}}^{\text{one-step}} \pm 1.96 \frac{\widehat{\text{StdDev}}(\dot{\mathbb{S}}_{\widehat{P}})}{n^{1/2}}$$

has 95% asymptotic coverage under conditions.

# Confidence intervals

$$\widehat{\mathbb{S}}^{\text{one-step}} \pm 1.96 \frac{\widehat{\text{StdDev}}(\dot{\mathbb{S}}_{\widehat{P}})}{n^{1/2}}$$

has 95% asymptotic coverage under conditions.

**Important condition:** efficient influence function is nondegenerate, i.e., $\text{StdDev}(\dot{\mathbb{S}}_P) > 0$.

# Confidence intervals

$$\widehat{\mathbb{S}}^{\text{one-step}} \pm 1.96 \frac{\widehat{\text{StdDev}}(\dot{\mathbb{S}}_{\widehat{P}})}{n^{1/2}}$$

has 95% asymptotic coverage under conditions.

**Important condition:** efficient influence function is nondegenerate, i.e., $\text{StdDev}(\dot{\mathbb{S}}_P) > 0$.

This **standard deviation converges to zero** in an important special case, namely under

$$H_0 : P_1 = P_0$$

# Confidence intervals

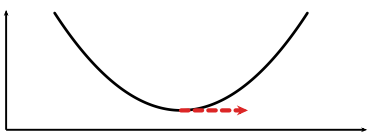
$$\widehat{\mathbb{S}}^{\text{one-step}} \pm 1.96 \frac{\widehat{\text{StdDev}}(\dot{\mathbb{S}}_{\widehat{P}})}{n^{1/2}}$$

has 95% asymptotic coverage under conditions.

**Important condition:** efficient influence function is nondegenerate, i.e., $\text{StdDev}(\dot{\mathbb{S}}_P) > 0$.

This **standard deviation converges to zero** in an important special case, namely under

$$H_0 : P_1 = P_0$$



Distributional treatment effects

Sinkhorn treatment effects

- Definition
- First-order estimation
- **Second-order estimation and testing** $H_0 : P_1 = P_0$

Numerical experiments

Conclusion and future work

# Testing the null via second-order influence functions

To test $H_{\mathbb{O}}$, we seek to use a von Mises approximation with **first-** and **second-** order terms:

$$\mathcal{S}(P) \approx \mathcal{S}(\widehat{P}) + \int \dot{\mathcal{S}}_{\widehat{P}}\, dP + \tfrac{1}{2} \int \ddot{\mathcal{S}}_{\widehat{P}}\, d(P \otimes P).$$

# Testing the null via second-order influence functions

To test $H_{\mathbb{0}}$, we seek to use a von Mises approximation with first- and second- order terms:

$$\mathcal{S}(P) \approx \mathcal{S}(\widehat{P}) + \int \dot{\mathcal{S}}_{\widehat{P}}\, dP \;+\; \tfrac{1}{2} \int \ddot{\mathcal{S}}_{\widehat{P}}\, d(P \otimes P)\,.$$

This would then suggest using the approximation

$$\mathcal{S}(P) \approx \;\; \mathcal{S}(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \dot{\mathcal{S}}_{\widehat{P}}(Z_i) \;\; + \;\; \frac{1}{2}\binom{n}{2}^{-1} \sum_{i<j} \ddot{\mathcal{S}}_{\widehat{P}}(Z_i, Z_j)\;.$$

# Testing the null via second-order influence functions

To test $H_{\mathbb{0}}$, we seek to use a von Mises approximation with first- and second- order terms:

$$\mathcal{S}(P) \approx \mathcal{S}(\widehat{P}) + \int \dot{\mathcal{S}}_{\widehat{P}}\, dP + \tfrac{1}{2} \int \ddot{\mathcal{S}}_{\widehat{P}}\, d(P \otimes P)\,.$$

This would then suggest using the approximation

$$\mathcal{S}(P) \approx \underbrace{\mathcal{S}(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \dot{\mathcal{S}}_{\widehat{P}}(Z_i)}_{\textbf{One-step estimator}} + \frac{1}{2}\binom{n}{2}^{-1} \sum_{i<j} \ddot{\mathcal{S}}_{\widehat{P}}(Z_i, Z_j)\,.$$

# Testing the null via second-order influence functions

To test $H_{\mathbb{0}}$, we seek to use a von Mises approximation with first- and second- order terms:

$$\mathcal{S}(P) \approx \mathcal{S}(\widehat{P}) + \int \dot{\mathcal{S}}_{\widehat{P}}\, dP + \tfrac{1}{2} \int \ddot{\mathcal{S}}_{\widehat{P}}\, d(P \otimes P)\,.$$

This would then suggest using the approximation

$$\mathcal{S}(P) \approx \underbrace{\mathcal{S}(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \dot{\mathcal{S}}_{\widehat{P}}(Z_i)}_{\widehat{\mathcal{S}}^{\text{one-step}}} + \frac{1}{2}\binom{n}{2}^{-1} \sum_{i<j} \ddot{\mathcal{S}}_{\widehat{P}}(Z_i, Z_j)\,.$$

# Testing the null via second-order influence functions

To test $H_{\mathbb{0}}$, we seek to use a von Mises approximation with first- and second- order terms:

$$\mathcal{S}(P) \approx \mathcal{S}(\widehat{P}) + \int \dot{\mathcal{S}}_{\widehat{P}}\, dP + \tfrac{1}{2} \int \ddot{\mathcal{S}}_{\widehat{P}}\, d(P \otimes P)\,.$$

This would then suggest using the approximation

$$\mathcal{S}(P) \approx \underbrace{\mathcal{S}(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \dot{\mathcal{S}}_{\widehat{P}}(Z_i)}_{\widehat{\mathcal{S}}^{\text{one-step}}} + \underbrace{\frac{1}{2}\binom{n}{2}^{-1} \sum_{i<j} \ddot{\mathcal{S}}_{\widehat{P}}(Z_i, Z_j)}_{\textbf{2nd-order correction}}\,.$$

# Testing the null via second-order influence functions

To test $H_{\mathbb{0}}$, we seek to use a von Mises approximation with first- and second- order terms:

$$\mathcal{S}(P) \approx \mathcal{S}(\widehat{P}) + \int \dot{\mathcal{S}}_{\widehat{P}}\, dP + \tfrac{1}{2} \int \ddot{\mathcal{S}}_{\widehat{P}}\, d(P \otimes P)\,.$$

