DIFFERENTIALLY PRIVATE LEARNERS FOR HETEROGENEOUS TREATMENT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Patient data is widely used to estimate heterogeneous treatment effects and 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 *doubly robust* 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 doubly robust 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 doubly robust 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). 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 involve 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 machine learning in medicine.*

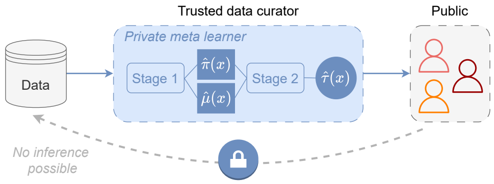


Figure 1: **Setting: CATE estimation under DP.** Only the trusted data curator can access the data, while published CATE estimates do not allow private information about individuals to be inferred.

To ensure the privacy of information contained in the training data of machine learning 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 nowadays widely used across different fields in machine learning (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 aim at CATE estimation from observational data under DP (see Fig. 1). Specifically, we propose DP-CATE, an output perturbation mechanism for doubly robust CATE estimation that satisfies differential privacy. Doubly robust estimators are generally preferred over standard CATE estimators, as they remain consistent even when either the outcome or the treatment selection model (i.e., the propensity model) is not correctly specified (e.g., Morzywolek et al., 2023).[1] Our DP framework is highly flexible and can be combined with all weighted doubly robust two-stage CATE learners, such as the R-learner (Nie & Wager, 2020). Further, our framework is model-agnostic and can be used with any ML model as a baselearner. To the best of our knowledge, we are the first to provide a framework for doubly robust CATE estimation under DP. DP-CATE is designed for two use cases relevant in practice:

① *Finitely-many queries:* 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 characteristics. In this setting, we treat the different CATE estimates as a potentially high-dimensional vector, for which we derive DP guarantees. Interestingly, we later employ a largely unexplored connection between model robustness and privacy, which allows us to base our DP-CATE for the doubly robust meta-learners on efficient influence functions (→ our Theorem 1).

② *Functional query:* 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. Hence, this requires querying the CATE function a large number of times but where the exact number is a priori unknown. In this case, the respective CATE vector would have an infinite dimension, and, as a result, privately releasing the complete CATE function cannot be performed in the same manner as in the first use case. As a remedy, we derive a tailored privacy framework for functional queries, where we make use of tools from functional analysis to calibrate a Gaussian process, which we then add to the CATE function estimated through RKHS regression (→ our Theorem 2).

Why is it non-trivial to derive privacy mechanisms for CATE estimation? Common DP strategies include perturbations of either the data, model, or output (e.g., Abadi et al., 2016; Chaudhuri et al., 2011). Yet, a naïve application of such perturbations would naturally violate causal assumptions or lead to CATE estimates that are biased. Furthermore, the CATE is an unobservable, functional quantity. However, common privatization mechanisms are only developed for vector-valued quantities. Thus, it is **not** possible follow the standard procedure of adding calibrated noise to the algorithm. Rather, we have to derive a novel, non-trivial framework that is tailored to our setting.

Our contributions:[2] (1) We propose a novel framework for CATE estimation that is differentially private and doubly robust. (2) We extend our framework to privately release both CATE estimates and even the complete CATE function. (3) We demonstrate our proposed framework for differentially private CATE estimation in experiments across various datasets.

2 Related work

We provide a brief overview of the different literature streams relevant to our work, namely, (i) CATE estimation, (ii) differential privacy, and (iii) works that adapt DP to treatment effect estimation.

CATE estimation: Popular methods for estimating CATE from observational data are the doubly robust meta-learners, such as the DR-learner (Kennedy, 2023a; van der Laan, 2006) and the R-learner (Nie & Wager, 2020). A strength of meta-learners is that these are model-agnostic approaches and can thus be instantiated with arbitrary machine learning models (e.g., neural networks). The learners have several additional benefits: (i) They are robust to model-misspecification of the first-stage nuisance estimators due to the Neyman-orthogonal loss function, which ensures first-order insensitivity to small perturbations. Hence, the meta-learners are also said to be debiased. (ii) The learners achieve quasi-oracle efficiency, even if the nuisance functions are estimated at slower rates.[3] As a result, the doubly robust estimators are asymptotically equivalent to the oracle estimator (= the one that has access to the oracle nuisance functions), thereby mitigating the finite-sample bias arising from the miss-specification of the nuisance functions (Mackey et al., 2018; Morzywolek et al., 2023). (iii) Additionally, the R-Learner is insensitive to slight overlap violations due to its inherent weighting of the pseudo-outcomes.

[1] For an introduction to the double-robustness property, see Section 3.1.

[2] All codes are available via our GitHub repository

[3] Informally, quasi-oracle efficiency means that the target model is learned almost equally well with either the estimated nuisance functions or the ground truth.

In this paper, we thus focus on differential privacy for doubly robust meta-learners. Throughout the paper, we derive our DP-CATE framework for the R-learner due to the above benefits. We later provide an extension to the DR-learner in Supplement B.

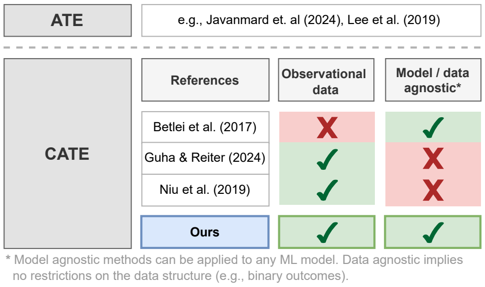
Differential privacy: DP ensures that the release of aggregated results does not reveal information about individual data samples, typically with strict theoretical results (Dwork, 2006; Dwork & Lei, 2009). As a result, DP has been employed in various fields of machine learning (e.g., Abadi et al., 2016; Bassily et al., 2014; Wang et al., 2019), but typically *outside* of CATE estimation. We discuss different strategies on how DP can be achieved in Supplement A.

In our setting, we later adopt output perturbation to CATE estimation. Output perturbation has two clear advantages in our task: (i) it can be applied to *any* machine learning model after training, which naturally fits the idea of meta-learners from above as model-agnostic approaches; and (ii) it leaves the original objective and the data unchanged. The latter is crucial because changes to the input or the objective (i.e., the estimand) arguably could violate causal assumptions and lead to biased results, respectively. However, an off-the-shelf application of output perturbation to CATE estimation is *not* possible due to various challenges (see Sec. 3.4); rather, a tailored approach must be derived for output perturbation of CATE estimation in order for privacy guarantees to hold.

DP in treatment effect estimation: The existing literature on differentially private methods for treatment effect estimation is sparse. We provide an overview in Fig. 2, where we group prior works by the underlying estimand: • *Average treatment effect (ATE).* Many works focus on privately estimating the ATE (e.g., Javanmard et al., 2024; Lee et al., 2019; Yao et al., 2024). Yet, the ATE is a much simpler causal quantity than the CATE. The ATE makes population-wide estimates and, therefore, unlike the CATE, does *not* allow to make individualized predictions about treatment effects.

• *Conditional average treatment effect (CATE).* The few works aimed at DP for CATE estimation have clear *limitations* (Fig. 2): (i) they are either restricted to interventional data from a randomized control trial and thus *not* applicable to observational data (Betlei et al., 2021); (ii) are *only* applicable to binary outcomes (Guha & Reiter, 2024); or (iii) require special private base learners for running the overall CATE estimation algorithm and are thus *not* model-agnostic (Niu et al., 2022). In particular, the latter work (Niu et al., 2022) is restricted to explainable boosting machines. Therefore, it is *not* applicable to other explainable models (e.g., linear regression or regression trees). Furthermore, it is ***not*** applicable to neural networks. In sum, none of the above methods provides a method for CATE estimation under DP where both observational data and different ML models can be used.

Research gap: So far, a DP framework for CATE estimation from observational data with general meta-learners is missing. We are thus the first to propose a framework for CATE estimation that is adheres to DP and doubly robust.



ATE	e.g., Javanmard et. al (2024), Lee et al. (2019)		
CATE	References	Observational data	Model / data agnostic*
	Betlei et al. (2017)	✗	✓
	Guha & Reiter (2024)	✓	✗
	Niu et al. (2019)	✓	✗
	Ours	✓	✓

* Model agnostic methods can be applied to any ML model. Data agnostic implies no restrictions on the data structure (e.g., binary outcomes).

Figure 2: Comparison of relevant literature.

3 Problem Formulation

Notation: Throughout our work, we write random variables in capital letters X with realizations x. We denote the probability distribution over X by P_X, where we omit the subscript whenever it is apparent from the context. We denote the probability mass function by $P(x) = P(X = x)$ for discrete X and the probability density function w.r.t. the Lebesgue measure by $p(x)$. We base our analysis on the potential outcomes framework (Rubin, 2005) and denote the potential outcome of intervention a by $Y(a)$.

Setting: We consider a dataset $D := \{(X_i, A_i, Y_i)\}_{i=1,\dots,n}$, consisting of observed confounders X in a bounded domain $\mathcal{X}$, a binary treatment $A \in \{0, 1\}$, and a bounded outcome $Y \in \mathcal{Y}$, where $Z_i := (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$, and $\mathcal{X}, \mathcal{Y}$ have bounded domains. Note that Y can be discrete or continuous. Let $\pi(x) := P(A = 1 \mid X = x)$ define the propensity score and $\mu(x, a) := \mathbb{E}[Y \mid X = x, A = a]$ the outcome function.

Estimand: Our objective is to estimate the conditional average treatment effect (CATE) $\tau(x)$ for specific groups of samples with covariates $X = x$: We make the standard assumptions for causal treatment effect estimation: positivity, consistency, and unconfoundedness (e.g., Curth & van der

Schaar, 2021; Feuerriegel et al., 2024; Rubin, 2005).[4] Then, CATE is identifiable as

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mu(x,1) - \mu(x,0). \tag{1}$$

In our work, we aim to estimate $\tau(x)$ by (i) using doubly robust meta-learners (Sec. 3.1) and (ii) by ensuring differential privacy (Sec. 3.2), which we briefly review in the following.

3.1 Doubly robust meta-learners for CATE estimation

To estimate the CATE from Equation 1, the common approach is to regress the difference in the potential outcomes $Y(1) - Y(0)$ on the confounders X. Thus, one considers the population risk for a working model $g \in \mathcal{G} : \mathcal{X} \mapsto \mathbb{R}$

$$L_P(g, \lambda(\pi)) = \mathbb{E}[\lambda(\pi(X))\,((\mu(X,1) - \mu(X,0)) - g(X))^2] \tag{2}$$

with respect to a *weight function* $\lambda(\cdot)$ of the propensity score to obtain the minimizer g^* over the L_2 Hilbert space (Hirano et al., 2003; Morzywolek et al., 2023). However, L_P cannot be directly estimated and, subsequently, minimized given the data D, as it depends on the unknown nuisance functions, π and μ. We can employ the estimated nuisance functions, $\hat{\pi}$ and $\hat{\mu}$, but then, their estimation errors propagate into the errors of estimation and, thus, minimization of L_P.

