DIFFERENTIALLY PRIVATE LEARNERS FOR HETEROGENEOUS TREATMENT EFFECTS

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

ABSTRACT

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

1 INTRODUCTION

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

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

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

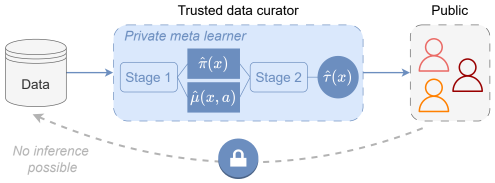


Figure 1: **Setting: CATE estimation under DP.** Only the trusted data curator can access the data, while published CATE estimates do not allow private information about individuals to be inferred.

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

In this paper, we estimate the CATE from observational data under DP (Fig. 1). Specifically, we propose DP-CATE, an output perturbation mechanism for Neyman-orthogonal CATE estimators that satisfies DP. Neyman-orthogonal estimators are generally preferred over standard plug-in CATE estimators, as they are less dependent on the estimation errors of nuisance functions (Morzywolek et al., 2023).

Our DP framework is highly flexible and can be combined with all weighted Neyman-orthogonal 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 base learner. To the best of our knowledge, we are the first to provide a framework for Neyman-orthogonal CATE estimation under differential privacy. Our `DP-CATE` is designed for two use cases relevant to medical 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 subgroups. 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 Neyman-orthogonality and privacy, which allows us to base our `DP-CATE` 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 add to the CATE function estimated through RKHS regression (→ our Theorem 3).

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 and/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 to 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:[1] (1) We propose a novel framework for CATE estimation that is differentially private and Neyman-orthogonal. (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 Neyman-orthogonal 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 and can thus be instantiated with arbitrary ML models (e.g., neural networks). Neyman-orthogonal learners (also known as debiased learners) have several additional benefits. (i) The learners achieve quasi-oracle efficiency[2], which is guaranteed due to the first-order insensitivity to errors in the estimation of nuisance functions. As a result, the estimators are asymptotically equivalent to the oracle estimator (= the estimator that has access to the oracle nuisance functions), thereby mitigating the finite-sample bias arising from the misspecification of the nuisance functions (Mackey et al., 2018; Morzywolek et al., 2023). (ii) The DR-learner is further doubly robust: it yields consistent estimation even when either the outcome or the propensity model is not correctly specified. (iii) The R-learner is less sensitive to overlap violations due to its inherent weighting of the pseudo-outcomes.

In this paper, we focus on DP for Neyman-orthogonal meta-learners. We derive a framework for the R-learner in the main paper and provide an extension to the DR-learner in Supplement B.

---

[1] The source code is available at our GitHub repository.

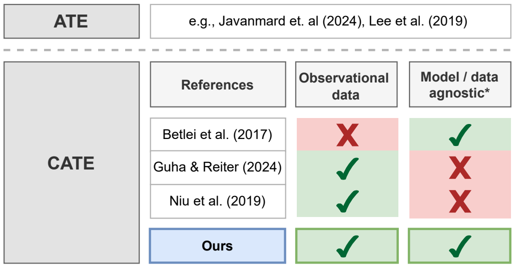
[2] Informally, quasi-oracle efficiency means that the target model is learned almost equally well using either the estimated nuisance functions or the ground truth.

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. For example, one strategy is output perturbation, where one adds calibrated random noise to the non-private model prediction prior to its release in order to ensure DP (e.g., Chaudhuri et al., 2011; Zhang et al., 2022).

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* ML 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.

DP in treatment effect estimation: The existing literature on DP 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. It makes population-wide estimates and, thus, 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): they (i) are either restricted to interventional data from RCTs 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 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 models such as neural networks. In sum, none of the above works provide a method for CATE estimation under DP where both observational data and different types of ML models can be used.



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

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

Figure 2: Comparison of relevant literature.

Research gap: So far, a DP framework for CATE estimation from observational data with meta-learners is missing. We are thus the first to propose a Neyman-orthogonal framework for CATE estimation that fulfills DP.

3 PROBLEM FORMULATION

Notation: We write random variables as 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 outcome under intervention a by $Y(a)$.