This would then suggest using the approximation

$$\mathcal{S}(P) \approx \underbrace{\mathcal{S}(\widehat{P}) + \frac{1}{n}\sum_{i=1}^{n} \dot{\mathcal{S}}_{\widehat{P}}(Z_i)}_{\widehat{\mathcal{S}}^{\text{one-step}}} + \underbrace{\frac{1}{2}\binom{n}{2}^{-1} \sum_{i<j} \ddot{\mathcal{S}}_{\widehat{P}}(Z_i, Z_j)}_{\frac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}}\,.$$

# Unfortunate fact: 2nd-order influence functions don't usually exist

*Statistical Science*
2014, Vol. 29, No. 4, 679–686
DOI: 10.1214/14-STS478


## Higher Order Tangent Spaces and Influence Functions

**Aad van der Vaart**

*Unfortunately, there is no such free lunch: one cannot seriously correct bias without seriously increasing the variance. Although Eq. 1.12 and the preceding heuristics are correct, they do not apply, as* **higher-order influence functions typically do not exist.** *Besides by a lack of invertibility of the map* $\eta \mapsto p_\eta$, **this is caused by failure of a higher order von Mises type representation.**

# Unfortunate fact: 2nd-order influence functions don't usually exist

IMS Collections
**Probability and Statistics: Essays in Honor of David A. Freedman**
Vol. 2 (2008) 335–421

DOI: 10.1214/193940307000000527

## Higher order influence functions and minimax estimation of nonlinear functionals

**James Robins[1], Lingling Li[1], Eric Tchetgen[1] and Aad van der Vaart[2]**

Robins et al. gave many interesting examples; **none have 2nd-order influence functions:**

# Unfortunate fact: 2nd-order influence functions don't usually exist

IMS Collections
**Probability and Statistics: Essays in Honor of David A. Freedman**
Vol. 2 (2008) 335–421

DOI: 10.1214/193940307000000527

## Higher order influence functions and minimax estimation of nonlinear functionals

**James Robins[1], Lingling Li[1], Eric Tchetgen[1] and Aad van der Vaart[2]**

Robins et al. gave many interesting examples; **none have 2nd-order influence functions:**

**Example 1a (Expected product of conditional expectations).** . . .

# Unfortunate fact: 2nd-order influence functions don't usually exist

IMS Collections
**Probability and Statistics: Essays in Honor of David A. Freedman**
Vol. 2 (2008) 335–421

DOI: 10.1214/193940307000000527

## Higher order influence functions and minimax estimation of nonlinear functionals

James Robins[1], Lingling Li[1], Eric Tchetgen[1] and Aad van der Vaart[2]

Robins et al. gave many interesting examples; **none have 2nd-order influence functions:**

**Example 1a (Expected product of conditional expectations).** ...

**Example 1b (Expected conditional covariance).** ...

# Unfortunate fact: 2nd-order influence functions don't usually exist

IMS Collections
**Probability and Statistics: Essays in Honor of David A. Freedman**
Vol. 2 (2008) 335–421

DOI: 10.1214/193940307000000527

## Higher order influence functions and minimax estimation of nonlinear functionals

James Robins[1], Lingling Li[1], Eric Tchetgen[1] and Aad van der Vaart[2]

Robins et al. gave many interesting examples; **none have 2nd-order influence functions:**

**Example 1a (Expected product of conditional expectations).** ...

**Example 1b (Expected conditional covariance).** ...

**Example 1c (Variance-weighted average treatment effect).** ...

# Unfortunate fact: 2nd-order influence functions don't usually exist

IMS Collections
**Probability and Statistics: Essays in Honor of David A. Freedman**
Vol. 2 (2008) 335–421

DOI: 10.1214/193940307000000527

## Higher order influence functions and minimax estimation of nonlinear functionals

James Robins[1], Lingling Li[1], Eric Tchetgen[1] and Aad van der Vaart[2]

Robins et al. gave many interesting examples; **none have 2nd-order influence functions:**

**Example 1a (Expected product of conditional expectations).** ...

**Example 1b (Expected conditional covariance).** ...

**Example 1c (Variance-weighted average treatment effect).** ...

⋮

# Unfortunate fact: 2nd-order influence functions don't usually exist

IMS Collections
**Probability and Statistics: Essays in Honor of David A. Freedman**
Vol. 2 (2008) 335–421

DOI: 10.1214/193940307000000527

## Higher order influence functions and minimax estimation of nonlinear functionals

James Robins[1], Lingling Li[1], Eric Tchetgen[1] and Aad van der Vaart[2]

Robins et al. gave many interesting examples; **none have 2nd-order influence functions:**

**Example 1a (Expected product of conditional expectations).** ...

**Example 1b (Expected conditional covariance).** ...

**Example 1c (Variance-weighted average treatment effect).** ...

⋮

**Example 4 (Confidence intervals for the optimal treatment strategy).** ...

# Settings I'm aware of where 2nd-order influence functions exist

**1)** **Parametric models** and projections onto them (Robins et al., 2008).

# Settings I'm aware of where 2nd-order influence functions exist

**1) Parametric models** and projections onto them (Robins et al., 2008).

**2) Maximum mean discrepancy** between pushforwards of (possibly unknown) functions, under the null that they're the same (Luedtke, Carone, et al., 2019).

- E.g., can test

$$H_0 : E[Y_1|X] \overset{\text{d}}{=} E[Y_0|X],$$

or could test

$$H_0 : \text{CATE}(X) \overset{\text{a.s.}}{=} 0.$$

# Settings I'm aware of where 2nd-order influence functions exist

**1) Parametric models** and projections onto them (Robins et al., 2008).

**2) Maximum mean discrepancy** between pushforwards of (possibly unknown) functions, under the null that they're the same (Luedtke, Carone, et al., 2019).

# Settings I'm aware of where 2nd-order influence functions exist

**1) Parametric models** and projections onto them (Robins et al., 2008).

**2) Maximum mean discrepancy** between pushforwards of (possibly unknown) functions, under the null that they're the same (Luedtke, Carone, et al., 2019).

**3) Squared Hilbert norms** of Hilbert-valued parameters $\psi$ that have an efficient influence function, under $H_0 : \psi(P) = 0$ (Luedtke and Chung, 2024; Agarwal et al., 2026).

- E.g., in the work I'm discussing today, we show this implies that

$$P \mapsto \mathrm{MMD}^2(P_1, P_0)$$

has a 2nd-order influence function.

# Settings I'm aware of where 2nd-order influence functions exist

**1) Parametric models** and projections onto them (Robins et al., 2008).

**2) Maximum mean discrepancy** between pushforwards of (possibly unknown) functions, under the null that they're the same (Luedtke, Carone, et al., 2019).

