LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

Munich Center for Machine Learning

Causal Machine Learning under Privacy Constraints

ELLIS Workshop on Safe and Secure Artificial Intelligence
Online via ZOOM – 20.05.2026

Valentyn Melnychuk

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

mcml Munich Center for Machine Learning

Causal Machine Learning under Privacy Constraints

ELLIS Workshop on Safe and Secure Artificial Intelligence
Online via ZOOM – 20.05.2026

Valentyn Melnychuk Maresa Schröder Stefan Feuerriegel

Published as a conference paper at ICLR 2025

DIFFERENTIALLY PRIVATE LEARNERS FOR HETEROGENEOUS TREATMENT EFFECTS

Maresa Schröder, Valentyn Melnychuk & Stefan Feuerriegel
LMU Munich
Munich Center for Machine Learning (MCML)
{maresa.schroeder,melnychuk,feuerriegel}@lmu.de

ABSTRACT

Patient data is widely used to estimate heterogeneous treatment effects and thus understand the effectiveness and safety of drugs. Yet, patient data includes highly sensitive information that must be kept private. In this work, we aim to estimate the conditional average treatment effect (CATE) from observational data under differential privacy. Specifically, we present DP-CATE, a novel framework for CATE estimation that is *Neyman-orthogonal* and further ensures *differential privacy* of the estimates. Our framework is highly general: it applies to any two-stage CATE meta-learner with a Neyman-orthogonal loss function, and any machine learning model can be used for nuisance estimation. We further provide an extension of our DP-CATE, where we employ RKHS regression to release the complete CATE function while ensuring differential privacy. We demonstrate our DP-CATE across various experiments using synthetic and real-world datasets. To the best of our knowledge, we are the first to provide a framework for CATE estimation that is Neyman-orthogonal and differentially private.

1 INTRODUCTION

Machine learning (ML) is increasingly used for estimating treatment effects from observational data (e.g., Baiardi & Naghi, 2024; Braun & Schwartz, 2024; Ellickson et al., 2023; Feuerriegel et al., 2024). Yet, this involves sensitive information about individuals, and, hence, methods are often needed to ensure privacy.

Motivating example: *Electronic health records (EHRs) are commonly used to estimate treatment effects and thus to personalize care. Yet, EHRs capture highly sensitive data about patients (Brothers & Rothstein, 2015). Hence, many regulations, such as the US Health Insurance Portability and Accountability Act (HIPAA), mandate strong privacy guarantees for ML in medicine.*

Figure 1: **Setting: CATE estimation under DP.** Only the trusted data curator can access the data, while published CATE estimates do not allow private information about individuals to be inferred.

To ensure the privacy of information contained in the training data of ML models, multiple *privacy mechanisms* have been introduced. Arguably, the most common mechanism is *differential privacy* (DP) (Dwork, 2006; Dwork & Lei, 2009). DP builds upon the idea of injecting noise into algorithms so that sufficient information about the complete population in a dataset is kept while safeguarding sensitive information about individuals. Importantly, DP enjoys stringent theoretical guarantees and is widely used across different fields (e.g., Abadi et al., 2016; Bassily et al., 2014; Wang et al., 2019).

However, methods for treatment effect estimation under DP are scarce. Existing work has primarily focused on the *average treatment effect* (ATE) (e.g., Lee et al., 2019; Ohnishi & Awan, 2023). However, the ATE fails to capture important variations in how different subgroups or individuals respond to treatments. Therefore, many applications such as personalized medicine are interested in the *conditional average treatment effect* (CATE) (e.g., Ballmann, 2015; Feuerriegel et al., 2024).

In this paper, we estimate the CATE from observational data under DP (Fig. 1). Specifically, we propose DP-CATE, an output perturbation mechanism for Neyman-orthogonal CATE estimators that

1

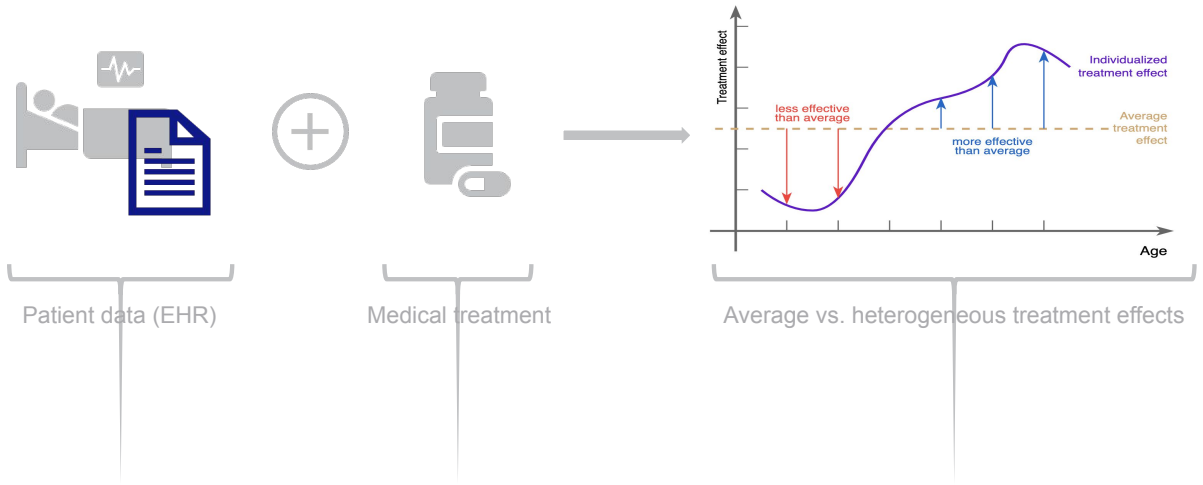
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

Why do we need differentially-private treatment effect estimation?

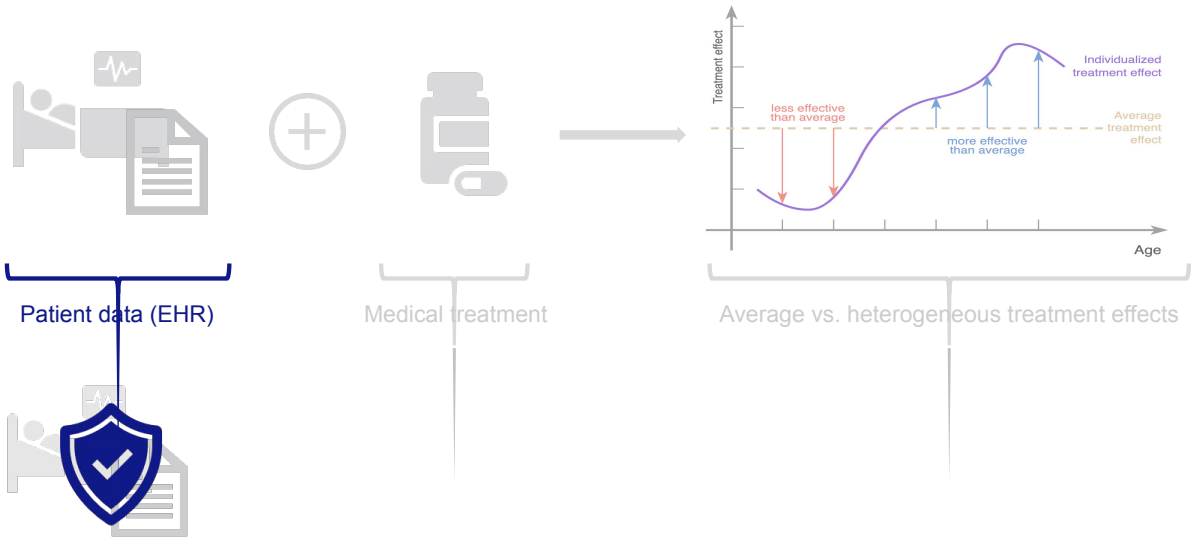
Privacy risks when estimating treatment effects from sensitive data