Setting: We consider a dataset $\bar{D} := \{(X_i, A_i, Y_i)\}_{i=1,\ldots,2n}$, consisting of observed confounders X in a bounded domain $\mathcal{X} \subseteq \mathbb{R}^q$, a binary treatment $A \in \{0, 1\}$, and a bounded discrete or continuous outcome $Y \in \mathcal{Y}$, where $Z_i := (X_i, A_i, Y_i) \sim P$ i.i.d., $Z_i \in \mathcal{Z}$. 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 individuals 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). [3] 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 Neyman-orthogonal meta-learners (Sec. 3.1) while (ii) ensuring differential privacy (Sec. 3.2), which we briefly review in the following.

---

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

3.1 NEYMAN-ORTHOGONAL META-LEARNERS FOR CATE ESTIMATION

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

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

where $\eta = (\mu, \pi)$ are nuisance functions, $\lambda(\cdot) > 0$ is a *weight function*, and $\Lambda(g)$ is a regularization term. We denote a population minimizer of $R_P(g, \eta, \lambda(\pi))$ with $g^*(\cdot; \eta) = \arg\min_{g \in \mathcal{G}} R_P(g, \eta, \lambda(\pi))$ (Hirano et al., 2003; Morzywolek et al., 2023). However, R_P cannot be directly estimated and subsequently minimized given the data $\bar{D}$, as it depends on the unknown nuisance functions π and μ. One could employ the estimated nuisance functions (i.e., $\hat{\pi}_{\bar{D}}$ and $\hat{\mu}_{\bar{D}}$) here, but then, their estimation errors propagate into the minimization of R_P.

A popular approach to circumvent the above problem is to use Neyman-orthogonal meta-learners. Formally, such meta-learners operate in two stages (e.g., Kennedy, 2023b; Nie & Wager, 2020). Let the dataset $\bar{D}$ be a disjoint union of the two subsets $\tilde{D}$ and D of size n, i.e., $\bar{D} = \tilde{D} \cup D$. In the first stage, the meta-learners estimate *nuisance functions* $\hat{\eta}_{\tilde{D}} = (\hat{\pi}_{\tilde{D}}, \hat{\mu}_{\tilde{D}})$ on dataset $\tilde{D}$, and, in the second stage, we minimize the adapted Neyman-orthogonal population risk function

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

with

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

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

on dataset D, where $\hat{\eta}_{\tilde{D}}$ is used in place of η (Morzywolek et al., 2023). The use of the loss function allows for quasi-oracle efficiency of the final estimator. In the following, we denote the estimated population risk R_P via a loss R_D that is dependent on the data D, namely

$$R_D(g, \eta, \lambda(\pi)) = \frac{1}{n}\sum_{i=1}^{n} \rho(A_i, \pi(X_i))\,(\phi(Z_i, \eta, \lambda(\pi(X_i))) - g(X_i))^2 \,+\, \Lambda(g). \tag{6}$$

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, Neyman-orthogonal meta-learners such as the R-learner achieve state-of-the-art performance (Curth & van der Schaar, 2021), and the orthogonality property makes the models less sensitive to the misspecification of the nuisance functions (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 does not significantly affect the estimated outcome (Dwork, 2006; Dwork & Lei, 2009). Specifically, for a given *privacy budget* ε, DP ensures that the 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*, denoted as $D \sim D^{'}$, if their Hamming distance equals one, i.e., $d_{\mathrm{H}}(D, D^{'}) = 1$.

Definition 1 (Differential privacy (Dwork & Lei, 2009)). *A function $\hat{f}_D(\mathbf{x}) : \mathcal{X}^d \mapsto \mathbb{R}^d$ trained on a 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*

$$P(\hat{f}_D(\mathbf{x}) \in S) \leq \exp(\varepsilon) \cdot P(\hat{f}_{D'}(\mathbf{x}) \in S) + \delta \qquad \textit{for all } \mathbf{x} \in \mathcal{X}^d. \tag{7}$$

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 (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*.