**3) Squared Hilbert norms** of Hilbert-valued parameters $\psi$ that have an efficient influence function, under $H_0 : \psi(P) = 0$ (Luedtke and Chung, 2024; Agarwal et al., 2026).

# Settings I'm aware of where 2nd-order influence functions exist

**1) Parametric models** and projections onto them (Robins et al., 2008).

**2) Maximum mean discrepancy** between pushforwards of (possibly unknown) functions, under the null that they're the same (Luedtke, Carone, et al., 2019).

**3) Squared Hilbert norms** of Hilbert-valued parameters $\psi$ that have an efficient influence function, under $H_0 : \psi(P) = 0$ (Luedtke and Chung, 2024; Agarwal et al., 2026).

**4) Second-order Hadamard differentiable maps** of expectation operators (Ren et al., 2001; Keller, 2006).

# Settings I'm aware of where 2nd-order influence functions exist

**1) Parametric models** and projections onto them (Robins et al., 2008).

**2) Maximum mean discrepancy** between pushforwards of (possibly unknown) functions, under the null that they're the same (Luedtke, Carone, et al., 2019).

**3) Squared Hilbert norms** of Hilbert-valued parameters $\psi$ that have an efficient influence function, under $H_0 : \psi(P) = 0$ (Luedtke and Chung, 2024; Agarwal et al., 2026).

**4) Second-order Hadamard differentiable maps** of expectation operators (Ren et al., 2001; Keller, 2006).

- Important example: (non-causal) **Sinkhorn divergence** (Goldfeld et al., 2024; Lavenant et al., 2024; Kokot et al., 2025)

# 2nd-order influence function of Sinkhorn treatment effect under $H_0$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n\,\ddot{\mathcal{S}}_{\widehat{P}},$$

the **usual first-order one-step estimator** plus a **second-order correction.**

# 2nd-order influence function of Sinkhorn treatment effect under $H_0$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

# 2nd-order influence function of Sinkhorn treatment effect under $H_0$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_0 : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \text{AIPWTransform}_P^{\otimes 2}(\, k_{P_0}\,)(z_1, z_2)$$

# 2nd-order influence function of Sinkhorn treatment effect under $H_{\mathbb{0}}$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_{\mathbb{0}} : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \text{AIPWTransform}_P^{\otimes 2}(k_{P_0})(z_1, z_2),$$

where $k_{P_1} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the

(uncentered) **2nd-order influence function of** $P_1 \mapsto \text{Sinkhorn}(P_1, P_0)$ at $P_1 = P_0$.

Explicit forms derived in Lavenant et al. (2024) and Kokot and Luedtke (2025).

# 2nd-order influence function of Sinkhorn treatment effect under $H_0$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_0 : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \text{AIPWTransform}_P^{\otimes 2}(\, k_{P_0}\,)(z_1, z_2),$$

# 2nd-order influence function of Sinkhorn treatment effect under $H_0$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_0 : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \big(\,\text{AIPWTransform}_P \otimes \text{AIPWTransform}_P\,\big)(\,k_{P_0}\,)(z_1, z_2),$$

# 2nd-order influence function of Sinkhorn treatment effect under $H_0$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_0 : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \big(\text{AIPWTransform}_P \otimes \text{AIPWTransform}_P\big)(\,k_{P_0}\,)(z_1, z_2),$$

where

$\text{AIPWTransform}_P(f) =$ EIF of the Avg Treatment Effect on $f(Y)$, $\quad \forall f : \mathcal{Y} \to \mathbb{R}$

# 2nd-order influence function of Sinkhorn treatment effect under $H_{\mathbb{0}}$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_{\mathbb{0}} : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \big(\text{AIPWTransform}_P \otimes \text{AIPWTransform}_P\big)\big(k_{P_0}\big)(z_1, z_2),$$

where

$$\text{AIPWTransform}_P(f) : (x, a, y) \mapsto \begin{array}{l} \frac{2a-1}{P(A=a\,|\,x)}\,\{f(y) - E_P[f(Y)\,|\,a, x]\} \\ + E_P[f(Y)\,|\,A=1, x] - E_P[f(Y)\,|\,A=0, x] - \text{ATE}_P(f). \end{array}$$

# 2nd-order influence function of Sinkhorn treatment effect under $H_{\mathbb{0}}$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n\,\ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_{\mathbb{0}} : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \big(\,\text{AIPWTransform}_P \otimes \text{AIPWTransform}_P\,\big)(\,k_{P_0}\,)(z_1, z_2)$$

# 2nd-order influence function of Sinkhorn treatment effect under $H_{\emptyset}$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_{\emptyset} : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_P(z_1, z_2) = \big(\,\text{AIPWTransform}_P \otimes \text{AIPWTransform}_P\,\big)(\,k_{P_0}\,)(z_1, z_2)$$

**Potential problem:** When $\widehat{P}_1 \neq \widehat{P}_0$, 2nd-order influence function $\ddot{\mathcal{S}}_{\widehat{P}}$ need not exist.

# 2nd-order influence function of Sinkhorn treatment effect under $H_{\mathbb{0}}$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n\,\ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_{\mathbb{0}} : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_{P}(z_1, z_2) = \big(\,\text{AIPWTransform}_{P} \otimes \text{AIPWTransform}_{P}\,\big)(\,k_{P_0}\,)(z_1, z_2)$$

**Potential problem:** When $\widehat{P}_1 \neq \widehat{P}_0$, 2nd-order influence function $\ddot{\mathcal{S}}_{\widehat{P}}$ need not exist.

↳ **Good news!** We show can use an obvious extension: replace $P$ by $\widehat{P}$ above.

# 2nd-order influence function of Sinkhorn treatment effect under $H_{\mathbb{0}}$

**Recall:** a (nonparametric) **second-order influence function** yields

$$\mathcal{S}(P) \approx \widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\,\mathbb{U}_n\,\ddot{\mathcal{S}}_{\widehat{P}}\,.$$

We show this (nonparametric) influence function exists under $H_{\mathbb{0}} : P_1 = P_0$, taking the form

$$\ddot{\mathcal{S}}_{\widehat{P}}(z_1, z_2) = \big(\,\text{AIPWTransform}_{\widehat{P}} \otimes \text{AIPWTransform}_{\widehat{P}}\,\big)\big(\,k_{\widehat{P}_0}\,\big)(z_1, z_2)$$

**Potential problem:** When $\widehat{P}_1 \neq \widehat{P}_0$, 2nd-order influence function $\ddot{\mathcal{S}}_{\widehat{P}}$ need not exist.