Patient data is widely used to estimate **heterogeneous treatment effects** and understand the effectiveness and safety of drugs

- HTEs capture important variations in how different subgroups or individuals respond to treatments
- Estimation often based on observational data, e.g., electronic health records (EHRs), containing a multitude of clinical measurements, diagnoses, and medications as well as patient demographics

Privacy risks when estimating treatment effects from sensitive data

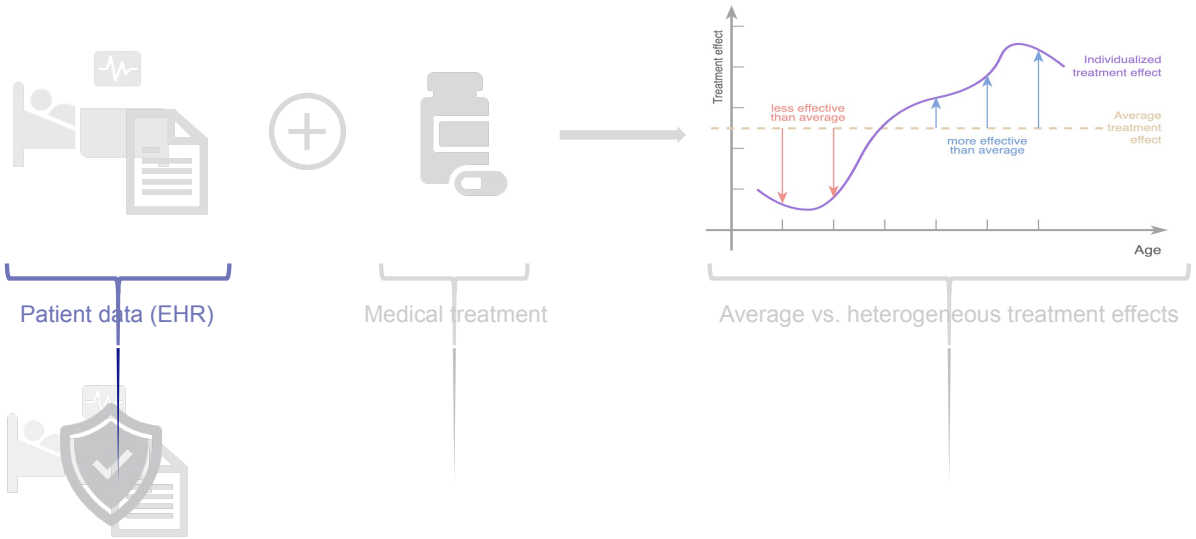


Patient data is widely used to estimate **heterogeneous treatment effects** and understand the effectiveness and safety of drugs

- HTEs capture important variations in how different subgroups or individuals respond to treatments
- Estimation often based on observational data, e.g., electronic health records (EHRs), containing a multitude of clinical measurements, diagnoses, and medications as well as patient demographics

Problem: Patient data includes highly **sensitive information**. Many regulations, such as the US Health Insurance Portability and Accountability Act (HIPAA), mandate strong privacy guarantees for ML in medicine.

Privacy risks when estimating treatment effects from sensitive data



Patient data is widely used to estimate **heterogeneous treatment effects** and understand the effectiveness and safety of drugs

- HTEs capture important variations in how different subgroups or individuals respond to treatments
- Estimation often based on observational data, e.g., electronic health records (EHRs), containing a multitude of clinical measurements, diagnoses, and medications as well as patient demographics

Problem: Patient data includes highly **sensitive information**. Many regulations, such as the US Health Insurance Portability and Accountability Act (HIPAA), mandate strong privacy guarantees for ML in medicine.

De-identification of datasets to ensure privacy?

1. **Pseudonymization:**
   Replace direct identifiers (e.g., names, IDs) with random tokens
   Problem: re-identification often possible via quasi-identifiers (e.g., ZIP code, date of birth), linkage attacks across datasets
2. **Masking:**
   Remove or partially mask data columns
   Problem: quickly destroys predictive signal, potentially misses to mask personally identifiable information, linkage attacks across datasets
3. **Generalization (k-anonymity):**
   Make each record indistinguishable from at least $k-1$ others on quasi-identifiers
   Problem: fragile against attackers with auxiliary info, hard for high-dimensional data, composition attacks across multiple k-anonymous releases

Differential privacy (DP)

Differential privacy (DP) ensures that the *inclusion or exclusion of data from any individual* does not significantly affect **any summary of the dataset** (= mechanism)

(ε, δ)- differential privacy:

A mechanism $\mathrm{f_D}: D \mapsto \mathbb{R}^{\mathrm{d}}$ on a dataset D is (ε, δ)-differentially private if, for all neighboring datasets $D, D' \in \mathrm{Z}^{\mathrm{n}}$ and all measurable $S \subseteq \mathbb{R}^{\mathrm{d}}$, it holds that

$$P\left(\mathrm{f_D} \in S\right) \leq \exp(\varepsilon) \cdot P\left(\mathrm{f_{D'}} \in S\right) + \delta.$$

- *Neighboring* datasets D and D' with Hamming distance $d_H(D, D') = 1$, denoted as $D \sim D$
- Mechanism $\hat{\mathrm{f}}_{\mathrm{D}}$ and $\hat{\mathrm{f}}_{\mathrm{D'}}$
- *Privacy budget* ε and *failure probability* δ
- DP ensures that the probability density of any summary on dataset D is ε-indistinguishable from the probability density of the same summary stemming from a neighboring dataset D' with probability of at least $1 - \delta$

Differential privacy (DP)

Differential privacy (DP) ensures that the *inclusion or exclusion of data from any individual* does not significantly affect **any summary of the dataset** (= mechanism)

(ε, δ)- differential privacy:

A mechanism $\mathrm{f_D}: D \mapsto \mathbb{R}^{\mathrm{d}}$ on a dataset D is (ε, δ)-differentially private if, for all neighboring datasets $D, D' \in \mathrm{Z}^{\mathrm{n}}$ and all measurable $S \subseteq \mathbb{R}^{\mathrm{d}}$, it holds that

$$P\,(\mathrm{f_D} \in S) \leq \exp(\varepsilon) \cdot P\,(\mathrm{f_{D'}} \in S) + \delta.$$

- *Neighboring* datasets D and D' with Hamming distance $d_H(D, D') = 1$, denoted as $D \sim D$
- Mechanism $\hat{\mathrm{f}}_{\mathrm{D}}$ and $\hat{\mathrm{f}}_{\mathrm{D}'}$
- *Privacy budget* ε and *failure probability* δ
- DP ensures that the probability density of any summary on dataset D is ε-indistinguishable from the probability density of the same summary stemming from a neighboring dataset D' with probability of at least $1 - \delta$

How can we achieve DP?

Output perturbation adds appropriately calibrated zero-centered noise (e.g., Gaussian noise) to the summary to perturb the prediction in a way that the predictions resulting from two neighboring databases cannot be differentiated

Gaussian noise mechanism:

Let $\mathrm{f}: \mathrm{D} \mapsto \mathbb{R}^{\mathrm{d}}$ be a mechanism on dataset D with l_2-sensitivity

$$\Delta_2(\mathrm{f}) = \sup_{\mathrm{D} \sim \mathrm{D}'} \|\mathrm{f_D} - \mathrm{f_{D'}}\|_2$$

and

$$\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_{\mathrm{d}}) \text{ for } \sigma \geq \frac{1}{\varepsilon} \sqrt{2 \ln(1.25/\delta)}\, \Delta_2(\mathrm{f}).$$

Then, $\mathrm{f_{DP}} = \mathrm{f_D} + \mathbf{U}$ preserves (ε, δ)-DP.