A popular approach to circumvent the above problem is to use doubly robust meta-learners. Formally, such meta-learners operate in two stages (e.g., Kennedy, 2023b; Nie & Wager, 2020). In the first stage, the meta-learners estimate *nuisance functions* $\hat{\eta} = (\hat{\pi}, \hat{\mu})$, and, in the second stage, we minimize the adapted Neyman-orthogonal population risk function

$$L_P(g, \eta, \lambda(\pi)) = \frac{1}{\mathbb{E}[\lambda(\pi(X))]}\mathbb{E}\left[\rho(A, \pi(X))\,(\phi(Z, \eta, \lambda(\pi(X))) - g(X))^2\right] \tag{3}$$

with $$\rho(a, \pi) := (a - \pi(x))\,\lambda^{'}(\pi(x)) + \lambda(\pi(x)) \quad \text{and} \tag{4}$$

$$\phi(z, \eta, \lambda(\pi)) := \frac{\lambda(\pi(x))}{\rho(a, \pi(x))}\,\frac{a - \pi(x)}{\pi(x)\,(1 - \pi(x))}(y - \mu(x,a)) + \mu(x,1) - \mu(x,0), \tag{5}$$

where $\hat{\eta}$ is used in place of η (Morzywolek et al., 2023). In the following, we denote the estimated population risk, L_P, as a loss dependent on the data D, L_D.

Later, we use the R-learner Nie & Wager (2020), which is given by $\lambda^{\mathrm{R}}(\pi) = \pi(x)\,(1 - \pi(x))$, due to its theoretical advantages (e.g., Neyman-orthogonality and oracle-efficiency). Importantly, doubly robust meta-learners such as the R-learner achieve state-of-the-art performance (Curth & van der Schaar, 2021), and the property of doubly robust makes the models insensitive to misspecification (e.g., Curth & van der Schaar, 2021; Kennedy, 2023b).

3.2 Differential privacy

Differential privacy (DP) ensures that the inclusion or exclusion of data from any individual data does not significantly affect the estimated outcome Dwork (2006); Dwork & Lei (2009). For a given *privacy budget* ε, the notion probability density of any outcome y on dataset $D \in \mathcal{Z}^n$ is ε-*indistinguishable* from the probability density of the same outcome y stemming from a neighboring dataset $D^{'} \in \mathcal{Z}^n$ with a probability of at least $1 - \delta$. The datasets D and $D^{'}$ are called *neighbors* if their Hamming distance equals one, i.e., $d_{\mathrm{H}}(D, D^{'}) = 1$. We will refer neighboring $D, D^{'}$ as $D \sim D^{'}$.

Definition 1 (Differential privacy (Dwork & Lei, 2009))**.** *A function* $\mathbf{f}_D : \mathcal{X}^d \mapsto \mathbb{R}^d$ *trained on dataset* D *is* (ε, δ)-differentially private *if, for all neighboring datasets* D, $D^{'} \in \mathcal{Z}^n$ *and all measurable* $S \subseteq \mathbb{R}^d$ *, it holds that*

$$P(\mathbf{f}_D(\mathbf{x}) \in S) \leq \exp(\varepsilon) \cdot P(\mathbf{f}_{D'}(\mathbf{x}) \in S) + \delta \qquad \textit{for all } \mathbf{x} \in \mathcal{X}^d. \tag{6}$$

One common strategy to ensure DP is *output perturbation* (Chaudhuri et al., 2011; Zhang et al., 2022), which we later tailor to CATE estimation as part of our framework. Intuitively, one perturbs the prediction in a way that the predictions resulting from two neighboring databases cannot be differentiated. It has been shown in the literature (e.g., Dwork & Roth, 2014) that adding appropriately calibrated zero-centered noise (e.g., Gaussian noise) to the prediction is sufficient to ensure differential privacy for *traditional*, supervised machine learning tasks (but not for CATE estimation, as we discuss later). This is stated in the following *Gaussian noise privacy mechanism*.

[4] We give more details on the standard causal assumptions and CATE estimation in Supplement A.3.

Definition 2 (Gaussian noise privacy mechanism (Dwork & Roth, 2014))**.** *Let* $\mathbf{f} : \mathcal{X}^d \mapsto \mathbb{R}^d$ *be a function with* l_2*-sensitivity* $\Delta_2(\mathbf{f}) = \sup_{D \sim D', \mathbf{x} \in \mathcal{X}^d} ||\mathbf{f}_D(\mathbf{x}) - \mathbf{f}_{D'}(\mathbf{x})||$ *and* $\mathbf{U} \sim \mathcal{N}(0, \sigma \boldsymbol{I}_d)$ *for* $\sigma \geq \frac{1}{\varepsilon}\sqrt{2 \ln{(1.25/\delta)}}\, \Delta_2(\mathbf{f})$*. Then, the output perturbation mechanism* $\mathcal{M}(D, \mathbf{f}, \varepsilon) = \mathbf{f}_D(\mathbf{x}) + \mathbf{U}$ *preserves* (ε, δ)*-differential privacy.*

Definition 2 describes how to ensure DP for a given prediction. However, this requires estimating the training sensitivity $\Delta_2(\mathbf{f})$ of the employed model f, which, for general function classes such as neural networks, is infeasible. Hence, this motivates our custom framework later.

3.3 Problem statement

In our work, we aim at doubly robust CATE estimation under differential privacy. Specifically, we aim to derive a (ε, δ)-differentially-private version of $g_D^* = \arg\min_{g \in \mathcal{G}} L_D(g, \hat{\eta}, \lambda(\hat{\pi}))$ of the form

$$\mathbf{g}_{\mathrm{DP}}^*(\mathbf{x}) = \mathbf{g}_D^*(\mathbf{x}) + r(\varepsilon, \delta, \Delta_2(\mathbf{g}_D^*)) \cdot \mathbf{U}, \tag{7}$$

where $\mathbf{g}_D^*(\mathbf{x}) = (g_D^*(x_1), \ldots, g_D^*(x_d))$, $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and where $r(\cdot)$ is a *calibration function*. Importantly, we consider *arbitrary* working model classes $\mathcal{G}$, such as all parameterizations of a given neural network architecture.

Our problem statement – and thus our framework – is intentionally flexible. (1) We assume that the propensity score is *not* known and, instead, is estimated from observational data. (2) We focus on doubly robust meta-learners because these are *model-agnostic* and can thus be seamlessly instantiated with any machine learning model, including neural networks. (3) Our derivations are general and, therefore, apply to *any* orthogonal loss of the form in Equation 3. Below, we present our DP-CATE framework for the R-learner due to its state-of-the-art performance. We additionally provide an extension of our framework for the DR-learner in Supplement B.

3.4 Challenges in differentially private CATE estimation

The above task is highly non-trivial due to **three main challenges**:

① **Causal assumptions:** To point-identify the CATE, one must make the standard causal assumptions of positivity, consistency, and unconfoundedness (Rubin, 2005). Yet, privacy mechanisms perturb different parts of the data (or the model), which will arguably violate the causal assumptions and thus lead to *biased* estimates if not properly accounted for.

② **Fundamental problem of causal inference:** Unlike supervised prediction tasks for observed outcomes, CATE estimation requires counterfactual quantities $Y(a)$. However, one observes *only* one outcome per individual, while the estimation target CATE $\tau(x)$ is *never* observed (Pearl, 2010). Thus, estimation errors can neither be observed. Blindly introducing noise to the estimation setup to guarantee DP might thus bias the CATE estimate in an uncontrollable way.

③ **CATE is a function:** Many privacy mechanisms (e.g., Dwork, 2006; Nissim et al., 2007) apply *only* to point estimates $f(x)$ or vectors of point estimates $\mathbf{f}(\mathbf{x})$, but *not* to functions f. Thus, standard privacy mechanisms are generally *not* applicable to the CATE, which is a function. It is unclear how to provide a model-agnostic CATE learner that returns a private function.

4 Our framework: DP-CATE

Overview: We now present our DP-CATE framework which is aimed at CATE estimation for *doubly robust* meta-learners under (ε, δ)*-differential privacy*. To address the challenges from above, we employ output perturbation, which is highly suitable to our purpose for two reasons. First, it allows us to ensure that causal assumptions are fulfilled even after perturbation (→ challenge ①). Second, it allows us to retain the abilities of existing CATE methods designed to address challenge ②.

Use cases: Our framework comes in two variants, relevant for different use cases in medical practice:

① DP-CATE **for finite queries (→ Sec. 4.1):** Here, we aim to report a number d of CATE estimations (e.g., treatment effects across different age groups). ⇒ *How do we solve this?* We draw upon the often overlooked shared property of robustness and privacy of machine learning models in terms of insensitivity to outliers and small measurement errors (Dwork & Lei, 2009). For this, we propose to calibrate the noise added during the perturbation with a function $r(\cdot)$ of the influence function of the meta-learner. Importantly, our approach preserves the double-robustness guarantees of the original non-private model and is fully model-agnostic.

② DP-CATE **for functional queries (→ Sec. 4.2):** Here, we release an estimate g^*_{DP} of the complete CATE function τ (→ challenge ③), which can then be queried arbitrarily often (e.g, as in clinical decision support systems). Yet, this is non-trivial: Existing output perturbation mechanisms only apply to scalar or finite-dimensional vector-valued outputs. Therefore, the above is only valid if the overall number of queries made to the algorithm is both finite and known before the perturbation. ⇒ *How do we solve this?* We derive a non-trivial output perturbation method based on Gaussian processes that is valid for all functional CATE estimates solving Equation 3, as long as the estimation in the second stage is performed through a Gaussian kernel regression.

4.1 DP-CATE FOR A FIXED NUMBER OF QUERIES

In this variant of DP-CATE, a total number of d CATE estimates should be released. Here, the number of queries, d, to the CATE function is known a priori. For notational simplicity, we thus can simply rewrite the d separate CATE estimates as a d-dimensional vector. We employ bold letters in the following section to emphasize that we are interested in a vectorized version of the CATE meta-learner g^*_D, i.e., $\mathbf{g}^*_D(\mathbf{x}) \in \mathbb{R}^d$.

We now derive a calibration function $r(\cdot)$ that is applicable to any doubly robust CATE meta-learner. Then, we employ $r(\cdot)$ to calibrate a noise vector $\mathbf{U}$ with respect to the privacy budget (ε, δ) to the model sensitivity. Finally, we perturb $\mathbf{g}^*_D(\mathbf{x})$ to fulfill DP through $\mathbf{g}^*_{\mathrm{DP}}(\mathbf{x}) = \mathbf{g}^*_D(\mathbf{x}) + r^{\mathrm{R}}(\varepsilon, \delta) \cdot \mathbf{U}$.

For this, we borrow ideas from the literature that observed similarities between robust statistics and differential privacy (e.g., Avella-Medina, 2021; Dwork & Lei, 2009) but which we carefully tailor to our setting in the following. Our idea is to employ the so-called *influence function* (IF) of the CATE estimation model to calibrate the noise. The IF allows us to quantify how sensitive the CATE is when we add noise as part of our output perturbation. Intuitively, the IF describes the effect of an infinitesimally small perturbation of the input z on the model output.

Definition 3. *Let T be a functional of a distribution that defines the parameter of interest, $\theta = T(F)$. The* influence function *(IF) of T at z under distribution F is defined as*

$$\mathrm{IF}(z, T; F) := \lim_{t \mapsto 0} \frac{T((1-t)F + t\delta_z) - T(F)}{t}, \tag{8}$$

where δ_z denotes the Dirac-delta functional at z. The gross-error sensitivity *of T at z under F is given by the supremum of the Euclidean norm of the influence functions*

$$\gamma(T, F) := \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; F)\|. \tag{9}$$

Next, we derive the IF of doubly robust meta-learners for the CATE. Observe that $T(F)$ depends on the data distribution directly through F and indirectly through the first stage functional $S(F)$.