Definition 2 (Gaussian noise privacy mechanism (Dwork & Roth, 2014))**.** *Let* $\hat{f}_D(\mathbf{x}) : \mathcal{X}^d \mapsto \mathbb{R}^d$ *be a function on dataset* D *with* l_2*-sensitivity* $\Delta_2(\hat{f}) = \sup_{D \sim D', \mathbf{x} \in \mathcal{X}^d} \|\hat{f}_D(\mathbf{x}) - \hat{f}_{D'}(\mathbf{x})\|_2$ *and* $\mathbf{U} \sim \mathcal{N}(0, \sigma \boldsymbol{I}_d)$ *for* $\sigma \geq \frac{1}{\varepsilon}\sqrt{2 \ln(1.25/\delta)}\, \Delta_2(\hat{f})$*. Then, the output perturbation mechanism returns* $\mathcal{M} : \hat{f}_{\mathrm{DP}}(\mathbf{x}) = \hat{f}_D(\mathbf{x}) + \mathbf{U}$ *that preserves* (ε, δ)*-differential privacy.*

Definition 2 describes how to ensure DP for a given prediction. However, this requires estimating the training sensitivity $\Delta_2(\hat{f})$ of the employed model $\hat{f}$, 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 Neyman-orthogonal CATE estimation under differential privacy. Specifically, we aim to derive an (ε, δ)-differentially private version of $\hat{g}_D(\cdot; \eta) = \arg\min_{g \in \mathcal{G}} R_D(g, \eta, \lambda(\pi))$ of the form

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \eta) = \hat{g}_D(\mathbf{x}; \eta) + r(\varepsilon, \delta, \hat{g}_D, \eta) \cdot \mathbf{U}, \tag{8}$$

where $\hat{g}_D(\mathbf{x}; \eta) = (\hat{g}_D(x_1; \eta), \ldots, \hat{g}_D(x_d; \eta))^\top$, $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and where $r(\cdot)$ is a *calibration function*. Importantly, we consider *arbitrary* working model classes $\mathcal{G}$.

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 Neyman-orthogonal 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 Eq. (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.

The above task is highly *non-trivial* as the identification of CATE relies on 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. Accounting for the bias introduced by perturbation is especially difficult for causal tasks, as CATE estimation requires counterfactual quantities $Y(0)$ and $Y(1)$ from which *only* one outcome per individual is observed, while differences, $Y(1) - Y(0)$, and, thus, the estimation target (CATE), $\tau(x)$, are *never* observed (Pearl, 2010). Instead of blindly introducing noise to the estimation setup to guarantee DP, we must cautiously calibrate the noise in a targeted setup to retain the consistency and efficiency of existing CATE estimation methods.

4 OUR FRAMEWORK: DP-CATE

Overview: We now present our DP-CATE framework aimed at CATE estimation for Neyman-orthogonal 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. Second, we retain the abilities of existing CATE estimation methods.

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):** We aim to report a number d of CATE estimates (e.g., treatment effects across different age groups). ⇒ *How do we solve this?* We draw upon the shared property of robustness and privacy of ML models in terms of insensitivity to outliers and small measurement errors (Dwork & Lei, 2009). We then propose calibrating the noise added with a function $r(\cdot)$ that depends on the meta-learner's influence function. Importantly, our model-agnostic approach preserves the quasi-oracle guarantees of the non-private model stemming from the orthogonal loss function.

② DP-CATE **for functional queries (→ Sec. 4.2):** We release an estimate $\hat{g}_{\mathrm{DP}}$ of the complete CATE function τ, which can then be queried arbitrarily often (e.g, as in clinical decision support systems). 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 an output perturbation method based on Gaussian processes that is valid for all functional CATE estimates solving Eq. (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

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

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

For this, we borrow ideas from the literature that observed similarities between differential privacy and robust statistics (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 a single observation in D influences CATE estimation and the model output. 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, $T = T(P) \in \mathbb{R}^d$. The* gross-error sensitivity *of T at z under P is given by the supremum of the l_2 norm $\|\cdot\|_2$ of the influence functions, i.e.,*

$$\gamma(T, P) := \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; P)\|_2, \tag{9}$$

where $\mathrm{IF}(z, T; P) = \frac{\mathrm{d}}{\mathrm{d}t}\big[T((1-t)P + t\delta_z)\big]\big|_{t=0}$ is an influence function (IF) of T at z under distribution with density/probability mass function P, and δ_z denotes the Dirac-delta function at z.

Next, we derive the IF of Neyman-orthogonal meta-learners for the CATE. Observe that $T(P)$ depends on the data distribution directly through P and indirectly through the first-stage estimation.