↳ **Good news!** We show can use an obvious extension: replace $P$ by $\widehat{P}$ above.

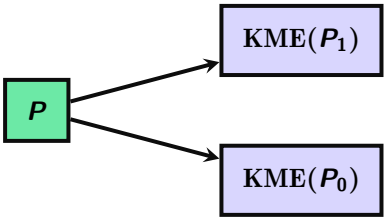
# Chain rule: strategy for establishing 2nd-order differentiability

**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

# Chain rule: strategy for establishing 2nd-order differentiability


**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

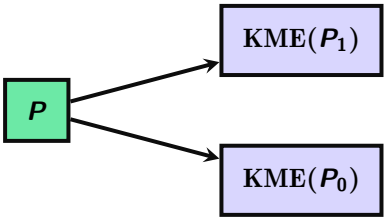
# Chain rule: strategy for establishing 2nd-order differentiability




**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

1) **Kernel mean embedding:** $P \mapsto \mathrm{KME}(P_a)$ (Luedtke and Chung, 2024)
   ↳ Sobolev kernel (fixed $s > d/2$)

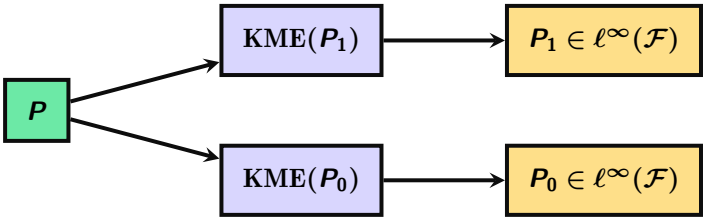
# Chain rule: strategy for establishing 2nd-order differentiability




**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

**1) Kernel mean embedding:** $P \mapsto \mathrm{KME}(P_a)$ (Luedtke and Chung, 2024)

# Chain rule: strategy for establishing 2nd-order differentiability

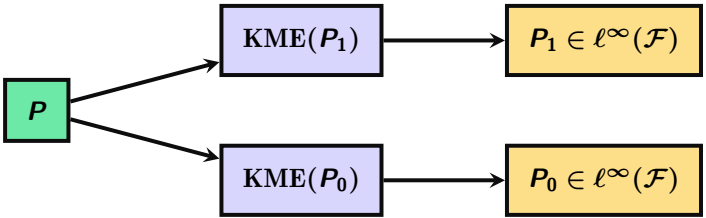



**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

1) **Kernel mean embedding:** $P \mapsto \mathrm{KME}(P_a)$ (Luedtke and Chung, 2024)

2) **Isometric embedding into** $\ell^\infty(\mathcal{F})$: $\mathrm{KME}(P_a) \mapsto (f \mapsto P_a f)$
   ↳ $\mathcal{F}$ = the unit ball in the Sobolev RKHS

# Chain rule: strategy for establishing 2nd-order differentiability

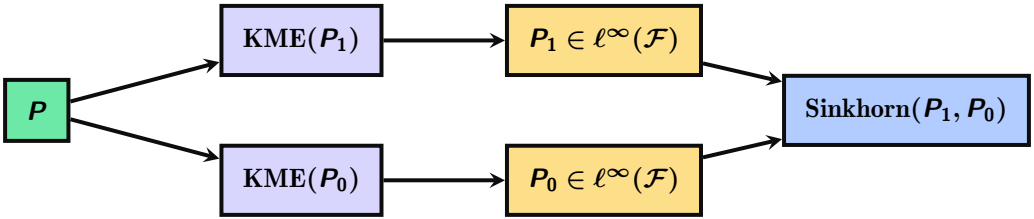



**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

1) **Kernel mean embedding:** $P \mapsto \mathrm{KME}(P_a)$ (Luedtke and Chung, 2024)

2) **Isometric embedding into** $\ell^\infty(\mathcal{F})$**:** $\mathrm{KME}(P_a) \mapsto (f \mapsto P_a f)$

# Chain rule: strategy for establishing 2nd-order differentiability




**Proof strategy:** Write estimand as a composition of smooth maps, and apply a **chain rule**.

1) **Kernel mean embedding:** $P \mapsto \mathrm{KME}(P_a)$ (Luedtke and Chung, 2024)

2) **Isometric embedding into $\ell^\infty(\mathcal{F})$:** $\mathrm{KME}(P_a) \mapsto (f \mapsto P_a f)$

3) **Sinkhorn divergence:** $(P_1, P_0) \mapsto \mathrm{Sinkhorn}(P_1, P_0)$ (Lavenant et al., 2024)

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$:

$$D^2\mathcal{S}_P = \; D^2 \,\text{Sinkhorn}\left(D\mu_P,\, D\mu_P\right) \;\; + \;\; D\,\text{Sinkhorn}\left(\; D^2\mu_P \;\right) .$$

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$:

$$D^2\mathcal{S}_P = \underbrace{D^2\,\text{Sinkhorn}\left(D\mu_P,\, D\mu_P\right)}_{\textbf{harmless}} + \underbrace{D\,\text{Sinkhorn}\left(\,D^2\mu_P\,\right)}_{\textbf{problematic}}.$$

⚠ $D^2\mu_P$ (causal curvature) usually **has no 2nd-order IF** (Robins et al., 2008).

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$:

$$D^2\mathcal{S}_P = \; D^2 \,\text{Sinkhorn}\left(D\mu_P,\, D\mu_P\right) \;+\; D\,\text{Sinkhorn}\left(\, D^2\mu_P \,\right) .$$

**Under** $H_0 : P_1 = P_0$, Sinkhorn is minimized, so its first derivative vanishes:

$$D\,\text{Sinkhorn}(\,\cdot\,) = 0$$

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$:

$$D^2\mathcal{S}_P = \; D^2\,\text{Sinkhorn}\left(D\mu_P,\, D\mu_P\right) \;+\; D\,\text{Sinkhorn}\left(\, D^2\mu_P \,\right).$$

**Under** $H_0 : P_1 = P_0$, Sinkhorn is minimized, so its first derivative vanishes:

$$D\,\text{Sinkhorn}(\,\cdot\,) = 0 \quad \Longrightarrow \quad D\,\text{Sinkhorn}\left(D^2\mu_P\right) = 0.$$

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$:

$$D^2\mathcal{S}_P = D^2\,\text{Sinkhorn}\left(D\mu_P, D\mu_P\right) + D\,\text{Sinkhorn}\left(D^2\mu_P\right).$$

**Under** $H_0 : P_1 = P_0$, Sinkhorn is minimized, so its first derivative vanishes

Only the **harmless term** survives, so a 2nd-order influence function $\ddot{\mathcal{S}}_P$ **exists under** $H_0$.