Differential privacy (DP)

Differential privacy (DP) ensures that the *inclusion or exclusion of data from any individual* does not significantly affect **any summary of the dataset** (= mechanism)

(ε, δ)- differential privacy:

A mechanism $\mathrm{f_D}: D \mapsto \mathbb{R}^{\mathrm{d}}$ on a dataset D is (ε, δ)-differentially private if, for all neighboring datasets $D, D' \in \mathrm{Z}^{\mathrm{n}}$ and all measurable $S \subseteq \mathbb{R}^{\mathrm{d}}$, it holds that

$$P\,(\mathrm{f_D} \in S) \leq \exp(\varepsilon) \cdot P\,(\mathrm{f_{D'}} \in S) + \delta.$$

- *Neighboring* datasets D and D' with Hamming distance $d_H(D, D') = 1$, denoted as $D \sim D$
- Mechanism $\hat{\mathrm{f}}_{\mathrm{D}}$ and $\hat{\mathrm{f}}_{\mathrm{D}'}$
- *Privacy budget* ε and *failure probability* δ
- DP ensures that the probability density of any summary on dataset D is ε-indistinguishable from the probability density of the same summary stemming from a neighboring dataset D' with probability of at least $1 - \delta$

How can we achieve DP?

Output perturbation adds appropriately calibrated zero-centered noise (e.g., Gaussian noise) to the summary to perturb the prediction in a way that the predictions resulting from two neighboring databases cannot be differentiated

Gaussian noise mechanism:

Let $\mathrm{f}: \mathrm{D} \mapsto \mathbb{R}^{\mathrm{d}}$ be a mechanism on dataset D with l_2-sensitivity

$$\Delta_2(\mathrm{f}) = \sup_{\mathrm{D} \sim \mathrm{D}'} \|\mathrm{f_D} - \mathrm{f_{D'}}\|_2$$

and

$$\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_{\mathrm{d}}) \text{ for } \sigma \geq \frac{1}{\varepsilon} \sqrt{2 \ln(1.25/\delta)}\, \Delta_2(\mathrm{f}).$$

Then, $\mathrm{f_{DP}} = \mathrm{f_D} + \mathbf{U}$ preserves (ε, δ)-DP.

Problem:

1) $\Delta_2(\mathrm{f})$ is difficult or even impossible to compute in general

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

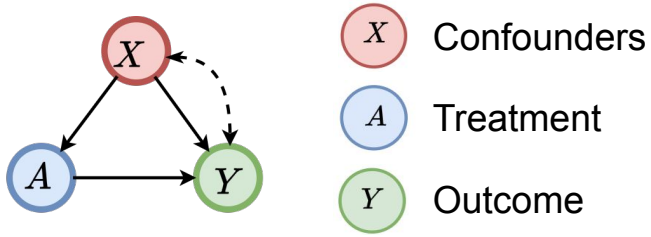
LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

Background: Efficient conditional average treatment effect estimation

Two-stage learners for conditional average treatment effect (CATE) estimation

- Dataset $\bar{D} \coloneqq \{(X_i, A_i, Y_i)\}_{i=1,\ldots,2n}$, confounders X (in bounded domain $\mathcal{X} \in \mathbb{R}^q$), binary treatment $A \in \{0,1\}$, bounded outcome $Y \in \mathcal{Y}$, $Z_i \coloneqq (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$.



X Confounders

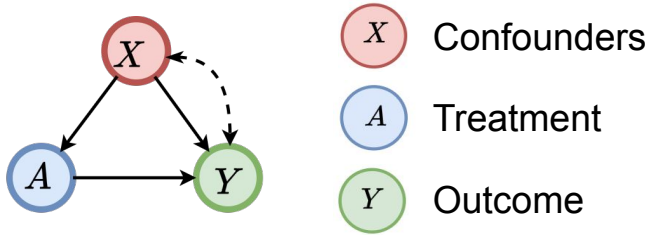
A Treatment

Y Outcome

- Propensity score $\pi(x) \coloneqq P(A = 1 | X = x)$
  Outcome function $\mu(x, a) \coloneqq \mathbb{E}[Y | X = x, A = a]$
  Potential outcome $Y(a)$

Two-stage learners for conditional average treatment effect (CATE) estimation

- Dataset $\bar{D} \coloneqq \{(X_i, A_i, Y_i)\}_{i=1,\dots,2n}$, confounders X (in bounded domain $\mathcal{X} \in \mathbb{R}^q$), binary treatment $A \in \{0,1\}$, bounded outcome $Y \in \mathcal{Y}$, $Z_i \coloneqq (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$.



- Propensity score $\pi(x) \coloneqq P(A = 1 | X = x)$
  Outcome function $\mu(x, a) \coloneqq \mathbb{E}[Y | X = x, A = a]$
  Potential outcome $Y(a)$

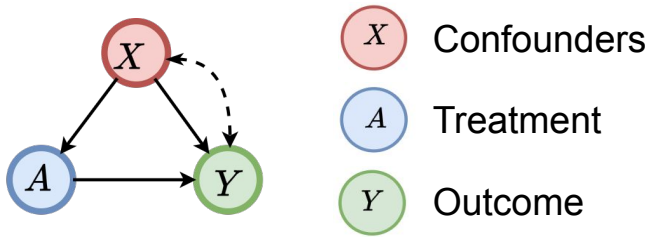
Target: Conditional average treatment effect (CATE)

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) | X = x]$$

Setting

Two-stage learners for conditional average treatment effect (CATE) estimation

- Dataset $\bar{D} \coloneqq \{(X_i, A_i, Y_i)\}_{i=1,\dots,2n}$, confounders X (in bounded domain $\mathcal{X} \in \mathbb{R}^q$), binary treatment $A \in \{0,1\}$, bounded outcome $Y \in \mathcal{Y}$, $Z_i \coloneqq (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$.



- Propensity score $\pi(x) \coloneqq P(A = 1 | X = x)$
  Outcome function $\mu(x, a) \coloneqq \mathbb{E}[Y | X = x, A = a]$
  Potential outcome $Y(a)$

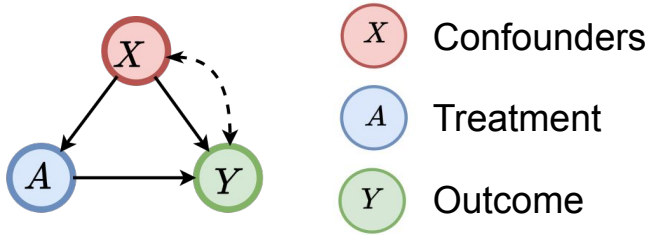
Target: Conditional average treatment effect (CATE)

$$\begin{aligned}\tau(x) &\coloneqq \mathbb{E}[Y(1) - Y(0) | X = x] \\ &= \mathbb{E}[Y | X = x, A = 1] - \mathbb{E}[Y | X = x, A = 0]\end{aligned}$$

under the standard causal assumptions

Two-stage learners for conditional average treatment effect (CATE) estimation

- Dataset $\bar{D} \coloneqq \{(X_i, A_i, Y_i)\}_{i=1,\dots,2n}$, confounders X (in bounded domain $\mathcal{X} \in \mathbb{R}^q$), binary treatment $A \in \{0,1\}$, bounded outcome $Y \in \mathcal{Y}$, $Z_i \coloneqq (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$.