Lemma 1. *The IF of the CATE meta-learners described by the loss in Equation 3 is equivalent to the IF of their second-stage estimation, i.e.,*

$$\mathrm{IF}((z_1, z_2), (g, \eta), F) = \mathrm{IF}(z_2, g, F). \tag{10}$$

Proof. We prove Lemma 1 in Supplement G. □

The above lemma has an important implication: the influence of the samples used for training the nuisance estimators can be neglected when ensuring the differential privacy of the overall CATE meta-learner. Put simply, the lemma reduces the complexity of the subsequent derivation: we only need to focus on output perturbation for the second stage of the meta-learners (but not the first stage).

We now state our main theorem for differentially private CATE estimation with a known number d of queries. The intuition behind how we construct the calibration function $r(\cdot)$ builds upon a result in Nissim et al. (2007) in which (ε, δ)-differential privacy is achieved through calibrating noise with respect to the smooth sensitivity of the prediction model in comparison to the commonly employed global sensitivity from Definition 2. However, calculating the smooth sensitivity is still difficult or even infeasible for general function classes. Nevertheless, we show that the smooth sensitivity of the doubly robust meta-learners can be upper bounded by the gross-error sensitivity $\gamma(g, F)$ of the second stage regression (see Lemma 4 in Supplement F).

Theorem 1 (DP-CATE for finite queries)**.** *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$ and $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ the estimated nuisance functions. Furthermore, let D be the training dataset with $|D| = n$. For $z = (a, x, y) \in \mathcal{Z}$ we define*

$$\mathbf{g}^*_{\mathrm{DP}}(\mathbf{x}) := \mathbf{g}^*_D(\mathbf{x}) + \sup_{z \in \mathcal{Z}} \left\| \rho(a, \hat{\pi}(x))(\phi(z, \hat{\eta}, \lambda(\hat{\pi}(x))) - g^*_D(x)) \right\| \cdot \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon\, n} \cdot \mathbf{U}, \tag{11}$$

where $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Then, $\mathbf{g}^*_{\mathrm{DP}}(\mathbf{x})$ is (ε, δ)-differentially private.

Proof. We prove Theorem 1 in Supplement G. To do so, we first state the IF of general weighted doubly robust meta-learners. Then, we calculate the gross-error sensitivity to show that the sample-size-weighted sensitivity upper-bounds the smooth sensitivity of the respective learner. □

We now derive an output perturbation mechanism $r^{\mathrm{R}}(\varepsilon, \delta)$ for the R-learner. For this, we simply leverage the above theorem, so that we can ensure that the outputs of the R-learner fulfill DP.

Lemma 2. *For the R-Learner, the d-dimensional (ε, δ)-differentially private CATE estimate is*

$$\mathbf{g}^*_{\mathrm{DP}}(x) = \mathbf{g}^*_D(x) + r^{\mathrm{R}}(\varepsilon, \delta) \cdot \mathbf{U}, \tag{12}$$

where $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Specifically, the calibration function $r^{\mathrm{R}}(\varepsilon, \delta)$ is given by

$$r^{\mathrm{R}}(\varepsilon, \delta) = \sup_{z \in \mathcal{Z}} \left\| \frac{a - \pi(x)}{\mathbb{E}[\pi(X)\,(1 - \pi(X))]} \Big(y - \mu_a(x) + (a - \pi(x))\,(\mu_1(x) - \mu_0(x) - g^*_R(x)) \Big) \right\| \cdot c(\varepsilon, \delta, n),$$

where $c(\varepsilon, \delta, n) := \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon n}$.

Proof. The result directly follows from Theorem 1 with the respective weighting function $\lambda(\pi(x))$ stated in Section 3. The influence function is derived in Lemma 5 in Supplement F. □

In sum, our DP-CATE is implemented in a straightforward manner: one simply trains a meta-learner g^*_D on the dataset D, computes the CATE estimates $\mathbf{g}^*_D(\mathbf{x})$, and then transforms them into private estimates $\mathbf{g}^*_{\mathrm{DP}}(\mathbf{x})$ by using the calibration function r together with random noise $\mathbf{U}$. In particular, DP-CATE is then a consistent estimator of the treatment effect. We present more details on the theoretical properties of DP-CATE in Supplement D.

Scalability: The computational complexity of calculating the gross error sensitivity is *independent* of the size of the dataset or the number of queries. Overall, our post-hoc framework does not alter the runtime of the ML model.

4.2 DP-CATE FOR COMPLETE CATE FUNCTIONS

In this variant of DP-CATE, we seek to privately release an estimate g of the complete CATE function τ. Note that we cannot leverage Theorem 1 due to challenge ③, because it is only applicable to CATE estimates but not to complete functions. Instead, we must now derive a tailored approach.

Intuitively, we need to find (I) a type of noise and (II) a respective calibration function that does not depend on d. More precisely, the added noise should be a function itself to guarantee privacy of the CATE function. For (I), we follow (Hall et al., 2013) add a calibrated Gaussian process to the predicted CATE. To start with, we first state the definition of a Gaussian process.[5]

Definition 4 (Gaussian process)**.** *A family of random variables $\{X_t\}_{t\in T}$ is a Gaussian process if for any subset $S \in T$, $\{X_t\}_{t\in S}$ has a Gaussian distribution. The process is entirely determined by its mean function $m(t) := \mathbb{E}[X_t]$ and covariance function $K(s,t) := \mathrm{Cov}(X_s, X_t)$.*

[5] We refer to Rasmussen & Williams (2006) for an excellent, in-depth introduction to Gaussian processes.

Now, we are left with answering question (II) from above: *how to calibrate* the Gaussian process noise to fulfill the differential privacy guarantees? For this, we make the following two important observations: (i) The Neyman-orthogonality of the loss function guarantees privacy with regard to the data used for the nuisance estimation in terms of the sensitivity of two neighboring datasets. Hence, we again need to simply focus on the second stage of the meta-learner. (ii) If g lies in a reproducing kernel Hilbert space (RKHS), we can calibrate the Gaussian process noise with respect to the RKHS norm using results from functional analysis (Hall et al., 2013). To ensure that g indeed lies in an RKHS, we can later follow prior research (e.g., Kennedy, 2023a; Singh et al., 2024) and model the second-stage estimation in our DP-CATE framework as a Gaussian kernel regression.

We now want to bound the difference of CATE functions trained on neighboring datasets with respect to the norm of the Hilbert space. Recall that differential privacy of function f under the Gaussian mechanism in Def. 2 requires knowledge about $\sup_{D\sim D', \mathbf{x}\in\mathcal{X}^d} ||\mathbf{f}_D(\mathbf{x}) - \mathbf{f}_{D'}(\mathbf{x})||_2$ to calibrate the Gaussain noise variable. Similarly, we now require knowledge about $\sup_{D\sim D'} ||f_D - f_{D'}||_{\mathcal{H}}$ where f specifies an RKHS regression to calibrate the Gaussian process (Hall et al., 2013). However, to the best of our knowledge, no closed-form solution for this quantity exists. We thus derive the following lemma as an extension of Hall et al. (2013) to our setting.

Lemma 3. *Let $\mathcal{H}$ denote the RKHS induced by the Gaussian kernel $K(x, y) = \frac{1}{(\sqrt{2\pi}h)^q}\exp(-\frac{\|x-y\|^2}{2h^2})$ for $x, y \in \mathbb{R}^q$, and let f_D be the optimal solution to the RKHS regression*

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda\|f\|_{\mathcal{H}}, \tag{13}$$

on dataset D with $|D| = n$ for $\ell(\hat{y}, y)$ being a convex and Lipschitz loss function in $\hat{y}$ with Lipschitz constant L. Then, we have

$$||f_D - f_{D'}||_{\mathcal{H}} \leq \frac{L}{\lambda n}\left(\sqrt{(2\pi)}h\right)^{-q}. \tag{14}$$

Proof. We prove Lemma 3 in Supplement G. □

We now can use the above results and present our output perturbation mechanism for CATE functions. To ensure DP, the meta-learners proceed as follows: (i) Stage 1: We estimate the nuisance functions μ and π through *any* parametric or non-parametric method of choice. (ii) Stage 2: We perform a Gaussian kernel regression to minimize Eq. 3. (iii) We calibrate a suitably chosen Gaussian process based on Lemma 3 and add the resulting function to the CATE function. The double robustness of our framework directly follows from (Kennedy, 2023a). We present the pseudo-code for DP-CATE in Alg. 1. The following theorem states our desired privacy guarantee.

Theorem 2 (DP-CATE for functional queries)**.** *Let $\hat{\mu}(a, x)$ and $\hat{\pi}(x)$ denote the nuisance estimators trained in stage 1. Let $z = (a, x, y)$ be a data sample from dataset D with $|D| = n$ and $x \in \mathbb{R}^q$. Let $\mathcal{H}$ denote the RKHS induced by the kernel $K(x, y) = \frac{1}{\sqrt{2\pi h}}\exp(-\frac{\|x-y\|^2}{2h^2})$, and let $\ell(\cdot)$ be a convex and Lipschitz loss function with Lipschitz constant L. Furthermore, we define g^* as the second stage regression solving Equation 3 via*

$$g_D^* \in \arg\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\big[\rho(a_i, \pi(x_i))\left(\phi(z_i, \hat{\eta}, \lambda(\pi(x_i))) - g(x_i)\right)\big] + \lambda\|g\|^2_{\mathcal{H}} \tag{15}$$

$$\textit{for} \qquad \rho(a, \pi(x)) := (a - \pi(x))\lambda^{'}(\pi(x)) + \lambda(\pi(x)) \tag{16}$$

$$\textit{and} \qquad \phi(z, \eta, \lambda(\pi(x))) := \frac{\lambda(\pi(x))}{\rho(a, \pi(x))}\frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}(y - \mu(x, a)) + \mu(x, 1) - \mu(x, 0). \tag{17}$$

Furthermore, let G be the sample path of a zero-centered Gaussian process with covariance function $K(\cdot)$. Then, (ε, δ)-differential privacy is guaranteed by

$$g^*_{\text{DP}} := g^*_D + \frac{4L\sqrt{2\ln(2/\delta)}}{\left(\sqrt{2\pi}h\right)^q\lambda n\varepsilon}\cdot G \tag{18}$$

Proof. We prove Theorem 2 in Supplement G. □

We make a practical comment. The above theorem requires a convex Lipschitz loss $\ell(\cdot)$. There are many suitable loss functions (e.g., the squared loss on bounded domains, a trimmed squared loss, or the Huber loss). For many losses, the Lipschitz constant is data-independent and directly computable from the loss function.[6] We require the second stage in the meta-learner to be a Gaussian kernel

[6] For example, for the L1 loss, the Lipschitz constant L equals 1; for the Huber loss, L equals the loss parameter δ; and, for the truncated L2 loss, the constant equals the gradient at the truncation value.

regression, which is widely used in causal inference (e.g., Kennedy, 2023a; Singh et al., 2024). Nevertheless, our DP-CATE is still fairly flexible in that any Neyman-orthogonal meta-learner can be used (e.g., R-learner, DR-learner) and that *any* ML model can be used for nuisance estimation.

Scalability: Our post-hoc privatization framework does not increase the runtime of the ML model. Of note, the complexity of Alg. 1 is free of iterative calculations and therefore does *not* scale with the number of queries d. We present an iterative version of Alg. 1 in Supplement C.

5 Experiments

Implementation: Our DP-CATE is model-agnostic and highly flexible. Therefore, we instantiate our DP-CATE with multiple versions of the R-leaner (Nie et al., 2021) where we vary the underlying base learners. Hence, we implement the pseudo-outcome regression in the second stage as a random forest (RF) and a neural network (NN), respectively. We estimate the nuisance functions through neural networks. This is recommended as one typically allows for flexibility in the nuisance functions (Curth & van der Schaar, 2021). Details on implementation and training are in Supplement H. We emphasize again that there do not exist suitable baselines for DP-CATE. In Supplement E, we compare DP-CATE against DP-EBM Niu et al. (2022) and a naïve method based on k-anonymization, yet these are designed for different settings or different objectives.