# Weak convergence under the null

**Theorem** (informal): Suppose $H_0 : P_1 = P_0$ and

**1)** the extension $\ddot{\mathbb{S}}_{\widehat{P}}$ is $L^2$-consistent for $\ddot{\mathbb{S}}_P$;

**2)** a doubly robust remainder is $o_p(n^{-1/2})$;

**3)** the third-order von Mises remainder is $o_p(n^{-1})$.

Then, the **2nd-order estimator** satisfies

$$
\begin{aligned}
& n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n\ddot{\mathbb{S}}_{\widehat{P}} - \text{Sinkhorn}(P_1, P_0)\right] \\
&\quad = \tfrac{n}{2}\mathbb{U}_n\ddot{\mathbb{S}}_P + o_p(1) \\
&\quad \xrightarrow{d} \sum_{j=1}^{\infty} \frac{\lambda_j}{2}\left[\chi^2_j(1) - 1\right].
\end{aligned}
$$

# Weak convergence under the null

**Theorem** (informal): Suppose $H_0 : P_1 = P_0$ and

**1)** the extension $\ddot{\mathbb{S}}_{\widehat{P}}$ is $L^2$-consistent for $\ddot{\mathbb{S}}_P$;

**2)** a doubly robust remainder is $o_p(n^{-1/2})$;

**3)** the third-order von Mises remainder is $o_p(n^{-1})$.

Then, the **2nd-order estimator** satisfies

$$
\begin{aligned}
& n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n\ddot{\mathbb{S}}_{\widehat{P}} - \text{Sinkhorn}(P_1, P_0)\right] \\
&\quad = \tfrac{n}{2}\mathbb{U}_n\ddot{\mathbb{S}}_P + o_p(1) \quad \leftarrow \textbf{asymptotically quadratic} \\
&\quad \xrightarrow{d} \sum_{j=1}^{\infty} \frac{\lambda_j}{2}\left[\chi_j^2(1) - 1\right].
\end{aligned}
$$

# Weak convergence under the null

**Theorem** (informal): Suppose $H_0 : P_1 = P_0$ and

**1)** the extension $\ddot{\mathbb{S}}_{\widehat{P}}$ is $L^2$-consistent for $\ddot{\mathbb{S}}_P$;

**2)** a doubly robust remainder is $o_p(n^{-1/2})$;

**3)** the third-order von Mises remainder is $o_p(n^{-1})$.

Then, the **2nd-order estimator** satisfies

$$
\begin{aligned}
& n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n\ddot{\mathbb{S}}_{\widehat{P}} - \text{Sinkhorn}(P_1, P_0)\right] \\
&\quad = \tfrac{n}{2}\mathbb{U}_n\ddot{\mathbb{S}}_P + o_p(1) \\
&\quad \xrightarrow{d} \sum_{j=1}^{\infty} \frac{\lambda_j}{2}\left[\chi_j^2(1) - 1\right].
\end{aligned}
$$

$\leftarrow$ **usual degenerate U-stat limit**

# Weak convergence under the null

**Theorem** (informal): Suppose $H_0 : P_1 = P_0$ and

**1)** the extension $\ddot{\mathbb{S}}_{\widehat{P}}$ is $L^2$-consistent for $\ddot{\mathbb{S}}_P$; ← **arbitrarily slow rate allowed**

**2)** a doubly robust remainder is $o_p(n^{-1/2})$; ← $n^{-1/4}$ **nuisance rates**

**3)** the third-order von Mises remainder is $o_p(n^{-1})$. ← $n^{-1/3}$ **nuisance rates**

Then, the **2nd-order estimator** satisfies

$$
\begin{aligned}
& n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n\ddot{\mathbb{S}}_{\widehat{P}} - \text{Sinkhorn}(P_1, P_0)\right] \\
&\quad = \tfrac{n}{2}\mathbb{U}_n\ddot{\mathbb{S}}_P + o_p(1) \\
&\quad \xrightarrow{d} \sum_{j=1}^{\infty} \frac{\lambda_j}{2}\left[\chi_j^2(1) - 1\right].
\end{aligned}
$$

# Other theoretical properties

**Under the null:** Previous slide ensures proper type 1 error if test rejects when

$$n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_{\widehat{P}}\right] > (1-\alpha) \text{ quantile of U-stat limit}$$

# Other theoretical properties

**Under the null:** Previous slide ensures proper type 1 error if test rejects when

$$n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_{\widehat{P}}\right] > (1-\alpha) \text{ quantile of U-stat limit}$$

**Under a fixed alternative:** Under similar nuisance rate conditions to those used earlier,

$$\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_{\widehat{P}} = \widehat{\mathbb{S}}^{\text{one-step}} + o_p(n^{-1/2}).$$

Hence, **2nd-order estimator** inherits asymptotic normality of the **one-step estimator**.

# Other theoretical properties

**Under the null:** Previous slide ensures proper type 1 error if test rejects when

$$n\left[\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_{\widehat{P}}\right] > (1-\alpha) \text{ quantile of U-stat limit}$$

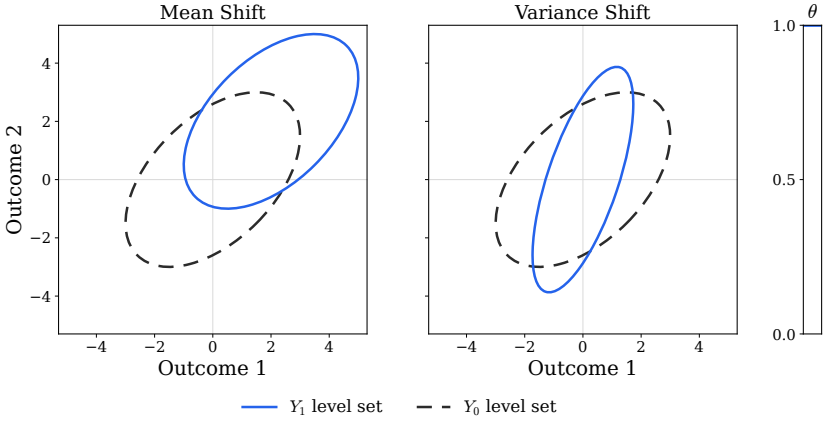
**Under a fixed alternative:** Under similar nuisance rate conditions to those used earlier,

$$\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_{\widehat{P}} = \widehat{\mathbb{S}}^{\text{one-step}} + o_p(n^{-1/2}).$$

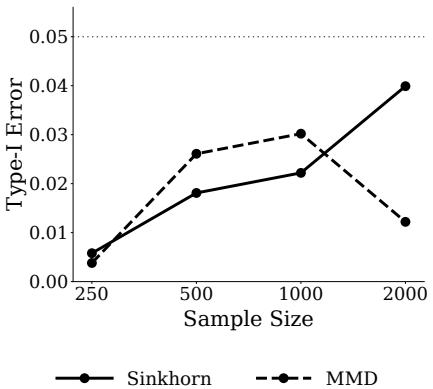
Hence, 2nd-order estimator inherits asymptotic normality of the one-step estimator.