X Confounders

A Treatment

Y Outcome

- Propensity score $\pi(x) \coloneqq P(A = 1 | X = x)$
  Outcome function $\mu(x, a) \coloneqq \mathbb{E}[Y | X = x, A = a]$
  Potential outcome $Y(a)$

Target: Conditional average treatment effect (CATE)

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) | X = x]$$
$$= \mathbb{E}[Y | X = x, A = 1] - \mathbb{E}[Y | X = x, A = 0]$$

under the standard causal assumptions

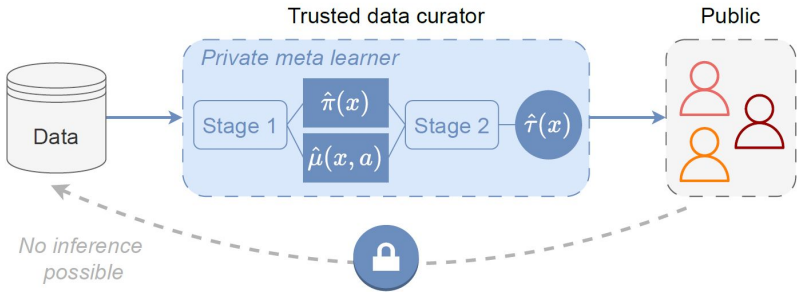
CATE meta-learners (2-stage learners)

Model-agnostic CATE estimation algorithms, that can be implemented with arbitrary machine learning algorithms

Step 1: Fit nuisance functions $\hat{\eta} = (\hat{\pi}, \hat{\mu})$

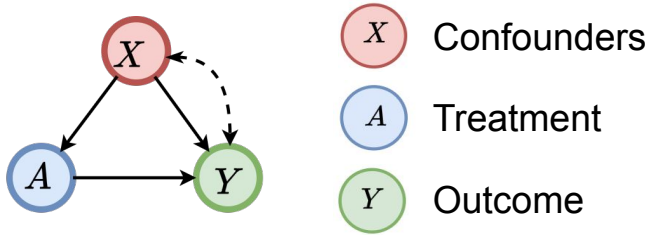
Step 2: Construct and minimize risk $R_D(g, \hat{\eta})$ s.t.,

$$\hat{\tau}(x) = \arg\min_{g \in G} R_D(g, \hat{\eta})$$



Two-stage learners for conditional average treatment effect (CATE) estimation

- Dataset $\bar{D} \coloneqq \{(X_i, A_i, Y_i)\}_{i=1,\dots,2n}$, confounders X (in bounded domain $\mathcal{X} \in \mathbb{R}^q$), binary treatment $A \in \{0,1\}$, bounded outcome $Y \in \mathcal{Y}$, $Z_i \coloneqq (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$.



- Propensity score $\pi(x) \coloneqq P(A = 1 | X = x)$
  Outcome function $\mu(x, a) \coloneqq \mathbb{E}[Y | X = x, A = a]$
  Potential outcome $Y(a)$

Target: Conditional average treatment effect (CATE)

$$\tau(x) \coloneqq \mathbb{E}[Y(1) - Y(0) | X = x]$$
$$= \mathbb{E}[Y | X = x, A = 1] - \mathbb{E}[Y | X = x, A = 0]$$

under the standard causal assumptions

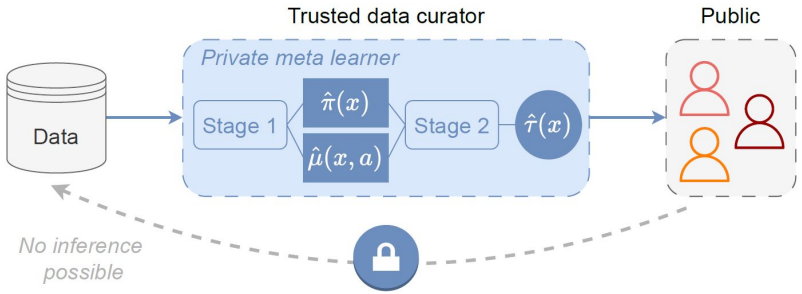
CATE meta-learners (2-stage learners)

Model-agnostic CATE estimation algorithms, that can be implemented with arbitrary machine learning algorithms

Step 1: Fit nuisance functions $\hat{\eta} = (\hat{\pi}, \hat{\mu})$

Step 2: Construct and minimize risk $R_D(g, \hat{\eta})$ s.t.,

$$\hat{\tau}(x) = \arg\min_{g \in G} R_D(g, \hat{\eta})$$



Task: *Efficiently* estimate $\tau(x)$ by
i. using *Neyman-orthogonal* meta-learners, while
ii. ensuring differential privacy in a *model-agnostic* manner

Neyman-orthogonal CATE meta-learners

General estimation framework

We aim to estimate CATE

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

through regressing the difference in the potential outcomes on X based on population risk for working model $g \in \mathcal{G}$

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\Big[\lambda(\pi(X))\big((\mu(1, X) - \mu(0, X)) - g(X)\big)^2\Big] + \Lambda(g)$$

- nuisance functions $\eta = (\mu, \pi)$
- weight function $\lambda(\cdot) > 0$
- regularization term $\Lambda(g)$

Neyman-orthogonal CATE meta-learners

General estimation framework

We aim to estimate CATE

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

through regressing the difference in the potential outcomes on X based on population risk for working model $g \in \mathcal{G}$

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\left[\lambda(\pi(X))\big((\mu(1, X) - \mu(0, X)) - g(X)\big)^2\right] + \Lambda(g)$$

- nuisance functions $\eta = (\mu, \pi)$
- weight function $\lambda(\cdot) > 0$
- regularization term $\Lambda(g)$

Problem

- $R_P(g, \eta, \lambda(\pi))$ cannot be estimated and minimized due to **unknown nuisances** $\eta = (\mu, \pi)$
- Directly employing the estimated $\hat{\mu}, \hat{\pi}$ leads to **error propagation** and **bias**

Neyman-orthogonal CATE meta-learners

General estimation framework

We aim to estimate CATE

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

through regressing the difference in the potential outcomes on X based on population risk for working model $g \in \mathcal{G}$

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\left[\lambda(\pi(X))\big((\mu(1,X) - \mu(0,X)) - g(X)\big)^2\right] + \Lambda(g)$$

- nuisance functions $\eta = (\mu, \pi)$
- weight function $\lambda(\cdot) > 0$
- regularization term $\Lambda(g)$

Problem

- $R_P(g, \eta, \lambda(\pi))$ cannot be estimated and minimized due to **unknown nuisances** $\eta = (\mu, \pi)$
- Directly employing the estimated $\hat{\mu}, \hat{\pi}$ leads to **error propagation** and **bias**

Neyman-orthogonal risk functions

- quasi-oracle efficiency through orthogonalization of the risk
- unbiasedness of the final estimator, even when one nuisance is mis-specified

Neyman-orthogonal CATE meta-learners

General estimation framework

We aim to estimate CATE

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

through regressing the difference in the potential outcomes on X based on population risk for working model $g \in \mathcal{G}$

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\left[\lambda(\pi(X))\big((\mu(1,X) - \mu(0,X)) - g(X)\big)^2\right] + \Lambda(g)$$

- nuisance functions $\eta = (\mu, \pi)$
- weight function $\lambda(\cdot) > 0$
- regularization term $\Lambda(g)$

Problem

- $R_P(g, \eta, \lambda(\pi))$ cannot be estimated and minimized due to **unknown nuisances** $\eta = (\mu, \pi)$
- Directly employing the estimated $\hat{\mu}, \hat{\pi}$ leads to **error propagation** and **bias**

Neyman-orthogonal risk functions

- quasi-oracle efficiency through orthogonalization of the risk
- unbiasedness of the final estimator, even when one nuisance is mis-specified

Orthogonal risk function

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\left[\rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2\right] + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a, \pi(x))} \frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}\big(y - \mu(x, a)\big) + \mu(x, 1) - \mu(x, 0)$$