Performance metrics: As explained in Sec. 2, there are **no** flexible CATE meta-learners that ensure DP. Hence, there is **no** suitable baseline for our task. Hence, we perform experiments to primarily show the applicability of our DP-CATE across different privacy budgets. We expect that the prediction performance will increase with the increasing privacy budget and then approach the prediction performance of a non-private CATE learner. We measure the performance via the *precision in estimation of heterogeneous effect* (PEHE) with regard to the true CATE (e.g., Hill, 2011).

5.1 Evaluation on synthetic datasets

Synthetic datasets: Due to the fundamental problem of causal inference, counterfactual outcomes are never observable in real-world datasets. Therefore, we follow common practice in evaluating our framework on synthetic datasets, which allows us to have access to ground-truth CATE and thus to the PEHE (e.g., Kennedy, 2023a; Nie & Wager, 2020).We evaluate DP-CATE for 300 queries.

We consider two settings with different treatment effect complexity following the data generation mechanism in Oprescu et al. (2019). • **Dataset 1** contains two confounders from which only one influences CATE. This setting allows us to visualize the CATE function and the effect of privatization on the prediction for varying covariate values. • **Dataset 2** contains 30 confounders and higher-dimensional influences on the CATE. Details on the data-generating mechanism are in Supplement H.

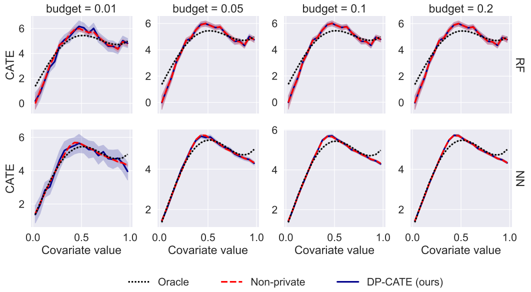


Figure 3: **Dataset 1** (finite queries). Predictions under different base learners and privacy budgets.

Results for finite queries: • **Dataset 1**: Fig. 3 shows the predictions for different base learner specifications and different privacy budgets. We make the following observations: (1) Our DP-CATE performs as expected: with increasing privacy budget, the predictions become less 'noisy' and converge to those of the non-private CATE. (2) Our DP-CATE shows consistent patterns for different base learners. For example, the predictions under both RF and NN are almost identical, showing the flexibility of our framework. • **Dataset 2**: Fig. 5 shows now the PEHE. Note that, here, we directly compare DP-CATE for finite queries (combined with RF and NN) and DP-CATE for functional queries (kernel regression). Again, we find that our DP-CATE performs as expected: the PEHE decreases with increasing privacy budget and converges against that of the non-private learner, showing that our DP-CATE makes correct predictions.

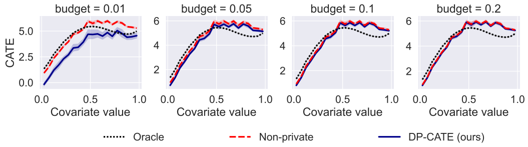


Figure 4: **Dataset 1** (functional queries). Predictions under different privacy budgets.

Results for functional queries: • **Dataset 1**: Fig. 4 shows the predictions of DP-CATE for functional queries across different privacy budgets. We observe similar behavior as in the case of finite queries: With increasing privacy budget, the predictions converge to those of the non-private learner. • **Dataset 2**: Fig. 5 shows the results for more complex dataset. As before, we observe that DP-CATE for functions behaves in the same way as DP-CATE for finite queries. Overall, our findings are consistent across datasets and different base learner specifications.

5.2 Evaluation on medical datasets

Medical datasets: We demonstrate the applicability of DP-CATE to medical datasets by using the **MIMIC-III** dataset (Johnson et al., 2016) and the **TCGA** dataset (Weinstein et al., 2013). MIMIC-III contains real-world health records from patients admitted to intensive care units at large hospitals. We aim to predict a patient's red blood cell count after being treated with mechanical ventilation. The Cancer Genome Atlas (TCGA) dataset contains a large collection of gene expression data from patients with different cancer types. We assign a treatment indicator based on the gene expression level and aim to predict a constant effect across all expression levels. Details are in Supplement H.

Results: • **MIMIC-III:** Fig. 6 reports the predictions of the CATE against different levels of hematocrit and different privacy budgets. Here, we have $d = 1312$ queries (i.e., the size of the test set). We observe a positive relationship as stipulated by domain knowledge in medicine. Our DP-CATE framework works as desired: for smaller privacy budgets, more noise should be added, which is also reflected in a larger variation of the predictions. • **TCGA:** Fig. 7 shows again that our DP-CATE is effective on a large number of queries (i.e., $d = 2659$, which corresponds to the size of the test set): we observe that, with increasing privacy budgets, the CATE predictions become more precise and approach those of the non-private learner as expected. Overall, the loss in precision due to the privacy guarantees (i.e., when comparing DP-CATE to the non-private learner) is fairly small. We provide further experimental results in Supplements B and I.

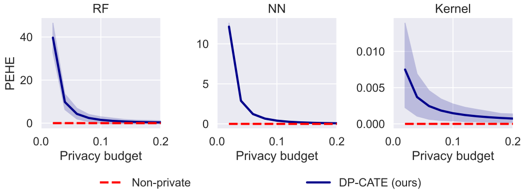


Figure 5: **Dataset 2**. Prediction errors under different privacy budgets of DP-CATE (finite) on RF and NN and DP-CATE (functional) combined with kernel regression. Plots centered at the PEHE of the non-private learner.

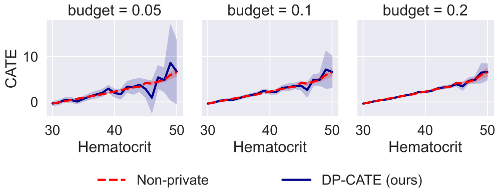


Figure 6: **MIMIC-III** (finite queries). Our DP-CATE generates private estimates of the effect of ventilation for different levels of hematocrit.

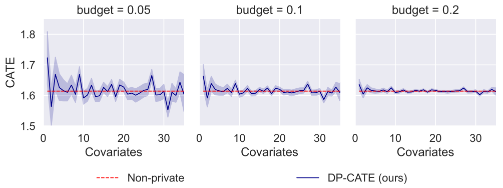


Figure 7: **TCGA** (finite queries). DP-CATE consistently estimates the constant treatment effect across the sum of all covariate values.

Takeaways: *The prediction error of DP-CATE decreases with larger privacy budgets as desired and converges to the non-private error, confirming that our framework makes precise CATE predictions.*

6 Discussion

Applicability: We provide a general framework for differentially private and doubly robust CATE estimation from observational data. First, our DP-CATE is carefully designed for observational data, which are common in medical applications (Feuerriegel et al., 2024). Second, DP-CATE is applicable to various doubly robust meta-learners (e.g., R-learner, DR-learner), which are widely used in practice. Third, DP-CATE allows different use cases: one can release either a certain number of CATE estimates or even the complete CATE function (e.g., as in clinical decision support systems).

Extension to the DR-Learner: Our derivations in Sec. 4 focus on the popular R-learner due to its favorable theoretical properties. Nevertheless, our DP-CATE can be applied to any other doubly robust meta-learner. In Supplement B, we thus provide an extension to the DR-learner. Therein, we also provide additional numerical experiments to show the applicability of our framework.

Conclusion: Ensuring the privacy of sensitive information in treatment effect estimation is mandated for ethical and legal reasons. Here, we provide the first framework for differentially private CATE estimation from observational using meta-learners.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Conference on Computer and Communications Security*, 2016.

Anish Agarwal and Rahul Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint*, 2107.02780, 2024.

Marco Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

Anna Baiardi and Andrea A. Naghi. The value added of machine learning to causal inference: evidence from revisited studies. *The Econometrics Journal*, 27(2):213–234, 2024.

Karla V. Ballmann. Biomarker: Predictive or prognostic? *Journal of Clinical Oncology*, 33(33): 3968–3971, 2015.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *IEEE Annual Symposium on Foundations of Computer Science*, 2014.

Artem Betlei, Théophane Gregoir, Thibaud Rahier, Alois Bissuel, Eustache Diemert, and Massih-Reza Amini. Differentially private individual treatment effect estimation from aggregated data. *hal*, 2021.

Michael Braun and Eric M. Schwartz. Where a-b testing goes wrong: How divergent delivery affects what online experiments cannot (and can) tell you about how customers respond to advertising. 2024.

Kyle B. Brothers and Mark A. Rothstein. Ethical, legal and social implications of incorporating personalized medicine into healthcare. *Personalized Medicine*, 12(1):43–51, 2015.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*. 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing*. 2009.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Paul B. Ellickson, Wreetabrata Kar, and James C. Reeder, III. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, 42(4): 704–728, 2023.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Dennis Frauen, Konstantin Hess, and Stefan Feuerriegel. Model-agnostic meta-learners for estimating heterogeneous treatment effects over time. *arXiv preprint*, 2407.05287, 2024.

Kazuto Fukuchi, Quang Khai Tran, and Jun Sakuma. Differentially private empirical risk minimization with input perturbation. In *Discovery Science*, 2017.

Sharmistha Guha and Jerome P. Reiter. Differentially private estimation of weighted average treatment effects for binary outcomes. *arXiv preprint*, 2408.14766v1, 2024.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14:703–727, 2013.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Ta-Wei Huang and Eva Ascara. Debiasing treatment effect estimation for privacy-protected data: A model auditing and calibration approach. *SSRN*, 2023.

Roger Iyengar, Joseph P. Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *IEEE Symposium on Security and Privacy*, 2019.

Adel Javanmard, Vahab Mirrokni, and Jean Pouget-Abadie. Causal inference with differentially private (clustered) outcomes. *arXiv preprint*, 2308.00957v2, 2024.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint*, 2203.06469, 2023a.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2), 2023b.

Daniel Kifer, Adam Smith, and Thakurta Abhradeep. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, 2012.

Si Kai Lee, Luigi Gresele, Mijung Park, and Krikamol Muandet. Privacy-preserving causal inference via inverse probability weighting. *arXiv preprint*, 1905.12592v2, 2019.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning (ICML)*, 2018.

Pawel Morzywolek, Johan Decruyenaere, and Stijn Vansteelandt. On weighted orthogonal learners for heterogeneous treatment effects. *arXiv preprint*, 2303.12687, 2023.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations (ICLR)*, 2021.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2020.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *ACM Symposium on Theory of Computing*, 2007.

Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal Learning and Reasoning*, 2022.

Yuki Ohnishi and Jordan Awan. Locally private causal inference for randomized experiments. *arXiv preprint*, 2301.01616v4, 2023.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning (ICML)*, 2019.

Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning (ICML)*, 2023.

Judea Pearl. The foundations of causal inference. *Sociological Methodology*, 40(1):75–149, 2010.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian process for machine learning*. Adaptive computation and machine learning. The MIT Press, London, England, 3. print edition, 2006.

Rachel Redberg, Antti Koskela, and Yu-Xiang Wang. Improving the privacy and practicality of objective perturbation for differentially private linear learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Donald. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 2005.

Jonas Schweisthal, Dennis Frauen, Mihaela van der Schaar, and Stefan Feuerriegel. Meta-learners for partially-identified treatment effects across multiple environments. In *International Conference on Machine Learning (ICML)*, 2024.