↳ In particular, this ensures the **test has power against fixed alternatives**.

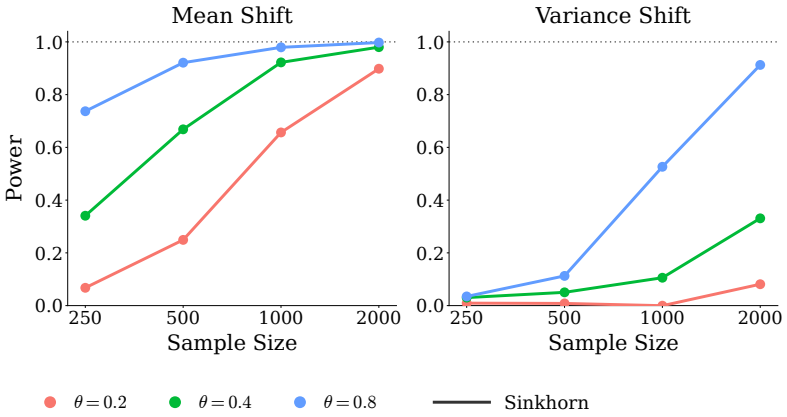
# Simulation setting: detecting 2d mean and variance shifts



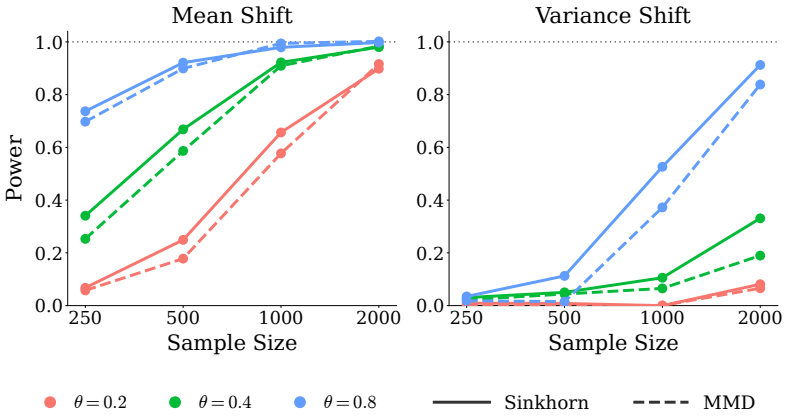
# Simulation results: type 1 error



# Simulation results: power



# Simulation results: power



# Recap

**New estimand:** the **Sinkhorn treatment effect**

**1st- and 2nd- order influence functions** for this new estimand

**Debiased estimators/tests**

# Extension to conditional distributional treatment effects

Rather than testing equality **marginally**, might want to test it **conditionally**:

$$H_0^{\text{cond}} : P_{Y(1)\,|\,X} = P_{Y(0)\,|\,X} \text{ a.s.}$$

# Extension to conditional distributional treatment effects

Rather than testing equality **marginally**, might want to test it **conditionally**:

$$H_0^{\text{cond}} : P_{Y(1)\,|\,X} = P_{Y(0)\,|\,X} \text{ a.s.}$$

Jain et al. (2026) test $H_0^{\text{cond}}$ using **unconditional** counterfactual mean embeddings.
↳ **Key insight:** Under our identifiability conditions,

$$H_0^{\text{cond}} \iff \underbrace{P_{Y(1),X} = P_{Y(0),X}}_{\text{unconditional null}} .$$

# Extension to conditional distributional treatment effects

Rather than testing equality **marginally**, might want to test it **conditionally**:

$$H_0^{\text{cond}} : P_{Y(1)\,|\,X} = P_{Y(0)\,|\,X} \text{ a.s.}$$

Jain et al. (2026) test $H_0^{\text{cond}}$ using **unconditional** counterfactual mean embeddings.
↳ **Key insight:** Under our identifiability conditions,

$$H_0^{\text{cond}} \iff P_{Y(1),X} = P_{Y(0),X} \ .$$

Same idea applies to Sinkhorn treatment effects:

1) **Augment outcome:** $Y^{\text{aug}}(a) = (Y(a), X)$.
2) **Test** $H_0^{\text{cond}}$ via $\text{Sinkhorn}\big(P_{Y^{\text{aug}}(1)}, P_{Y^{\text{aug}}(0)}\big)$.

# Open question

Our theory focused on the case where the entropic **regularization level $\varepsilon$** is fixed.*

*In the paper, we also showed how to data-adaptively select $\varepsilon$ over a fixed grid via an aggregation procedure. This drew inspiration from Schrab et al. (2023), which aggregated test statistics in MMD context.

# Open question

Our theory focused on the case where the entropic **regularization level** $\varepsilon$ is fixed.

**What can be said if $\varepsilon \to 0$ as $n \to \infty$, or if $\varepsilon = 0$?**

# Thank you!

# References I


Agarwal, M. and A. Luedtke (2026). "Sinkhorn treatment effects: A causal optimal transport measure". In: *Int Conf Mach Learn*.

Balakrishnan, S., E. Kennedy, and L. Wasserman (2025). "Conservative inference for counterfactuals". In: *Journal of Causal Inference* 13.1, p. 20230071.

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006). "What mean impacts miss: Distributional effects of welfare reform experiments". In: *American Economic Review* 96.4, pp. 988–1012.

Fawkes, J. et al. (2022). "Doubly robust kernel statistics for testing distributional treatment effects". In: *arXiv preprint arXiv:2212.04922*.

Feydy, J. et al. (2019). "Interpolating between optimal transport and mmd using sinkhorn divergences". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR, pp. 2681–2690.

Goldfeld, Z. et al. (2024). "Limit theorems for entropic optimal transport maps and Sinkhorn divergence". In: *Electronic Journal of Statistics* 18.1, pp. 980–1041.