Neyman-orthogonal CATE meta-learners

General estimation framework

We aim to estimate CATE

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

through regressing the difference in the potential outcomes on X based on population risk for working model $g \in \mathcal{G}$

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\Big[\lambda(\pi(X))\big((\mu(1,X) - \mu(0,X)) - g(X)\big)^2\Big] + \Lambda(g)$$

- nuisance functions $\eta = (\mu, \pi)$
- weight function $\lambda(\cdot) > 0$
- regularization term $\Lambda(g)$

Problem

- $R_P(g, \eta, \lambda(\pi))$ cannot be estimated and minimized due to **unknown nuisances** $\eta = (\mu, \pi)$
- Directly employing the estimated $\hat{\mu}, \hat{\pi}$ leads to **error propagation** and **bias**

Neyman-orthogonal risk functions

- quasi-oracle efficiency through orthogonalization of the risk
- unbiasedness of the final estimator, even when one nuisance is mis-specified

Orthogonal risk function

$$R_P(g, \eta, \lambda(\pi)) = \mathbb{E}\big[\rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2\big] + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a,\pi(x))}\frac{a - \pi(x)}{\pi(x)(1-\pi(x))}\big(y - \mu(x,a)\big) + \mu(x,1) - \mu(x,0)$$

Metal-learner framework

We can build various (weighted) learners through different choices of $\lambda(\cdot)$

1. **DR-Learner:** (unweighted learner)
   - $\lambda\big(\pi(x)\big) = 1$ for all x
   - Then, $\rho(a, \pi(x)) = 1$ for all x and $\phi(z, \eta, \lambda(\pi))$ equals the known pseudo-outcome of the DR-Learner
2. **R-Learner:** (overlap-weighted learner)
   - $\lambda\big(\pi(x)\big) = \pi(x)(1 - \pi(x))$
   - Then, $\rho(a, \pi(x)) = (a - \pi(x))^2$ and

     $$\phi(z, \eta, \lambda(\pi)) = \frac{y - \mu(x,a)}{a - \pi(x)} + \mu(x,1) - \mu(x,0)$$

     gives a pseudo-outcome presentation of the R-Learner

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

Our contribution: Differentially-private CATE meta-learners

Efficient CATE estimation under DP

Model-agnostic Neyman-orthogonal CATE estimation under (ε, δ)-DP

Let the dataset $\overline{D}$ be a disjoint union of the two subsets and $D, \widetilde{D}$ of size n. We employ tow-stage meta-learners that first estimate nuisance functions $\hat{\eta}_{\widetilde{D}} = (\hat{\pi}_{\widetilde{D}}, \hat{\mu}_{\widetilde{D}})$ on $\widetilde{D}$ and then minimize an adapted Neyman-orthogonal risk function

$$\hat{g}_D(\cdot;\eta) = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2 + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a,\pi(x))} \frac{a - \pi(x)}{\pi(x)(1-\pi(x))}\big(y - \mu(x,a)\big) + \mu(x,1) - \mu(x,0),$$

estimated on D with $\eta = \hat{\eta}_{\widetilde{D}}$.

Aim: Find a calibration term $r(\varepsilon, \delta, \hat{g}_D, \eta)$, such that

$$\hat{g}_{\mathrm{DP}}(\mathbf{x};\eta) = \hat{g}_D(\mathbf{x};\eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}$$

for $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ is (ε, δ)-differentially private.

Efficient CATE estimation under DP

Model-agnostic Neyman-orthogonal CATE estimation under (ε, δ)-DP

Let the dataset $\overline{D}$ be a disjoint union of the two subsets and $D, \widetilde{D}$ of size n. We employ tow-stage meta-learners that first estimate nuisance functions $\hat{\eta}_{\widetilde{D}} = (\hat{\pi}_{\widetilde{D}}, \hat{\mu}_{\widetilde{D}})$ on $\widetilde{D}$ and then minimize an adapted Neyman-orthogonal risk function

$$\hat{g}_D(\cdot;\eta) = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2 + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a, \pi(x))} \frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}\big(y - \mu(x, a)\big) + \mu(x, 1) - \mu(x, 0),$$

estimated on D with $\eta = \hat{\eta}_{\widetilde{D}}$.

Aim: Find a calibration term $r(\varepsilon, \delta, \hat{g}_D, \eta)$, such that

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \eta) = \hat{g}_D(\mathbf{x}; \eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}$$

for $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ is (ε, δ)-differentially private.

Gaussian noise mechanism:

Let $\mathrm{f}: \mathrm{D} \longmapsto \mathbb{R}^{\mathrm{d}}$ be a mechanism on dataset D with l_2-sensitivity

$$\Delta_2(\mathrm{f}) = \sup_{\mathrm{D} \sim \mathrm{D}', \mathrm{x} \in \mathcal{X}^{\mathrm{d}}} \|\mathrm{f}_{\mathrm{D}} - \mathrm{f}_{\mathrm{D}'}\|_2$$

and

$$\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_{\mathrm{d}}) \text{ for } \sigma \geq \frac{1}{\varepsilon} \sqrt{2 \ln(1.25/\delta)}\, \Delta_2(\mathrm{f}).$$

Then,

$$\mathrm{f}_{\mathrm{DP}} = \mathrm{f}_{\mathrm{D}} + \mathbf{U}$$

preserves (ε, δ)-DP.

Efficient CATE estimation under DP

Model-agnostic Neyman-orthogonal CATE estimation under (ε, δ)-DP

Let the dataset $\overline{D}$ be a disjoint union of the two subsets and $D, \widetilde{D}$ of size n. We employ tow-stage meta-learners that first estimate nuisance functions $\hat{\eta}_{\widetilde{D}} = (\hat{\pi}_{\widetilde{D}}, \hat{\mu}_{\widetilde{D}})$ on $\widetilde{D}$ and then minimize an adapted Neyman-orthogonal risk function

$$\hat{g}_D(\cdot;\eta) = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2 + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a,\pi(x))} \frac{a - \pi(x)}{\pi(x)(1-\pi(x))}\big(y - \mu(x,a)\big) + \mu(x,1) - \mu(x,0),$$

estimated on D with $\eta = \hat{\eta}_{\widetilde{D}}$.

Aim: Find a calibration term $r(\varepsilon, \delta, \hat{g}_D, \eta)$, such that

$$\hat{g}_{\mathrm{DP}}(\mathbf{x};\eta) = \hat{g}_D(\mathbf{x};\eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}$$

for $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ is (ε, δ)-differentially private.

Note: CATE is a function. However, the above **output perturbation** generally only applies to **finite-dimensional outputs**.

Efficient CATE estimation under DP

Model-agnostic Neyman-orthogonal CATE estimation under (ε, δ)-DP

Let the dataset $\overline{D}$ be a disjoint union of the two subsets and $D, \widetilde{D}$ of size n. We employ tow-stage meta-learners that first estimate nuisance functions $\hat{\eta}_{\widetilde{D}} = (\hat{\pi}_{\widetilde{D}}, \hat{\mu}_{\widetilde{D}})$ on $\widetilde{D}$ and then minimize an adapted Neyman-orthogonal risk function

$$\hat{g}_D(\cdot;\eta) = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2 + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a,\pi(x))} \frac{a - \pi(x)}{\pi(x)(1-\pi(x))}\big(y - \mu(x,a)\big) + \mu(x,1) - \mu(x,0),$$

estimated on D with $\eta = \hat{\eta}_{\widetilde{D}}$.

Aim: Find a calibration term $r(\varepsilon, \delta, \hat{g}_D, \eta)$, such that

$$\hat{g}_{\mathrm{DP}}(\mathbf{x};\eta) = \hat{g}_D(\mathbf{x};\eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}$$

for $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ is (ε, δ)-differentially private.