R. Singh, L. Xu, and A. Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516, 2024.

Mark J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

Kyle Wang, Michael J. Eblan, Allison M. Deal, Matthew Lipner, Timothy M. Zagar, Yue Wang, Panayiotis Mavroidis, Carrie B. Lee, Brian C. Jensen, Julian G. Rosenman, Mark A. Socinski, Thomas E. Stinchcombe, and Lawrence B. Marks. Cardiac toxicity after radiotherapy for stage III non-small-cell lung cancer: Pooled analysis of dose-escalation trials delivering 70 to 90 gy. *Journal of Clinical Oncology*, 35(13):1387–1394, 2017.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract. In *Conference on Health, Inference, and Learning (CHIL)*, 2020.

Yue Wang, Daniel Kifer, and Jaewoo Lee. Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, 9, 2019.

John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Leon Yao, Paul Yiming Li, and Jiannan Lu. Privacy-preserving quantile treatment effect estimation for randomized controlled trials. *arXiv preprint*, 2401.14549v1, 2024.

Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Mikhail Zhelonkin, Marc G. Genton, and Elvezio Ronchetti. On the robustness of two-stage estimators. *Statistics & Probability Letters*, 82(4):726–732, 2012.

A Additional background material

A.1 Differential privacy

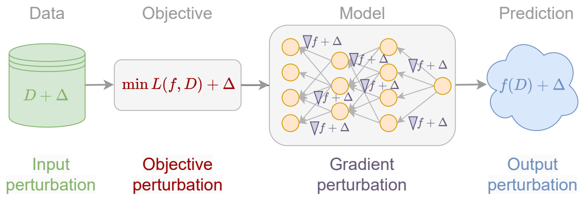


Figure 8: Privacy mechanisms in the machine learning workflow.

DP mechanisms: There are four main strategies of DP mechanisms (see Fig. 8): (i) *Input perturbation* independently randomizes each data sample before model training (e.g., Fukuchi et al., 2017). (ii) *Objective perturbation* adds a random term to the objective and releases the respective minimizer (e.g., Iyengar et al., 2019; Kifer et al., 2012; Redberg et al., 2023). The mechanisms in this field commonly make strong assumptions on the smoothness or convexity of the objective. (iii) *Gradient perturbation* clips, aggregates, and adds noise to the gradient updates in each step of gradient descent methods during model training (e.g., Abadi et al., 2016; Wang et al., 2017; 2019). (iv) *Output perturbation* adds noise to the non-private model prediction before its release (e.g., Chaudhuri et al., 2011; Zhang et al., 2022). All stated mechanisms are general strategies that must be carefully adapted to our CATE estimation setting. In our work, we employ output perturbation as it is most compatible with learning causal effects.

Choice of DP mechanism: Our DP-CATE framework employs *output perturbation* to achieve DP. Output perturbation is highly suitable for our setting since (i) it ensures that causal assumptions are fulfilled even after perturbation (challenge ①), and (ii) it retains the power of existing CATE estimation methods for addressing the fundamental problem of causal inference (challenge ②). The other DP strategies discussed above fail to fulfill the requirements.

Specifically, input perturbation might introduce confounding bias or violate the consistency assumption. For gradient and objective perturbation, the convergence of the model might be unclear. Furthermore, objective perturbation might fail to achieve the targeted privacy guarantee if the model does not converge to the exact global minimum in finite time (Iyengar et al., 2019). Gradient perturbation results in a non-trivial privacy overhead and does not align with our goal of providing a model-agnostic meta-learning framework (Redberg et al., 2023).

A.2 Extended related work

The only existing method targeting our setting was proposed by Niu et al. (2022). The authors provide an algorithm for differentially private CATE estimation through existing CATE meta-learners. However, the method necessitates special private base learners for the separate sub-algorithms in each stage of the meta-learner. It is thus *not agnostic* to any choice of ML method for the first- and second-stage regressions. Furthermore, it has been shown that privatizing separate parts of causal estimators can result in biased causal estimates (Ohnishi & Awan, 2023).

A different line of work proposes *locally differentially private* (LDP) algorithms (Agarwal & Singh, 2024; Huang & Ascara, 2023; Ohnishi & Awan, 2023). This notion of privacy becomes necessary if the central data curator cannot be trusted. During data collection, calibrated noise is added to each sample before adding it to the database. However, the perturbed data might violate the assumptions to identify causal treatment effects. Furthermore, this notion of privacy significantly reduces the predictive accuracy of the estimators (Huang & Ascara, 2023). Thus, whenever the data curator is a trusted party (as assumed in our work), global differential privacy is sufficient and should be the notion of choice as it is less accuracy-compromising than its local counterpart.

A.3 Theory on CATE estimation

The estimation of causal quantities, such as the conditional average treatment effect $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, involves counterfactual quantities $Y(a)$, as only one outcome per individual can be observed. Here, $Y(a)$ is the potential outcome that would hypothetically be observed if a decision a is taken.

Due to the above, identification of causal effects from observational data necessitates the following three assumptions common in the literature (e.g., Curth & van der Schaar, 2021; Feuerriegel et al., 2024):

1. Consistency: The potential outcome $Y_i(a = k)$ equals the observed factual outcome Y_i when individual i was assigned treatment k.
2. Positivity/overlap: The treatment assignment is not deterministic. Specifically, there exists a positive probability for each possible combination of features to be assigned to both the treated and the untreated group, i.e., $\exists\ \kappa > 0$ such that $\kappa < \pi(x) < 1 - \kappa$ for all $X = x \in \mathcal{X}$.
3. Unconfoundedness: Conditioned on the observed covariates, the treatment assignment is independent of the potential outcomes, i.e., $Y(0), Y(1) \perp\!\!\!\perp A|X$. Specifically, there are no unobserved variables (confounders) influencing both the treatment assignment and the outcome.

Importantly, the above assumptions are standard in the literature. Further, the assumptions are necessary for consistent causal effect estimation for *any* machine learning model. Then, CATE is identifiable as

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mu(x, 1) - \mu(x, 0), \tag{19}$$

where $\mu(x, a) = \mathbb{E}[Y \mid X = x, A = a]$. To estimate τ, one could thus train a machine learning model that estimates the aforementioned conditional expectation and then calculates the difference in conditional expectations for a given $X = x$. This is commonly referred to as *plug-in* method, yet which has several drawbacks, as we outline below. Rather, the preferred way to estimate the CATE is through meta-learners.

Meta-learners define model-agnostic algorithms which can be implemented with arbitrary machine learning algorithms. Therefore, those learners are flexible and commonly employed in practice. The plug-in estimation approach that descried above thus resembles a meta-leaner. CATE meta-learners can be classified into four different categories, depending on the ways they leverage the data: one-step learners, regression-adjusted two-stage learners, propensity-weighted two-stage learners, and doubly-robust tow-stage learners (Curth & van der Schaar, 2021). Causal meta-learners for specialized tasks, such as partial identification or treatment effects over time, have also seen increasing interest recently (e.g., Frauen et al., 2024; Oprescu et al., 2023; Schweisthal et al., 2024)

We now discuss each type of meta-learner and their potential drawbacks in more detail:

1. One-step plug-in learner: Here, ML models are trained to predict $\mu(x, a)$, either one single model for both treatment values or two different models, $\mu(x, 1)$, $\mu(x, 0)$. Then, the CATE is estimated directly as $\tau(x) = \mu(x, 1) - \mu(x, 0)$.
2. Two-stage regression-adjusted learner: In the observed data, the difference between factual and counterfactual outcomes is never present. Therefore, two-stage learners construct *pseudo-outcomes* as surrogates, which equal the CATE in expectation. The regression-adjusted learner designs the pseudo-outcome through a reweighting based on the function μ, which is estimated in the first stage. A misspecification of μ results in a biased CATE estimator.
3. Two-stage propensity-weighted learner: Here, the pseudo-outcome is constructed based on the Horvitz-Thompson transformation. Only the propensity function π needs to be estimated in the first step. A misspecification of π results in a biased CATE estimator.
4. Two-stage doubly robust learner: Different from the above two-stage learners, this learner is *unbiased* if either the propensity function π or the outcome regressions μ are correctly specified. Specifically, the final estimation is insensitive to small perturbations in the nuisance functions. This property is achieved through a Neyman-orthogonal loss function such as Eq. 3.

Hence, we focus on the two-stage doubly robust learner throughout our work due to the clear advantages.

B EXTENSION TO THE DR-LEARNER

In our main paper, we focused on the R-Learner for the derivations and the experiments. However, DP-CATE is applicable to any weighted two-stage doubly robust CATE meta-learner. Therefore, we now provide an extension to the DR-Learner.

For the DR-Learner, the weight function simplifies to $\lambda^{\mathrm{DR}}(\pi(x)) = 1$. Therefore, the private DR-Learner for a fixed number of queries is given below.

Lemma 2b. *The d-dimensional (ε, δ)-differentially-private CATE estimated through the DR-Learner is given by*

$$\mathbf{g}^*_{\mathrm{DP}}(x) = \mathbf{g}^*_D(x) + r^{\mathrm{DR}}(\varepsilon, \delta) \cdot \mathbf{U}, \tag{20}$$

where $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Specifically, the calibration function $r^{\mathrm{DR}}(\varepsilon, \delta)$ is given by

$$r^{\mathrm{DR}}(\varepsilon, \delta) = \sup_{z \in \mathcal{Z}} \left\| \frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}(y - \mu_a(x)) + \mu_1(x) - \mu_0(x) - g^*_{DR}(x) \right\| \cdot c(\varepsilon, \delta, n), \tag{21}$$

where $c(\varepsilon, \delta, n) := \frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon n}$.

Proof. We prove Lemma 2b in Supplement G. □

Observation: The noise calibration necessary for the privatization requires maximizing the influence function. For the DR-Learner, the IF includes the inverse of $\pi(x)(1 - \pi(x))$. Although we assume $\pi(x)$ to be bounded away from zero and one (the causal overlap assumption), maximizing over this term can lead to a very large calibration factor. This might limit the applicability of the DR-Learner for differentially-private CATE estimation.

Below, we evaluate the private DR-Learner in the same manner as the R-Learner in the main paper. We observe the expected behavior in Fig. 9. The noise added for privatizing the output is extremely enlarged compared to the R-Learner in Fig. 3. Note the varying scale of the y-axis.

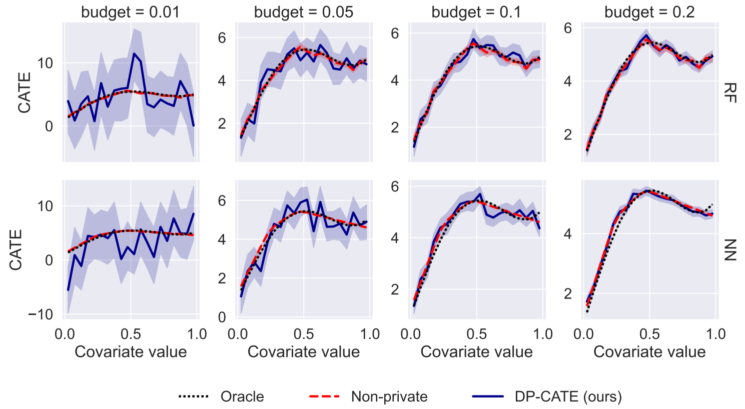


Figure 9: Evaluation of DP-CATE for finite queries the on DR-Learner for different base-learner specifications on dataset 1.