Jain, S. and A. Luedtke (2026). "Conditional distributional treatment effects: Doubly robust estimation and testing". In: *Int Conf Mach Learn*.

# References II


Keller, H. H. (2006). *Differential calculus in locally convex spaces*. Springer.

Kokot, A. and A. Luedtke (2025). "Coreset selection for the Sinkhorn divergence and generic smooth divergences". In: *arXiv preprint arXiv:2504.20194*.

Lavenant, H. et al. (2024). "The Riemannian geometry of Sinkhorn divergences". In: *arXiv preprint arXiv:2405.04987*.

Luedtke, A., M. Carone, and M. J. van der Laan (2019). "An omnibus non-parametric test of equality in distribution for unknown functions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81.1, pp. 75–99.

Luedtke, A. and I. Chung (2024). "One-step estimation of differentiable hilbert-valued parameters". In: *The Annals of Statistics* 52.4, pp. 1534–1563.

Martinez Taboada, D., A. Ramdas, and E. Kennedy (2023). "An efficient doubly-robust test for the kernel treatment effect". In: *Advances in Neural Information Processing Systems* 36, pp. 59924–59952.

Muandet, K. et al. (2021). "Counterfactual mean embeddings". In: *Journal of Machine Learning Research* 22.162, pp. 1–71.

# References III


Ramdas, A., N. García Trillos, and M. Cuturi (2017). "On wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2, p. 47.

Ren, J.-J. and P. K. Sen (2001). "Second order Hadamard differentiability in statistical applications". In: *Journal of multivariate analysis* 77.2, pp. 187–228.

Robins, J. et al. (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and statistics: essays in honor of David A. Freedman*. Vol. 2. Institute of Mathematical Statistics, pp. 335–422.

Van der Vaart, A. (2014). "Higher order tangent spaces and influence functions". In: *Statistical Science*, pp. 679–686.

# Extra Slides

# Optimal transport maps

Also expressible via (randomized) **transport maps** from control to treatment outcomes:

$$W_2^2(P_1, P_0) = \underset{T}{\text{minimize}}\ \mathbb{E}\left[\| T(Y_0) - Y_0 \|_2^2\right] \quad \text{subject to} \quad T(Y_0) \overset{d}{=} Y_1 .$$

# Optimal transport maps

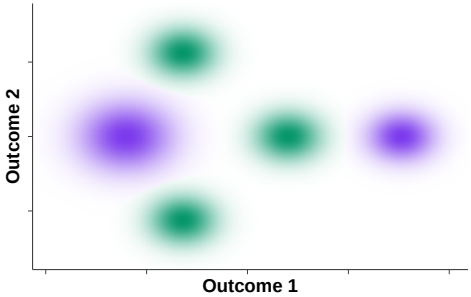
Also expressible via (randomized) transport maps from control to treatment outcomes:

$$W_2^2(P_1, P_0) = \underset{T}{\text{minimize}}\ \mathbb{E}\big[\| T(Y_0) - Y_0 \|_2^2\big] \quad \text{subject to} \quad T(Y_0) \stackrel{d}{=} Y_1 .$$

# Optimal transport maps

Also expressible via (randomized) transport maps from **control** to **treatment** outcomes:

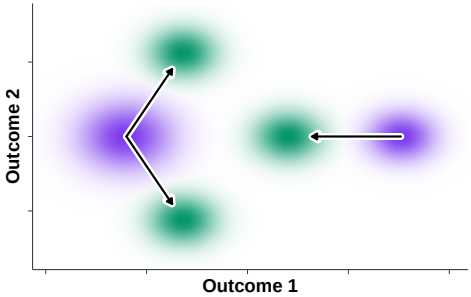
$$W_2^2(P_1, P_0) = \underset{T}{\text{minimize}}\ \mathbb{E}\left[\|T(Y_0) - Y_0\|_2^2\right] \quad \text{subject to } T(Y_0) \overset{d}{=} Y_1.$$



# Optimal transport maps

Also expressible via (randomized) transport maps from **control** to **treatment** outcomes:

$$W_2^2(P_1, P_0) = \underset{T}{\text{minimize}}\ \mathbb{E}\left[\|T(Y_0) - Y_0\|_2^2\right] \quad \text{subject to } T(Y_0) \overset{d}{=} Y_1.$$



# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$ (along a score $s$):

$$D^2\mathcal{S}_P(s,s) = D^2 \,\text{Sinkhorn}_{\mu(P)}\left(D\mu_P(s),\, D\mu_P(s)\right) \;+\; D\,\text{Sinkhorn}_{\mu(P)}\left(\, D^2\mu_P\,(s,s)\,\right).$$

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$ (along a score $s$):

$$D^2\mathcal{S}_P(s,s) = \underbrace{D^2 \,\text{Sinkhorn}_{\mu(P)}\left(D\mu_P(s),\, D\mu_P(s)\right)}_{\textbf{harmless}} + \underbrace{D\,\text{Sinkhorn}_{\mu(P)}\left(\,D^2\mu_P(s,s)\,\right)}_{\textbf{problematic}}.$$

⚠ $D^2\mu_P$ (causal curvature) usually **has no 2nd-order IF** (Robins et al., 2008).

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$ (along a score $s$):

$$D^2\mathcal{S}_P(s,s) = D^2\,\text{Sinkhorn}_{\mu(P)}\big(D\mu_P(s),\, D\mu_P(s)\big) \;+\; D\,\text{Sinkhorn}_{\mu(P)}\big(\,D^2\mu_P(s,s)\,\big)\,.$$

**Under** $H_0 : P_1 = P_0$, Sinkhorn is minimized, so its first derivative vanishes:

$$D\,\text{Sinkhorn}_{\mu(P)}(\,\cdot\,) = 0$$

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\mathrm{KME}(P_1), \mathrm{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$ (along a score $s$):

$$D^2\mathcal{S}_P(s,s) = D^2\,\text{Sinkhorn}_{\mu(P)}\left(D\mu_P(s),\, D\mu_P(s)\right) + D\,\text{Sinkhorn}_{\mu(P)}\left(\, D^2\mu_P\,(s,s)\,\right).$$

**Under** $H_0 : P_1 = P_0$, Sinkhorn is minimized, so its first derivative vanishes:

$$D\,\text{Sinkhorn}_{\mu(P)}(\,\cdot\,) = 0 \quad \Longrightarrow \quad D\,\text{Sinkhorn}_{\mu(P)}\left(D^2\mu_P(s,s)\right) = 0.$$

# What's special about the null?

$\mathcal{S} = \text{Sinkhorn} \circ \mu$, with **counterfactual mean embedding** $\mu(P) = \big(\text{KME}(P_1), \text{KME}(P_0)\big)$.