Note: CATE is a function. However, the above **output perturbation** generally only applies to **finite-dimensional outputs**.

Two distinct use-cases

1. **Finite queries:** We aim to report a number d of CATE estimates (e.g., treatment effects across different age groups)

Method outline

Efficient CATE estimation under DP

Model-agnostic Neyman-orthogonal CATE estimation under (ε, δ)-DP

Let the dataset $\overline{D}$ be a disjoint union of the two subsets and $D, \widetilde{D}$ of size n. We employ tow-stage meta-learners that first estimate nuisance functions $\hat{\eta}_{\widetilde{D}} = (\hat{\pi}_{\widetilde{D}}, \hat{\mu}_{\widetilde{D}})$ on $\widetilde{D}$ and then minimize an adapted Neyman-orthogonal risk function

$$\hat{g}_D(\cdot;\eta) = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2 + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a,\pi(x))} \frac{a - \pi(x)}{\pi(x)(1-\pi(x))}\big(y - \mu(x,a)\big) + \mu(x,1) - \mu(x,0),$$

estimated on D with $\eta = \hat{\eta}_{\widetilde{D}}$.

Aim: Find a calibration term $r(\varepsilon, \delta, \hat{g}_D, \eta)$, such that

$$\hat{g}_{\mathrm{DP}}(\mathbf{x};\eta) = \hat{g}_D(\mathbf{x};\eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}$$

for $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ is (ε, δ)-differentially private.

Note: CATE is a function. However, the above **output perturbation** generally only applies to **finite-dimensional outputs**.

Two distinct use-cases

1. **Finite queries:** We aim to report a number d of CATE estimates (e.g., treatment effects across different age groups)
2. **Functional queries:** We release an estimate $\hat{g}_{\mathrm{DP}}$ of the complete CATE function τ, which can then be queried arbitrarily often (e.g, as in clinical decision support systems)

Efficient CATE estimation under DP

Model-agnostic Neyman-orthogonal CATE estimation under (ε, δ)-DP

Let the dataset $\overline{D}$ be a disjoint union of the two subsets and $D, \widetilde{D}$ of size n. We employ tow-stage meta-learners that first estimate nuisance functions $\hat{\eta}_{\widetilde{D}} = (\hat{\pi}_{\widetilde{D}}, \hat{\mu}_{\widetilde{D}})$ on $\widetilde{D}$ and then minimize an adapted Neyman-orthogonal risk function

$$\hat{g}_D(\cdot;\eta) = \arg\min_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \rho\big(A_i, \pi(X_i)\big)\big(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i)\big)^2 + \Lambda(g)$$

with

$$\rho(a, \pi(x)) \coloneqq \big(a - \pi(x)\big)\lambda'\big(\pi(x)\big) + \lambda(\pi(x))$$

and

$$\phi(z, \eta, \lambda(\pi)) \coloneqq \frac{\lambda(\pi(x))}{\rho(a, \pi(x))} \frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}\big(y - \mu(x, a)\big) + \mu(x, 1) - \mu(x, 0),$$

estimated on D with $\eta = \hat{\eta}_{\widetilde{D}}$.

Aim: Find a calibration term $r(\varepsilon, \delta, \hat{g}_D, \eta)$, such that

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \eta) = \hat{g}_D(\mathbf{x}; \eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}$$

for $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$ is (ε, δ)-differentially private.

Note: CATE is a function. However, the above **output perturbation** generally only applies to **finite-dimensional outputs**.

Two distinct use-cases

1. **Finite queries:** We aim to report a number d of CATE estimates (e.g., treatment effects across different age groups)
2. **Functional queries:** We release an estimate $\hat{g}_{\mathrm{DP}}$ of the complete CATE function τ, which can then be queried arbitrarily often (e.g, as in clinical decision support systems)

Advantages:

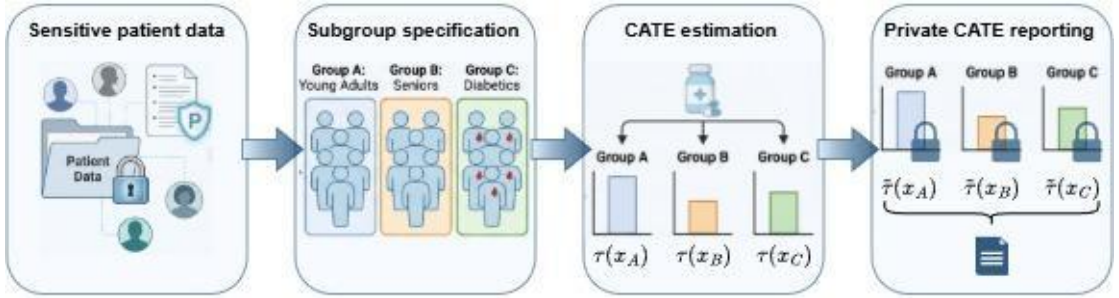
1. Highly **flexible** approach that can be combined with all weighted Neyman-orthogonal two-stage CATE learners
2. **Model-agnostic** approach that can be used with various ML models as base learners in both stages
3. Retains the **quasi-oracle efficiency** of the original CATE learner

Method – Case 1

Finite number of queries

Example: Reporting research findings about medical studies that involve sensitive data requires that finitely many CATE values are estimated, such as treatment effects of a drug for various patient subgroups.

Setting: Total number of d CATE estimates (d known a-priori)
⇒ we rewrite the d separate CATE estimates as a d -dimensional vector



Finite number of queries

Example: Reporting research findings about medical studies that involve sensitive data requires that finitely many CATE values are estimated, such as treatment effects of a drug for various patient subgroups.

Setting: Total number of d CATE estimates (d known a-priori)
⇒ we rewrite the d separate CATE estimates as a d -dimensional vector

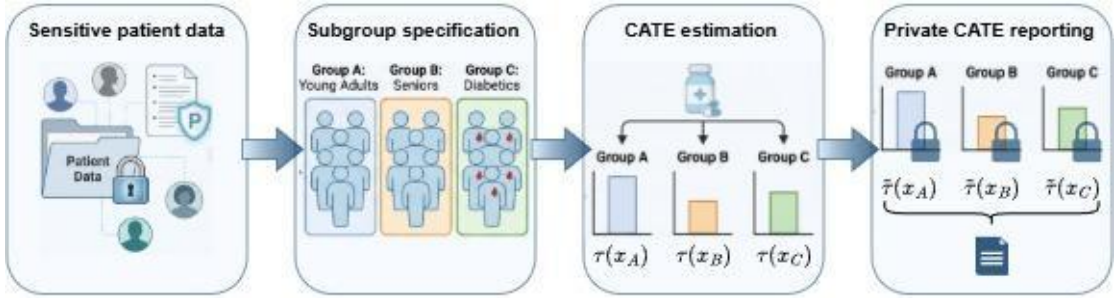
Intuition: The **influence function (IF)** allows us to quantify how a single observation in influences the estimation and the model output. Intuitively, the IF describes the **effect of an infinitesimally small perturbation** of the input z on the model output.

Observation: The sensitivity $\Delta_2(\hat{f})$ of the second-stage model necessary for applying the Gaussian mechanism can be **upper bounded by the gross-error sensitivity (GES)** of the second-stage regression
⇒ employ GES **to ensure DP**

Definition: *Let T be a functional of a distribution that defines the parameter of interest, $T = T(P) \in \mathbb{R}^d$. The* gross-error sensitivity *of T at z under P is given by the supremum of the l_2-norm of the IF of T at z at P, i.e.,*

$$\gamma(T, P) := \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; P)\|_2,$$

where $\mathrm{IF}(z, T; P) = \frac{d}{dt}[T((1-t)P + t\delta_z)]\Big|_{t=0}$, *and δ_z the Dirac-delta function.*



Method – Case 1

Finite number of queries

Example: Reporting research findings about medical studies that involve sensitive data requires that finitely many CATE values are estimated, such as treatment effects of a drug for various patient subgroups.