In Fig. 10, we show the performance of our DP-CATE for functional queries in combination with the DR-Learner. We see the same behavior as for DP-CATE for a finite number of queries. It is noteworthy that the predictions of DP-CATE for functional queries coverage even more rapidly to the non-private predictions for increasing privacy budget than the predictions of DP-CATE for finite queries.

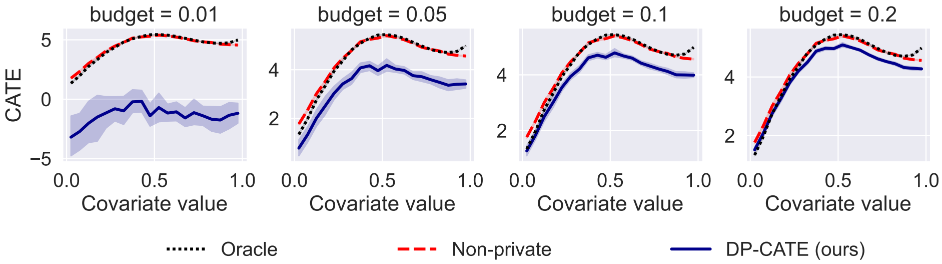


Figure 10: Evaluation of DP-CATE for functional queries on the DR-Learner on dataset 1.

C Similarities and differences of the finite an functional DP-CATE framework

C.1 Discussion

If one only wants to release private CATE estimates once, both approaches ① and ② are applicable. Nevertheless, the second approach called "functional approach" can also be employed for iteratively querying the function, which is especially of interest to medical practitioners aiming to assess the treatment effect of a drug for various patients with different characteristics. Put simply, when companies want to release a decision support system to guide treatment decisions of individual patients. Such treatment decisions are made based on the entire CATE model, then the "functional" approach is preferred. In contrast, the first approach (called "finite-query approach") is preferred whenever only a few CATE values should be released. This is relevant for researchers (or practitioners) who may want to share the treatment effectiveness for a certain number of subgroups (but not for individual patients).

The functional approach requires sampling from a Gaussian process. Depending on whether one aims to report finitely many queries once through this approach or iteratively query the function, the sampling procedure from the Gaussian process G differs. We highlight the differences in the type of queries in the following:

- **Simultaneous finitely many queries:** For querying the function only once with a finite amount of queries, sampling from a Gaussian process implies sampling from the prior distribution of the process. In empirical applications, this means that one samples from a multivariate normal distribution. Therefore, the noise added in the functional approach is similar to the finite-query approach. However, the approaches ① and ② are not the same, as the noise added in the functional approach is correlated, whereas the noise variables in finite-query approach are independent. Therefore, the functional approach might result in what appears to be a consistent under- or overestimation of the target. Still, both approaches guarantee privacy.

- **Iteratively querying the function:** In this setting, sampling from a Gaussian process implies sampling from the posterior distribution of the process. Specifically, if no query has been made to the private function yet, the finite-query approach proceeds by providing the first private CATE estimate of query x_1. Observe that the privatization of every further iterative query x_i needs to account for the information leakage through answering former queries. Thus, sampling from a Gaussian process now relates to sampling from the posterior distribution. To do so, it is necessary to keep track of and store former queries $x_1, \ldots, x_{i-1}$ and the privatized outputs. This setting is entirely different from our finite setting approach, in which we propose adding Gaussian noise scaled by gross-error sensitivity.

C.2 Algorithms for DP-CATE functions

Private outputs of the function g^{DP} in Theorem 2 can be released in two ways: (i) the standard *batch* setting presented in Algorithm 1, in which the private function outputs a private vector of CATE estimates *once* and (ii) the *iterative* or *online* setting, in which the function is queried iteratively, outputting one private CATE estimate at a time. Below, we provide an alternative algorithm to apply Theorem 2 in an iterative way.

Algorithm 1: Pseudo-code of out DP-CATE for functions

```
Input: CATE meta-learner $g_D^*$ trained on dataset $D$ with $|D| = n$, Gaussian kernel matrix
       $(K(x_i, x_j))_{i,j=1}^d$, query $\mathbf{x}_{\text{query}} \in \mathbb{R}^d$, privacy budget $\varepsilon$, $\delta$, Lipschitz const. $L$, ridge regularization
       $\lambda$
Output: Privatized CATE function $\mathbf{g}_{\text{DP}}^*$
$q \leftarrow length(x_i)$;
/* Calculate calibration term $r$                                  */
$r \leftarrow (4L\sqrt{2\log(2/\delta)})/((\sqrt{2\pi}h)^q \cdot \lambda n\varepsilon)$;
/* Sample from Gaussian process                                    */
$\mathbf{U} \sim \mathcal{N}(\mathbf{0}_d, (K(x_i, x_j))_{i,j=1,\dots,d})$;
/* Return private estimates                                        */
$\mathbf{g}_{\text{DP}}^*(\mathbf{x}_{\text{query}}) \leftarrow \mathbf{g}_D^*(\mathbf{x}_{\text{query}}) + r \cdot \mathbf{U}$;
```

Iterative approach: If no query has been made to the private function yet, we can employ Algorithm 2 to provide a private CATE estimate $g^{\text{DP}}(x_1)$, where x_1 denotes the first query. Specifically, Algorithm 2 samples from a Gaussian Process *prior* by sampling a suitable multivariate Gaussian noise variable. Observe that the privatization of every further iterative query x_i needs to account for the information leakage through answering former queries. Thus, sampling from a Gaussian Process now relates to sampling from the *posterior* distribution. To do so, it is necessary to keep track and store former queries $x_1, \dots, x_{i-1}$ and the privatized outputs $G_i = (g^{\text{DP}}(x_1), \dots, g^{\text{DP}}(x_{i-1}))^T$.

Algorithm 2: Pseudo-code of DP-CATE for functions (iterative setting)

```
Input: CATE meta-learner $g_D^*$ trained on dataset $D$ with $|D| = n$, Gaussian kernel matrix conditioned on
       the former queries $x_1, \dots, x_{i-1}$ $C_i := (K(x_k, x_l))_{k,l=1}^{i-1}$, former private outputs
       $G_i = (g^{\text{DP}}(x_1), \dots, g^{\text{DP}}(x_{i-1}))^T$, new query $x_i$, privacy budget $\varepsilon$, $\delta$, Lipschitz const. $L$, ridge
       regularization $\lambda$
Output: Privatized new prediction $g_{\text{DP}}^*(x_i)$
if $i$=1 then
    /* Apply Algorithm 1                                           */
end
else
    /* Calculate pairwise kernel vector                            */
    $V_i \leftarrow (K(x_1, x_i), \dots, K(x_{i-1}, x_i))^T$;
    /* Sample from Gaussian process posterior                      */
    $s \sim \mathcal{N}(V_i^T C_i^{-1} G_i, , K(x_i, x_i) - V_i^T C_i^{-1} V_i)$;
    /* Return private estimate                                     */
    $g^{\text{DP}}(x_i) \leftarrow s$;
end
```

A note on complexity: Algorithm 2 requires storage and iterative updating of the outcome vector G_i and the inverse matrix C_i^{-1}. For an increasing amount of queries, the computational complexity of Alg. 2 will thus grow. This poses a limitation of our approach for settings with many iterative queries.

D THEORETICAL EVALUATION OF PROPOSED ALGORITHMS

To show that our DP-CATE provides a valid but private estimator, we need to analyze the consistency of our framework. First, observe that we focus on consistent CATE learners as the underlying non-private estimators. Let g denote the non-private base meta-learner and τ the true CATE. Then, by consistency of the estimator, we have

$$\|g - \tau\| \to 0, \; n \to \infty. \tag{22}$$

Further, note that our framework does not alter the estimation procedure itself. Further, observe that the amount of noise added decreases in the sample size n. Let g^P denote the private estimator. Then, we have

$$\|g^P - g\| \to 0, \; n \to \infty. \tag{23}$$

Overall, we thus receive

$$\|g^P - \tau\| \leq \|g^P - g\| + \|g - \tau\| \to 0, \; n \to \infty. \tag{24}$$

Therefore, g^P is a consistent estimator of τ.

E COMPARISON TO BASELINES

We highlight again that other DP methods are either not model-agnostic (Niu et al., 2022), or restricted to RCT data or to data with binary outcomes (Betlei et al., 2021; Guha & Reiter, 2024). Therefore, the baselines are not applicable to general settings as in our paper. In other words, powerful baselines with theoretical DP guarantees are missing.

E.1 COMPARISON TO DP-EBM (NIU ET AL., 2022)

Nevertheless, in the following, we compare DP-CATE to DP-EBM (Niu et al., 2022), as the method is – in principle – applicable to the datasets we employ. We report the results on dataset 1 in Fig. 11. Please note the different ranges of the y-axis representing the CATE.

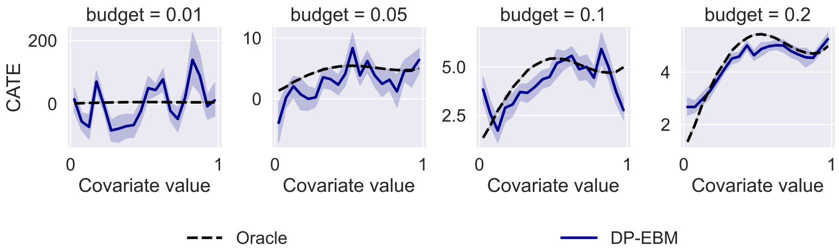


Figure 11: Evaluation of the baseline DP-EBM for finite queries (300) for various privacy budgets.

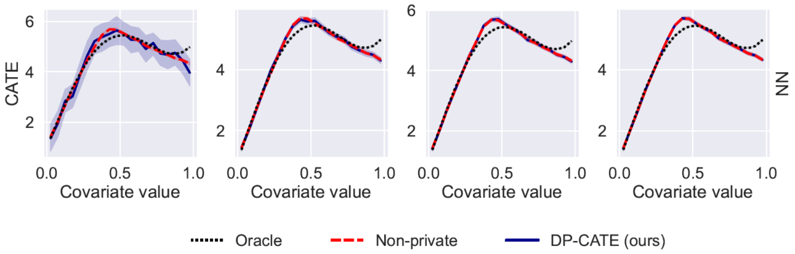


Figure 12: Evaluation of DP-CATE for finite queries (300) for various privacy budgets.

We find that the error induced by privatization from DP-EBM is – by far – worse than the error induced by our DP-CATE in Fig. 12 (we already reported Fig. 12 in our main paper). Especially for small privacy budgets, DP-EBM is not reliable. We explain this is due to the need for privatizing both stages in the DP-EBM framework. In sum, this confirms that our method is superior to existing methods for differentially private CATE estimation.

E.2 COMPARISON TO k-ANONYMIZATION

Furthermore, we compare our method against a very naïve baseline based on k-anonymization. As we show in the following, such naïve baselines have clear limitations in both theory and in our experiments. First, such k-anonymization essentially performs a data aggregation as part of the privatization technique, yet this can *break the consistency assumption* necessary for causal identifiability. Hence, this can lead to *biased* results. We observe that our method provides superior CATE estimates (while further offering theoretical guarantees to ensure DP) for almost all privacy budgets. Hence, our new experiments confirm again the effectiveness of our proposed method.