We now state our main theorem for differentially private CATE estimation with a known number d of queries. The intuition behind constructing 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 meta-learners can be upper bounded by the gross-error sensitivity γ of the second stage regression. [4]. Specifically, now $T(P) = g^*(\mathbf{x};\eta)$ and $T(D) = \hat{g}_D(\mathbf{x};\eta)$ if the nuisance functions were known; and $T(P) = g^*(\mathbf{x};\hat{\eta}_{\tilde{D}})$ and $T(D) = \hat{g}_D(\mathbf{x};\hat{\eta}_{\tilde{D}})$ if the nuisance functions are estimated.

Theorem 1 (DP-CATE for finite queries)**.** *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$ and $\hat{\eta}_{\mathrm{DP}} = (\hat{\mu}_{\mathrm{DP}}, \hat{\pi}_{\mathrm{DP}})$ the nuisance functions estimated on dataset $\tilde{D}$ in a $(\varepsilon/2, \delta/2)$-differentially private manner of choice.*[5] *Furthermore, let D be the training dataset for the second-stage model with $|D| = n$. For $z = (a, x, y) \in \mathcal{Z}$, we define*

$$\hat{g}_{\mathrm{DP}}(\mathbf{x};\hat{\eta}_{\mathrm{DP}}) := \hat{g}_D(\mathbf{x};\hat{\eta}_{\mathrm{DP}}) + \underbrace{\gamma\big(T, D\big) \cdot c(\varepsilon, \delta, n)}_{r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}})} \cdot \mathbf{U}, \qquad T(P) = g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}}), \tag{10}$$

$$\gamma\big(T, P\big) = \sup_{z \in \mathcal{Z}} \Big\| h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}})\, \rho(a, \hat{\pi}_{\mathrm{DP}}(x)) \big(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(x))) - g^*(x;\hat{\eta}_{\mathrm{DP}})\big) \Big\|_2, \tag{11}$$

where $\gamma\big(T, D\big)$ is a sample gross-error sensitivity with $T(D)$ instead of $T(P)$ in Eq. 9, $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $c(\varepsilon, \delta, n) := 5\sqrt{2\ln(n)\ln(2/\delta)}\big/(\varepsilon n)$, and where $h(g^*, \mathbf{x}, z, \hat{\eta}_{\mathrm{DP}}) \in \mathbb{R}^d$ and $g^*(\cdot;\hat{\eta}_{\mathrm{DP}})$ depend on the machine learning model employed for the second stage regression. Then, $\hat{g}_{\mathrm{DP}}(\mathbf{x};\hat{\eta}_{\mathrm{DP}})$ is (ε, δ)-differentially private.

Proof. We prove Theorem 1 in Supplement F. To do so, we first state the IF of general weighted Neyman-orthogonal 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. ☐

---

[4] See Lemma 5 in Supplement E

[5] As we need to query the nuisance functions multiple times, the privatization method needs to be suitable for functions, such as gradient perturbation through DP-SGD (Abadi et al., 2016).

In the case of the R-learner, Theorem 1 yields the following gross-error sensitivity $\gamma^{\mathrm{R}}\big(T, P\big)$:

$$\gamma^{\mathrm{R}}(T,P) = \sup_{z \in \mathcal{Z}} \left\| h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}}) \underbrace{(a - \hat{\pi}_{\mathrm{DP}}(x))^2}_{\rho(a, \hat{\pi}_{\mathrm{DP}})} \underbrace{\left(\frac{y - \hat{\mu}_{\mathrm{DP}}(x,a)}{a - \hat{\pi}_{\mathrm{DP}}(x)} + \hat{\mu}_{\mathrm{DP}}(x,1) - \hat{\mu}_{\mathrm{DP}}(x,0) \right.}_{\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))} \left. - g^*(x; \hat{\eta}_{\mathrm{DP}}) \right) \right\|_2 . \quad (12)$$

In the following, we present two corollaries stating the form of the function $h(g^*, \mathbf{x}, z, \hat{\eta}_{\mathrm{DP}})$ for parametric models and neural networks well as the non-parametric Nadaraya-Watson estimator.