Setting: Total number of d CATE estimates (d known a-priori)
⇒ we rewrite the d separate CATE estimates as a d -dimensional vector

Intuition: The **influence function (IF)** allows us to quantify how a single observation in influences the estimation and the model output. Intuitively, the IF describes the **effect of an infinitesimally small perturbation** of the input z on the model output.

Observation: The sensitivity $\Delta_2(\hat{f})$ of the second-stage model necessary for applying the Gaussian mechanism can be **upper bounded by the gross-error sensitivity (GES)** of the second-stage regression
⇒ employ GES **to ensure DP**

Definition: *Let T be a functional of a distribution that defines the parameter of interest, $T = T(P) \in \mathbb{R}^d$. The* gross-error sensitivity *of T at z under P is given by the supremum of the l_2-norm of the IF of T at z at P, i.e.,*

$$\gamma(T, P) \coloneqq \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; P)\|_2,$$

where $\mathrm{IF}(z, T; P) = \frac{d}{dt}[T((1-t)P + t\delta_z)]\Big|_{t=0}$, *and δ_z the Dirac-delta function.*

Theorem: *Let $\hat{\eta}_{\mathrm{DP}} = (\hat{\pi}_{\mathrm{DP}}, \hat{\mu}_{\mathrm{DP}})$ denote the nuisance functions estimated on $\tilde{D}$ in an $(\varepsilon/2, \delta/2)$-differentially private manner of choice (for functions). We define*

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) \coloneqq \hat{g}_D(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) + \gamma(T, D) \cdot c(\varepsilon, \delta, n) \cdot \mathbf{U},$$
$$T(P) = g^*(\mathbf{x}; \hat{\eta}_{DP}),$$

$\gamma(T, P) = \sup_{z \in \mathcal{Z}} \left\| h(g^*, \boldsymbol{x}, z, \hat{\eta}_{\mathrm{DP}}) \rho\left(a, \hat{\pi}_{\mathrm{DP}}(x)\right) \left(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(x))) - g^*(x; \hat{\eta}_{\mathrm{DP}})\right) \right\|_2$,

where $\gamma(T, D)$ is the sample gross-error-sensitivity, $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$, $c(\varepsilon, \delta, n) \coloneqq$

$\frac{5\sqrt{2\ln(n)\ln\left(\frac{2}{\delta}\right)}}{\varepsilon n}$ *and where $h(g^*, \boldsymbol{x}, z, \hat{\eta}_{\mathrm{DP}}) \in \mathbb{R}^d$, $g^*(\cdot; \hat{\eta}_{\mathrm{DP}})$ depend on the second-stage ML model . Then, $\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ is (ε, δ)-differentially private.*

Method – Case 1

Finite number of queries

Example: Reporting research findings about medical studies that involve sensitive data requires that finitely many CATE values are estimated, such as treatment effects of a drug for various patient subgroups.

Setting: Total number of d CATE estimates (d known a-priori)
⇒ we rewrite the d separate CATE estimates as a d -dimensional vector

Intuition: The **influence function (IF)** allows us to quantify how a single observation in influences the estimation and the model output. Intuitively, the IF describes the **effect of an infinitesimally small perturbation** of the input z on the model output.

Observation: The sensitivity $\Delta_2(\hat{f})$ of the second-stage model necessary for applying the Gaussian mechanism can be **upper bounded by the gross-error sensitivity (GES)** of the second-stage regression
⇒ employ GES **to ensure DP**

Definition: *Let T be a functional of a distribution that defines the parameter of interest, $T = T(P) \in \mathbb{R}^d$. The* gross-error sensitivity *of T at z under P is given by the supremum of the l_2-norm of the IF of T at z at P, i.e.,*

$$\gamma(T, P) \coloneqq \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; P)\|_2,$$

where $\mathrm{IF}(z, T; P) = \frac{d}{dt}[T((1-t)P + t\delta_z]\Big|_{t=0}$, *and δ_z the Dirac-delta function.*

Theorem: *Let $\hat{\eta}_{\mathrm{DP}} = (\hat{\pi}_{\mathrm{DP}}, \hat{\mu}_{\mathrm{DP}})$ denote the nuisance functions estimated on $\tilde{D}$ in an $(\varepsilon/2, \delta/2)$-differentially private manner of choice (for functions). We define*

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) \coloneqq \hat{g}_D(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) + \gamma(T, D) \cdot c(\varepsilon, \delta, n) \cdot \mathbf{U},$$
$$T(P) = g^*(\mathbf{x}; \hat{\eta}_{DP}),$$
$$\gamma(T, P) = \sup_{z \in \mathcal{Z}} \left\| h(g^*, \boldsymbol{x}, z, \hat{\eta}_{\mathrm{DP}}) \rho\left(a, \hat{\pi}_{\mathrm{DP}}(x)\right) \left(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(x))) - g^*(x; \hat{\eta}_{\mathrm{DP}})\right) \right\|_2,$$

where $\gamma(T, D)$ is the sample gross-error-sensitivity, $\mathbf{U} \sim \mathcal{N}(0, \sigma \mathbf{I}_d)$, $c(\varepsilon, \delta, n) \coloneqq \frac{5\sqrt{2\ln(n)\ln\left(\frac{2}{\delta}\right)}}{\varepsilon n}$ and where $h(g^*, \boldsymbol{x}, z, \hat{\eta}_{\mathrm{DP}}) \in \mathbb{R}^d$, $g^*(\cdot; \hat{\eta}_{\mathrm{DP}})$ depend on the second-stage ML model . Then, $\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ is (ε, δ)-differentially private.

Neyman-orthogonality and quasi-oracle efficiency

Privatization of the second-stage model preserves **Neyman-orthogonality**

$$\begin{aligned} &\|g^*(\cdot, \eta) - \hat{g}_{DP}(\cdot, \hat{\eta}_{\mathrm{DP}})\|_{L_2}{}^2 \\ &\lesssim R_P(g^*(\cdot, \hat{\eta}_{\mathrm{DP}}), \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) - R_P(g^*(\cdot, \eta), \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) + R_2(\hat{\eta}_{\mathrm{DP}}, \eta) \\ &+ \|g^*(\cdot, \hat{\eta}_{\mathrm{DP}}) - \hat{g}_D(\cdot, \hat{\eta}_{\mathrm{DP}})\|_{L_2}{}^2 + o_P(n^{-1}) \end{aligned}$$

Under additional regularity conditions on the privatization of the nuisance functions(e.g., gradient perturbation), **quasi-oracle efficiency** is achieved. If the original estimation of the nuisance functions is at rate of at least $o_P(n^{-1/4})$, the privatized estimation preserves this rate.