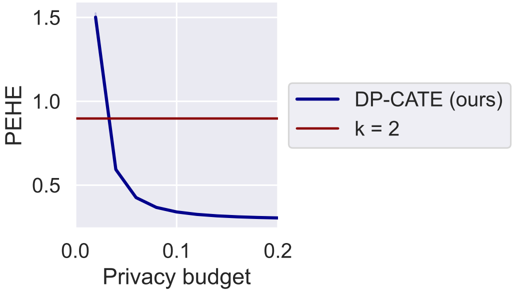


Figure 13: Comparison of our DP-CATE with a naïve data anonymization method based on k-anonymization in terms of PEHE.

F SUPPORTING LEMMAS

Lemma 4. *Let g be the optimizer of Equation 3 on dataset D with $|D| = n$ and covariate dimension d. Furthermore, let $\gamma(g, F)$ denote the gross error sensitivity of g. For the ξ-smooth sensitivity of the doubly robust meta-learner g with $\xi = \frac{\varepsilon}{4(d+2\log(2/\delta))}$, it holds*

$$SS_{\xi}(g, D) := \sup_{D'}\{\exp(-\xi d_{\mathrm{H}}(D, D^{'}))LS(g, D^{'}) \mid D^{'} \in \mathcal{Z}^n\} \leq \frac{\sqrt{\log(n)}}{n}\gamma(g, F), \tag{25}$$

where $\mathcal{Z}^n$ denotes the data domain and $LS(\cdot)$ the local sensitivity given by

$$LS(g, D) := \sup_{D'}\left\{\|g(D) - g(D^{'})\| \;\middle|\; d_{\mathrm{H}}(D, D^{'}) = 1\right\}. \tag{26}$$

Proof. For ease on notation, we denote the score function of Equation 3 by $\psi(z, g) := \frac{\rho(a,\pi_0(x))}{\mathbb{E}[\lambda(\pi_0(x))]}(\phi(z, \eta, \lambda(\pi(x))) - g(x))$ for $\phi(z, \eta, \lambda(\pi(x))) := \frac{\lambda(\pi(x))}{\rho(a,\pi(x))}\frac{a-\pi(x)}{\pi(x)(1-\pi(x))}(y - \mu(x, a)) + \mu(x, 1) - \mu(x, 0)$. First, note that $\psi(z, g)$ is differentiable w.r.t. g for all $z \in \mathcal{Z}$ with $\psi^{'}(z, g) = \frac{\partial\psi(z,g)}{\partial g} = -\frac{\rho(a,\pi_0(x))}{\mathbb{E}[\lambda(\pi_0(x))]}$. Since we assume that our data stems from a bounded domain, there exist constants K and L such that

$$\sup_{z\in\mathcal{Z}}\|\psi(z, g)\| \leq K \qquad \text{and} \qquad \sup_{z\in\mathcal{Z}}\|\psi^{'}(z, g)\| \leq L. \tag{27}$$

Therefore, $\psi(z, g)$ and $\psi^{'}(z, g)$ are uniformly bounded in $\mathcal{Z}$. Furthermore, it holds $M_F = M(g, F) = -\mathbb{E}_F[\psi^{'}(Z, g(F))] > 0$ for all empirical distributions $G_n \in \{G \mid D(G) \in \mathcal{Z}^n\}$ and the generative distribution $F = F_0$, since

$$M_F = \frac{\mathbb{E}[\rho(A, \pi_0(X))]}{\mathbb{E}[\lambda(\pi_0(x))]} = 1, \tag{28}$$

with $\mathbb{E}[\rho(A, \pi_0(X))] = 0 + \mathbb{E}[\lambda(\pi_0(x))]$. As a result, we can employ Lemmas 1 and 2 in Avella-Medina (2021) to show the desired upper bound of the smooth sensitivity. □

Lemma 5. *The influence functions of the DR-Learner and the R-Learner are given by*

$$\mathrm{IF}^{\mathrm{DR}}(z, g; F) = \frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}(y - \mu(x, a)) + \mu(x, 1) - \mu(x, 0) - g(x), \tag{29}$$

$$\mathrm{IF}^{\mathrm{R}}(z, g; F) = \frac{a - \pi(x)}{\mathbb{E}[\pi(x)(1 - \pi(x))]}\left(y - \mu(x, a) + (a - \pi(x))(\mu(x, 1) - \mu(x, 0) - g(x))\right) \tag{30}$$

Proof. By Lemma 1, the influence functions of the two-stage learner equal the influence function of the second-stage regression. Hirano et al. (2003) state a general IF for the weighted average treatment effect (WATE) as

$$\frac{\rho(a, \pi_0(x))}{\mathbb{E}[\lambda(\pi_0(x))]}\left(\frac{\lambda(\pi(x))}{\rho(a, \pi(x))}\frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}(y - \mu(x, a)) + \mu(x, 1) - \mu(x, 0) - g(x)\right) \tag{31}$$

for $\rho(a, \pi(x)) \neq 0$ with

$$\rho(a, \pi(x)) := (a - \pi(x))\lambda^{'}(\pi(x)) + \lambda(\pi(x)) \tag{32}$$

For the R-Learner it holds that $\lambda(\pi(x)) = \pi(x)(1 - \pi(x))$ and for the DR-Learner $\lambda(\pi(x)) = 1$. Since for the R-Learner $\rho(a, \pi(x)) = (a - \pi(x))^2$ and for the DR-Learner $\rho(a, \pi(x)) = 1$, the statement follows. □

G Proofs of the main theorems

G.1 Proofs of Lemmas 1 and 3

Proof of Lemma 1

Lemma 1. *The IF of the two-stage CATE estimators described by the loss in Equation 3 is equivalent to the IF of its second-stage estimation:*

$$\mathrm{IF}((z_1, z_2), (g, \eta), F) = \mathrm{IF}(z_2, g, F). \tag{33}$$

Proof. For ease of reading the proof, we rewrite $z_1 = z^{(1)}$ and $z_2 = z^{(2)}$. We employ a theorem for influence functions of general two-stage estimators by Zhelonkin et al. (2012): For a two-stage estimator with first- and second-stage score functions ψ_1 and ψ_2 and a function h, which is continuously piecewise differentiable in the second variable, fulfilling

$$\mathbb{E}_F[\psi_1(z^{(1)}, S(F))] = 0 \tag{34}$$

and

$$\mathbb{E}_F[\psi_2(z^{(2)}, h(z^{(1)}, S(F)), T(F))] = 0, \tag{35}$$

the influence function is given by

$$\mathrm{IF}(z, T, F) = M^{-1}\Bigg(\psi_2(z^{(2)}, h(z^{(1)}, S(F)), T(F)) \tag{36}$$

$$+ \int \frac{\partial}{\partial \theta}\psi_2(\tilde{z}^{(2)}, \theta, T(F))\frac{\partial}{\partial \nu}h(\tilde{z}^{(1)}, \nu)\,\mathrm{d}F(\tilde{z}) \cdot \mathrm{IF}(z, S, F)\Bigg) \tag{37}$$

where $M = -\int \frac{\partial}{\partial \xi}\psi_2(\tilde{z}^{(2)}, h(\tilde{z}^{(1)}, S(F)), \xi)\,\mathrm{d}F(\tilde{z})$. Here F denotes the distribution function of $z_i = (z_i^{(1)}, z_i^{(2)})$ where $z_i^{(j)} = (x_{ij}, y_{ij}), j = 1, 2, i = 1, \ldots n$. S denotes the functional of the first stage, and T is the functional of the second stage. The influence function of the first-stage estimation is given by $\mathrm{IF}(z, S, F)$.

For the doubly robust two-stage CATE estimators, we have $S(F) = (S_\mu(F), S_\pi(F)) = (\mu, \pi)$ (more precisely, the parameters describing the functions) and

$$h(z, (\mu, \pi)) = \frac{\lambda(\pi(x))(a - \pi(x))}{\pi(x)(1 - \pi(x))}(y - \mu(x, a)) + \mu(x, 1) - \mu(x, 0). \tag{38}$$

The score function of the second-stage regression is given by

$$\psi_2(z^{(2)}, h(z^{(1)}, S(F)), T(F)) = \psi_2(z^{(2)}, h(z^{(1)}, (\mu, \pi)), g) \tag{39}$$

$$= ((a - \pi(x))\lambda^{'}(\pi(x)) + \lambda(\pi(x)))(h(z^{(1)}, (\mu, \pi)) - g) \tag{40}$$

Observe that, with a slight abuse of nation, due to the Neyman-orthogonal loss function, it holds

$$\int \frac{\partial}{\partial \theta}\psi_2(\tilde{z}^{(2)}, \theta, T(F))\frac{\partial}{\partial \eta}h(\tilde{z}^{(1)}, \eta)\,\mathrm{d}F(\tilde{z}) = \int \frac{\partial \psi_2(\cdot, \cdot)}{\partial h}\begin{pmatrix} \frac{\partial h(\cdot,\cdot)}{\partial \mu} \\ \frac{\partial h(\cdot,\cdot)}{\partial \pi} \end{pmatrix}\mathrm{d}F(\tilde{z}) \tag{41}$$

$$= \mathbb{E}[\nabla_\eta(\psi_2)|_{\eta=\eta_0}] \tag{42}$$

$$= 0. \tag{43}$$

Furthermore, we have

$$M = -\int \frac{\partial}{\partial g}\psi_2(z^{(2)}, h(z^{(1)}, (\mu, \pi)), g)\,\mathrm{d}F(z) = -\int 1\,\mathrm{d}F(z) = -1. \tag{44}$$

Therefore, we receive

$$\mathrm{IF}((z^{(1)}, z^{(2)}), (g, \eta), F) = -1 \cdot [\psi_2(z^{(2)}, h(z^{(1)}, \eta), g) + 0] = \mathrm{IF}(z^{(2)}, g, F). \tag{45}$$

□

Proof of Lemma 3

Lemma 3. *Let $\mathcal{H}$ denote the RKHS induced by the Gaussian kernel $K(x, y) = \frac{1}{(\sqrt{2\pi}h)^q}\exp(-\frac{\|x-y\|^2}{2h^2})$ for $x, y \in \mathbb{R}^q$, and let f_D be the optimal solution to the RKHS regression*

$$\min_{f \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda\|f\|_{\mathcal{H}}, \tag{46}$$

on dataset D with $|D| = n$ for $\ell(\cdot)$ is a convex and Lipschitz loss function with Lipschitz constant L. Then

$$||f(D) - f(D^{'})||_{\mathcal{H}} \leq \frac{L}{\lambda n}(\sqrt{(2\pi)}h)^{-q}. \tag{47}$$

Proof. Observe that, for all $x \in \mathbb{R}^q$, the Gaussian kernel norm is given by $K(x,x) = \frac{1}{(\sqrt{2\pi}h)^q}$. Since l is convex and Lipschitz with constant L, it is L-admissable (see (Hall et al., 2013)). Therefore, we can employ a result from Hall et al. (2013) stating that the RKHS norm of minimizers of neighboring datasets can be bounded as

$$||f(D) - f(D^{'})||_{\mathcal{H}} \leq \frac{L}{\lambda n}\sqrt{\sup_{x} K(x,x)}. \tag{48}$$

With our observation above, the result follows. □

G.2 Proof of Theorem 1

Theorem 1 [DP-CATE for finite queries]. *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$ and $\hat{\eta} = (\hat{\mu}, \hat{\pi})$ the estimated nuisance functions. Furthermore, let the dataset D for the second-stage estimation s.t. $|D| = n$. Define*

$$g^{\mathrm{DP}}(\mathbf{x}) := g^{*}(\mathbf{x}) + \sup_{z\in\mathcal{Z}} \left\| \frac{\rho(a,\hat{\pi}(x))}{\mathbb{E}[\lambda(\hat{\pi}(X))]}(\phi(z,\hat{\eta},\lambda(\hat{\pi}(x))) - g^{*}(x)) \right\| \cdot \frac{5\sqrt{2\ln(n)\ln{(2/\delta)}}}{\varepsilon n} \cdot U, \tag{49}$$

where $U \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Thenk $g^{p}(\mathbf{x})$ is (ε,δ)-differentially private.