**Second-order chain rule** for $\ddot{\mathcal{S}}_P$ (along a score $s$):

$$D^2\mathcal{S}_P(s,s) = D^2\,\text{Sinkhorn}_{\mu(P)}\left(D\mu_P(s),\, D\mu_P(s)\right) \;+\; D\,\text{Sinkhorn}_{\mu(P)}\left(\, D^2\mu_P\,(s,s)\,\right).$$

**Under** $H_0 : P_1 = P_0$, Sinkhorn is minimized, so its first derivative vanishes

Only the **harmless term** survives, so a 2nd-order influence function $\ddot{\mathcal{S}}_P$ **exists under** $H_0$.

# Why the U-statistic dominates

Under $H_0 : P_1 = P_0$, our second-order estimator satisfies

$$\widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}} - \mathcal{S}(P) = \tfrac{1}{2}\mathbb{U}_n \ddot{\mathcal{S}}_P + \text{empirical process} + \text{U-process} + \text{von Mises remainder}.$$

Under conditions, **U-statistic dominates;** other three terms are each $o_p(n^{-1})$.

# Why the U-statistic dominates

Under $H_0 : P_1 = P_0$, our second-order estimator satisfies

$$\widehat{\mathbb{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_{\widehat{P}} - \mathbb{S}(P) = \tfrac{1}{2}\mathbb{U}_n \ddot{\mathbb{S}}_P + \text{empirical process} + \text{U-process} + \text{von Mises remainder}.$$

Under conditions, **U-statistic dominates;** other three terms are each $o_p(n^{-1})$.

**Lemma E.5 from the paper**

Suppose $H_0$ and

**1)** $\max_a \text{MMD}(\widehat{P}_a, P_a) = O_p(n^{-r})$ for some $r > 1/4$;

**2)** $\sup_{y \in \mathcal{Y}, a \in \{0,1\}} \int \left(\frac{\widehat{P}(a \mid x)}{P(a \mid x)} - 1\right)\left(\widehat{P}_{Y \mid a,x} - P_{Y \mid a,x}\right) k_{P_1}(\cdot, y)\, dP_X(x) = o_p(n^{-1/2})$.

Then,

$$\text{empirical process} = (P_n - P)\left(\dot{\mathbb{S}}_{\widehat{P}}(\cdot) + \int \ddot{\mathbb{S}}_{\widehat{P}}(z, \cdot)\, dP(z)\right) = o_p(n^{-1}).$$

# Why the U-statistic dominates

Under $H_0 : P_1 = P_0$, our second-order estimator satisfies

$$\widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}} - \mathcal{S}(P) = \tfrac{1}{2}\mathbb{U}_n \ddot{\mathcal{S}}_P + \text{empirical process} + \text{U-process} + \text{von Mises remainder}.$$

Under conditions, **U-statistic dominates;** other three terms are each $o_p(n^{-1})$.

**Lemma E.6 from the paper**

Suppose $H_0$ and

$$\|(I - P)^{\otimes 2} \ddot{\mathcal{S}}_{\widehat{P}} - \ddot{\mathcal{S}}_P\|_{L^2(P \otimes P)} \xrightarrow{P} 0.$$

Then, by Chebyshev's inequality for U-statistics,

$$\text{U-process} = \tfrac{1}{2}\,\mathbb{U}_n\big((I - P)^{\otimes 2}\,\ddot{\mathcal{S}}_{\widehat{P}} - \ddot{\mathcal{S}}_P\big) = o_p(n^{-1}).$$

# Why the U-statistic dominates

Under $H_0 : P_1 = P_0$, our second-order estimator satisfies

$$\widehat{\mathcal{S}}^{\text{one-step}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\mathcal{S}}_{\widehat{P}} - \mathcal{S}(P) = \tfrac{1}{2}\mathbb{U}_n \ddot{\mathcal{S}}_P + \text{empirical process} + \text{U-process} + \text{von Mises remainder}.$$

Under conditions, **U-statistic dominates;** other three terms are each $o_p(n^{-1})$.

**Lemma E.8 from the paper** ▷ *stated with slightly simpler but stronger conditions*

Suppose $H_0$ and

1) $\max_a \mathrm{MMD}(\widehat{P}_a, P_a) = O_p(n^{-r})$ for some $r > 1/3$;
2) $\max_a \int \left(\frac{P(a \mid x)}{\widehat{P}(a \mid x)} - 1\right)\left(P_{Y \mid a,x} - \widehat{P}_{Y \mid a,x}\right)\left(\upsilon_a^{(\widehat{P}_1, \widehat{P}_0)} - \upsilon_a^{(P_1, P_0)}\right) dP_X(x) = o_p(n^{-1})$;
3) $\max_a \int \left(\frac{P(a \mid x)}{\widehat{P}(a \mid x)} - 1\right)\left(P_{Y \mid a,x} - \widehat{P}_{Y \mid a,x}\right) K_{\widehat{P}_1}[\widehat{P}_1 - \widehat{P}_0]\, dP_X(x) = o_p(n^{-1})$;
4) $\max_{a,a'} \iint \left(\frac{P(a \mid x)}{\widehat{P}(a \mid x)} - 1\right)\left(\frac{P(a' \mid x')}{\widehat{P}(a' \mid x')} - 1\right)\left[\left(P_{Y \mid a,x} - \widehat{P}_{Y \mid a,x}\right) \otimes \left(P_{Y \mid a',x'} - \widehat{P}_{Y \mid a',x'}\right)\right] k_{\widehat{P}_1}\, dP_X(x)\, dP_X(x') = o_p(n^{-1})$;
5) $\sup_{s \in (0,1]} \| D^2\mathrm{Sinkhorn}[\widehat{P}_{1s}, \widehat{P}_{0s}] - D^2\mathrm{Sinkhorn}[\widehat{P}_{1s}, \widehat{P}_{1s}] \|_{\mathrm{op}} / s = o_p(n^{-1/3})$ with $\widehat{P}_{a,s} = (1-s)P_a + s\,\widehat{P}_a$.

Then,

$$\text{von Mises remainder} = \mathcal{S}(\widehat{P}) + (P - \widehat{P})\dot{\mathcal{S}}_{\widehat{P}} + \tfrac{1}{2}(P - \widehat{P})^{\otimes 2}\ddot{\mathcal{S}}_{\widehat{P}} - \mathcal{S}(P) = o_p(n^{-1}).$$