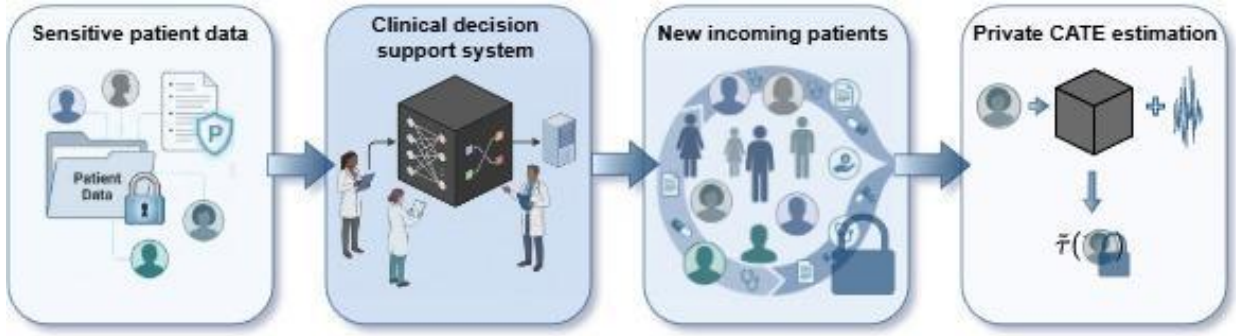
Method – Case 2

Complete CATE functions

Example: Medical researchers may want to have access to the complete CATE function. This is relevant when deploying a CATE function in clinical decision support systems where predictions about treatment effects are made for every incoming patient.

Setting: Privately release an estimate $\hat{g}_{\mathrm{DP}}(\cdot)$ of the complete CATE function $\tau(\cdot)$

⇒ Output perturbation not directly applicable



Complete CATE functions

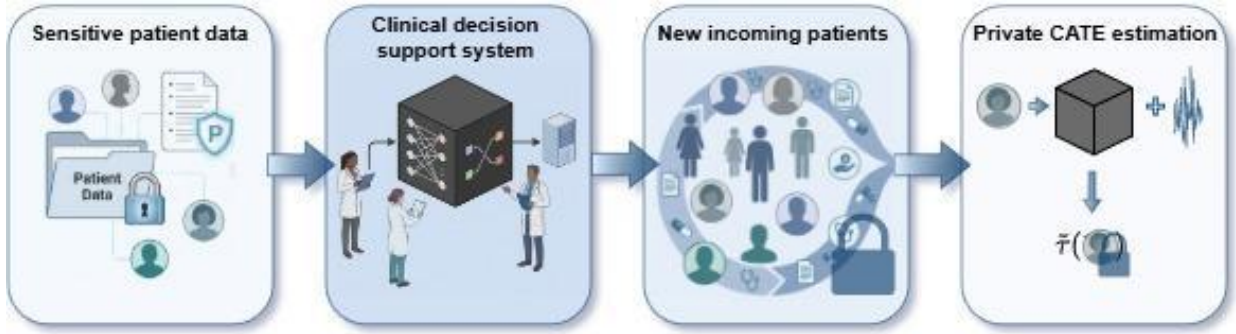
Example: Medical researchers may want to have access to the complete CATE function. This is relevant when deploying a CATE function in clinical decision support systems where predictions about treatment effects are made for every incoming patient.

Setting: Privately release an estimate $\hat{g}_{\mathrm{DP}}(\cdot)$ of the complete CATE function $\tau(\cdot)$
⇒ Output perturbation not directly applicable

Intuition: Need to find
i. a type of noise
ii. a calibration function that does not depend on the output dimension
⇒ The noise itself should be a function

Observation: The above can be addressed by the following
i. Employing a calibrated Gaussian process (GP) to privatize the CATE function
ii. If $\hat{g}_D(\cdot;\eta)$ lies in an RKHS, the GP noise can be calibrated wrt. the RKHS norm.

⇒ To ensure that $\hat{g}_D(\cdot;\eta)$ lies in an RKHS, the second-stage estimation in DP-CATE can be modelled as, e.g., a Gaussian kernel regression.



Method – Case 2

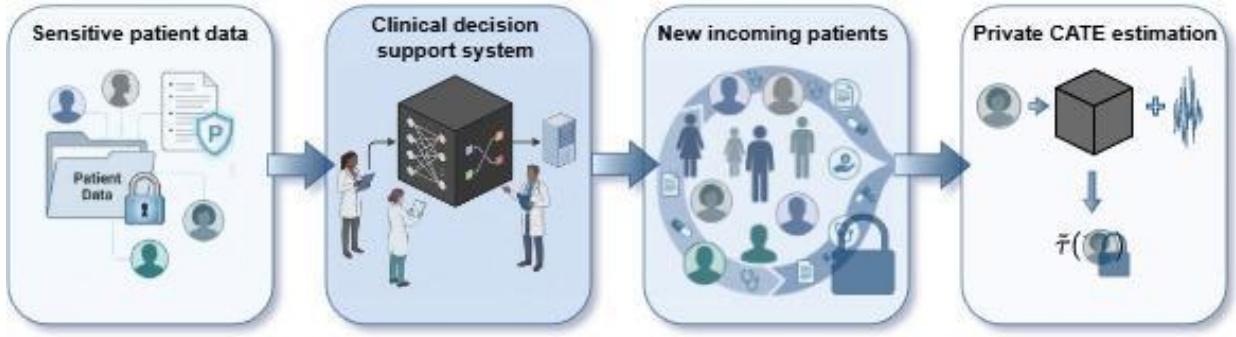
Complete CATE functions

Example: Medical researchers may want to have access to the complete CATE function. This is relevant when deploying a CATE function in clinical decision support systems where predictions about treatment effects are made for every incoming patient.

Setting: Privately release an estimate $\hat{g}_{\mathrm{DP}}(\cdot)$ of the complete CATE function $\tau(\cdot)$
⇒ Output perturbation not directly applicable

Intuition: Need to find
i. a type of noise
ii. a calibration function that does not depend on the output dimension
⇒ The noise itself should be a function

Observation: The above can be addressed by the following
i. Employing a calibrated Gaussian process (GP) to privatize the CATE function
ii. If $\hat{g}_D(\cdot;\eta)$ lies in an RKHS, the GP noise can be calibrated wrt. the RKHS norm.
⇒ To ensure that $\hat{g}_D(\cdot;\eta)$ lies in an RKHS, the second-stage estimation in DP-CATE can be modelled as, e.g., a Gaussian kernel regression.



Theorem: *Let $\hat{\eta}_{\mathrm{DP}} = (\hat{\pi}_{\mathrm{DP}}, \hat{\mu}_{\mathrm{DP}})$ denote the nuisance functions estimated on $\tilde{D}$ in an $(\varepsilon/2, \delta/2)$-differentially private manner of choice and let $x \in \mathcal{X} \subseteq \mathbb{R}^q$. Let $\mathcal{H}$ denote the RKHS induced by the kernel $K(x,x') = (\sqrt{2\pi}h)^{-q}\exp\left(\frac{-\|x-x'\|_2^2}{2h^2}\right)$ and let $l(\cdot,\cdot)$ be a convex and Lipschitz loss function with constant L. We define $\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}})$ as the second-stage regression model via*

$$\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) = \arg\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \rho\big(A_i, \hat{\pi}_{\mathrm{DP}}(X_i)\big) l\big(g(X_i), \phi(Z_i, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(X_i)))\big)^2 + \lambda \|g\|_{\mathcal{H}}^2.$$

Furthermore, let $U(\cdot) \in \mathcal{H}$ be the sample path of a zero-centred GP with covariance function $K(x,x')$. Then, (ε,δ)-DP is guaranteed by

$$\hat{g}_{\mathrm{DP}}(\cdot;\hat{\eta}_{\mathrm{DP}}) \coloneqq \hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) + \sup_{(a,x)\in\{0,1\}\times\mathcal{X}} \left[\rho\big(a, \hat{\pi}_{\mathrm{DP}}(x)\big)\right] \cdot \frac{4L\sqrt{2\ln(2/\delta)}}{(\sqrt{2\pi}h)^q \lambda \varepsilon n} \cdot U(\cdot).$$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF AI IN MANAGEMENT

mcml Munich Center for Machine Learning

Thank you for attention!

PDF of the paper @ ICLR 2025

Valentyn Melnychuk

(Soon to be) PostDoc @
Institute for AI in Management
LMU Munich