Proof. First, observe that, if we had access to the ξ-smooth sensitivity $SS_{\xi}(g, D)$ of the doubly robust meta-learner g with $\xi = \frac{\varepsilon}{4(d+2\log(2/\delta))}$, the estimator g^{DP} with

$$g^{\mathrm{DP}}(\mathbf{x}) = g^{*}(\mathbf{x}) + \frac{5\sqrt{2\log(2/\delta)}}{\varepsilon} SS_{\xi}(g,D) \cdot U, \tag{50}$$

where $U \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $\mathbf{x} \in \mathbb{R}^d$, would fulfill (ε,δ) differential privacy (see (Avella-Medina, 2021; Nissim et al., 2007)). However, the ξ-smooth sensitivity is highly difficult, or even impossible, to compute for general function classes such as neural networks. Therefore, we seek an upper bound on $SS_{\xi}(g,D)$ to ensure that the privacy guarantees stay valid while making it feasible to compute the calibration function $r(\cdot)$.

By Lemma 4, we find such an upper bound through

$$SS_{\xi}(g,D) \leq \frac{\sqrt{\log(n)}}{n}\gamma(g,F), \tag{51}$$

where $\gamma(g,F)$ denotes the gross-error sensitivity of g on data distribution F

$$\gamma(g,F) = \sup_{z\in\mathcal{Z}} \|\mathrm{IF}(z,(g,\eta),F)\|. \tag{52}$$

By Lemma 1 and Hirano et al. (2003), we know that

$$\mathrm{IF}(z,(g,\eta),F) = \mathrm{IF}(z,g,F) \tag{53}$$

$$= \frac{\rho(a,\pi_0(x))}{\mathbb{E}[\lambda(\pi_0(x))]}\left(\frac{\lambda(\pi(x))}{\rho(a,\pi(x))}\frac{a-\pi(x)}{\pi(x)(1-\pi(x))}(y-\mu(x,a)) + \mu(x,1) - \mu(x,0) - g(x)\right) \tag{54}$$

for

$$\rho(a,\pi(x)) := (a - \pi(x))\lambda^{'}(\pi(x)) + \lambda(\pi(x)). \tag{55}$$

With the definition of $\phi(z,\eta,\lambda(\pi(x))) := \frac{\lambda(\pi(x))}{\rho(a,\pi(x))}\frac{a-\pi(x)}{\pi(x)(1-\pi(x))}(y-\mu(x,a)) + \mu(x,1) - \mu(x,0)$, we receive

$$\mathrm{IF}(z,(g,\eta),F) = \frac{\rho(a,\pi_0(x))}{\mathbb{E}[\lambda(\pi_0(x))]}(\phi(z,\eta,\lambda(\pi(x))) - g(x)). \tag{56}$$

Taking the supremum over all samples from the dataspace $\mathcal{Z}$ and evaluating the IF at the trained function g^{*}. $\hat{\pi}$ and $\hat{\mu}$ states the desired result. □

G.3 Proof of Theorem 2

Theorem 2 [DP-CATE for functional queries]. Let $\hat{\mu}(a,x)$, $\hat{\pi}(x)$ denote the nuisance estimators trained in stage 1 and $z = (a,x,y)$ a data sample from dataset D with $|D| = n$ and $x \in \mathbb{R}^q$. Let $\mathcal{H}$ denote the RKHS induced by kernel $K(x,y) = \frac{1}{\sqrt{2\pi}h}\exp(-\frac{\|x-y\|^2}{2h^2})$, and let $\ell(\cdot)$ be a convex and Lipschitz loss function with Lipschitz constant L. Furthermore, define g^{*} as the second stage regression solving Equation 3 through

$$g^{*} \in \arg\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(\rho(a_i,\pi(x_i))\phi(z_i,\hat{\eta},\lambda(\pi(x_i))) - g(x_i)\big) + \lambda\|g\|_{\mathcal{H}}^{2} \tag{57}$$

for

$$\rho(a, \pi(x)) := (a - \pi(x))\lambda^{'}(\pi(x)) + \lambda(\pi(x)) \tag{58}$$

and

$$\phi(z, \eta, \lambda(\pi(x))) := \frac{\lambda(\pi(x))}{\rho(a, \pi(x))} \frac{a - \pi(x)}{\pi(x)(1 - \pi(x))}(y - \mu(x, a)) + \mu(x, 1) - \mu(x, 0). \tag{59}$$

Furthermore, let G be the sample path of a zero-centered Gaussian process with covariance function $K(\cdot)$. Then the function

$$g^{\mathrm{DP}} := g^{*} + \frac{4L\sqrt{2\ln{(2/\delta)}}}{(\sqrt{2\pi}h)^{q}\lambda n\varepsilon} \cdot G \tag{60}$$

is (ε, δ)-differentially-private.

Proof. Let G be the sample path of a zero-centered Gaussian process with covariance function $K(\cdot)$. For our proof, we make use of Corollary 9 by Hall et al. (2013):

For $f \in \mathcal{H}$, where $\mathcal{H}$ is the RKHS corresponding to the kernel K, the release of

$$f_D^p = f_D + \frac{\Delta c(\delta)}{\varepsilon}G \tag{61}$$

is (ε, δ)-differentially private for

$$c(\delta) \geq \sqrt{2\log(\frac{2}{\delta})} \tag{62}$$

and

$$\Delta \geq \sup_{D \sim D'} \|f_D - f_{D'}\|_{\mathcal{H}}. \tag{63}$$

Therefore, for $\Delta^{*} = \sup_{D \sim D'} \|g_D - g_{D'}\|_{\mathcal{H}}$,

$$g^{\mathrm{DP}} = g^{*} + \frac{\Delta^{*}\sqrt{2\log(2/\delta)}}{\varepsilon}G \tag{64}$$

is (ε, δ)-differentially private. From Lemma 3, we know that

$$\sup_{D \sim D'} \|g_D - g_{D'}\|_{\mathcal{H}} \leq \frac{L}{\lambda n}(\sqrt{(2\pi)}h)^{-q}. \tag{65}$$

Thus, the desired result follows. □

H Experiments

H.1 Synthetic dataset generation

We consider two different data-generation settings with different complexity. Both settings follow the mechanism described in (Oprescu et al., 2019):

$$X_i \sim \mathcal{U}[0,1]^p, \tag{66}$$

$$A_i = \mathbf{1}\{(X^T\beta)_i \geq \eta_i\}, \tag{67}$$

$$Y_i = \theta(X_i)A_i + (X^T\gamma)_i + \epsilon_i, \tag{68}$$

where $\eta_i, \epsilon_i \sim \mathcal{U}[-1,1]$ and β, γ have support with values drawn from $\mathcal{U}[0,0.3]$ and $\mathcal{U}[0,1]$. The dimension of the covariates is set to $p = 2$ for Dataset 1 and $p = 30$ for Dataset 2. In Dataset 1, the conditional treatment effect $\theta(x)$ is defined as

$$\theta(x) = \exp(2x_0) + 3\sin(4x_0). \tag{69}$$

In Dataset 2, $\theta(x)$ is defined as

$$\theta(x) = \exp(2x_0) + 3\sin(4x_1). \tag{70}$$

For each setting, we draw 3000 samples which we split into train (90%) and test (10%) sets.

H.2 Medical datasets

MIMIC-III: We showcase DP-CATE on the MIMIC-III dataset (Johnson et al., 2016), which includes electronic health records (EHRs) from patients admitted to intensive care units. We extract 8 confounders (heart rate, sodium, red blood cell count, glucose, hematocrit, respiratory rate, age, gender) and a binary treatment (mechanical ventilation) using an open-source preprocessing pipeline (Wang et al., 2020). We define the outcome variable as the red blood cell count after treatment which we adapt to be more responsive to the treatment ventilation. To extract features from the patient trajectories in the EHRs, we sample random time points and average the value of each variable over the ten hours prior to the sampled time point. All samples with missing values and outliers are removed from the dataset. We define samples with values smaller than the 0.1st percentile or larger than the 99.9th percentile of the corresponding variable as outliers. Our final dataset contains 14719 samples, which we split into train (90%) and test (10%) sets.

TCGA: The Cancer Genome Atlas (TCGA) dataset (Weinstein et al., 2013) contains a comprehensive and diverse collection of gene expression data collected from patients with different types of cancer. We consider the gene expression measurements of the 4,000 genes with the highest variability which we employ as our features X. The study cohort of consisted of 9659 patients. We model the binary treatment based on the sum of the 10 covariates with the highest variance and assign a constant treatment effect in the sum of the covariates.

H.3 Implementation details

Our experiments are implemented in Python. We provide our code in our GitHub repository: `https://anonymous.4open.science/r/PrivateCATE_anonymous-4B34/README.md`.

Our DP-CATE is model-agnostic and highly flexible. Therefore, we implement multiple versions of the R- and the DR-Leaner with varying base learners. For the outcome and the propensity estimation, we always employ a multilayer perceptron regression and classification model, respectively. The models consist of one layer of width 32 with ReLu activation function and were optimized via Adam at a learning rate of 0.01 and batch size 128.

For our experiments on the finite DP-CATE, we implement the pseudo-outcome regression in the second stage as a Random forest (RF) and a linear regression (LR) model with default parameter specifications and a neural network (NN) with two hidden layers of width 32 with ReLu activation function trained in the same manner as the nuisance models. In the experiments on our functional DP-CATE, we employ a Gaussian kernel ridge regression with $m = 50$ basis functions and default regularization parameter $\lambda = 1$. Furthermore, our functional DP-CATE requires the specification of the Lipschitz constant L. This constant is either known based on the employed loss, e.g., $L = 1$ for the L1 loss, or can be upper-bounded. In our settings, we employ the L2 loss on a bounded domain. Therefore, although the L2 loss itself is not Lipschitz, we can calculate L numerically as the upper bound of the domain. We did not perform hyperparameter optimization, as our model-agnostic framework is applicable to any prediction model.

Our framework requires calculating the supremum of the influence functions. We implemented the maximization problem through mathematical optimization using the L-BFGS-B, a limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm for solving nonlinear optimization problems with bounded variables. The solver was run with default parameters.

I Further results

Below, we present further experimental results underlining our findings in Section 5.

In Fig. 14 and Fig. 15, we show the estimation performance in terms of PEHE of DP-CATE for finite queries compared to the underlying non-private model (in absolute values and relative to the PEHE of the non-private learner). Our results align with our findings in the main paper.

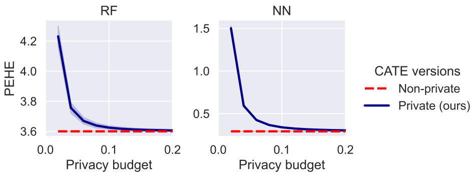


Figure 14: PEHE of DP-CATE for finite queries with respect to privacy budget ε on dataset 1

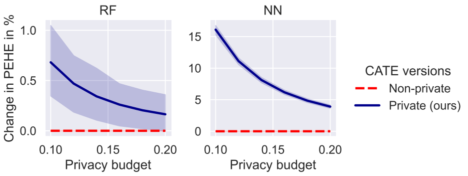


Figure 15: DP-CATE PEHE in % of the non-private PEHE for $\varepsilon \in [0.1, 0.2]$ on dataset 1.