Corollary 1 (Parametric second-stage regression)**.** *If the second-stage regression is a smooth parametric model, namely* $\mathcal{G} = \{g(x;\theta) : \mathcal{X} \mapsto \mathbb{R}, \theta \in \Theta \subseteq \mathbb{R}^p\}$*, then in Theorem 1*

$$g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) = g(\cdot; \theta^*) \quad \textit{and} \quad h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}}) = 2\, J_\theta\big[g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})\big] \cdot H_\theta^{-1} \cdot \nabla_\theta\big[g^*(x; \hat{\eta}_{\mathrm{DP}})\big], \quad (13)$$

where $\theta^* = \arg\min_{\theta \in \Theta} R_P(g, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))$*;* $J_\theta \in \mathbb{R}^{d \times p}$ *is a Jacobian matrix wrt.* θ*;* $H_\theta = \nabla^2_\theta\big[R_P(g^*, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))\big] \in \mathbb{R}^{p \times p}$ *is a Hessian matrix; and* ∇_θ *is a gradient.*

Proof. We prove Corollary 1 in Supplement F. □

A note on applicability: Corollary 1 requires access to and invertibility of H_θ. This might restrict the applicability of certain deep neural networks for the second-stage regression. However, we note that, in Theorem 1, we are only interested in predicting the CATE at certain values x. This motivates the local CATE estimation through kernel weighting. We present our alternative approach below.

Corollary 2 (Non-parametric second-stage regression)**.** *If the second-stage regression is a kernel ridge regression with* $\Lambda(g) = \lambda \|g\|^2_{\mathcal{H}}$*, where* $\mathcal{G} = \mathcal{H}$ *is a reproducing kernel Hilbert space (RKHS) induced by a kernel* $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$*, then in Theorem 1*

$$g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) = \big(\mathbb{L}_\rho + \lambda \mathbb{I}\big)^{-1} S(\cdot) \quad \textit{and} \quad h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}}) = \big(\mathbb{L}_\rho + \lambda \mathbb{I}\big)^{-1} K(\cdot, x)(\mathbf{x}), \quad (14)$$

where $\mathbb{L}_\rho : \mathcal{H} \mapsto \mathcal{H}$ *is a weighted covariance operator* $\mathbb{L}_\rho f(\cdot) = \mathbb{E}\big[\rho(A, \hat{\pi}_{\mathrm{DP}}(X))\, K(\cdot, X) f(X)\big]$*;* $\lambda \mathbb{I} f(\cdot) = \lambda f(\cdot)$ *is a scaling operator;* $S \in \mathcal{H}$ *is a cross-covariance functional* $S(\cdot) = \mathbb{E}\big[\rho(A, \hat{\pi}_{\mathrm{DP}}(X))\, K(\cdot, X)\, Y\big]$.

Proof. We prove Corollary 2 in Supplement F. □

Scalability: The complexity of calculating $\gamma(\cdot)$ is *independent* of the size of the dataset once the second-stage model $\hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}})$ is fitted.

Theorem 2 (Quasi-oracle efficiency of DP-CATE)**.** *The privatization of the nuisance functions does not affect the quasi-oracle efficiency of the final CATE estimator. Specifically, for machine learning models* f *with sufficiently fast convergence rate in* $o_P(n^{-1/4})$*, privatization through output or gradient perturbation still retains the stated rate, rendering the final estimate efficient:*

$$\|g^*(\cdot; \eta) - \hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{DP}})\|^2_{L_2} \lesssim R_P\big(g^*(\cdot; \hat{\eta}_{\mathrm{DP}}), \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})\big) - R_P\big(g^*(\cdot; \eta), \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})\big) \quad (15)$$

$$+ R_2(\hat{\eta}_{\mathrm{DP}}, \eta) + O_P(n^{-2}) + \|g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) - \hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{D}})\|^2_{L_2}, \quad (16)$$

where

$$\|\hat{\eta}_{\mathrm{DP}} - \eta\|^2_{L_2} \leq \|\hat{\eta}_{\mathrm{DP}} - \hat{\eta}_D\|^2_{L_2} + \|\hat{\eta}_D - \eta\|^2_{L_2} \quad (17)$$

Proof. We prove Theorem 2 in Supplement F. □

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 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 calibration function that does not depend on d. More precisely, the added noise should be a function itself to guarantee the privacy of the CATE function. For (I), we propose to add a calibrated Gaussian process (GP) to the predicted CATE. [6]

---

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

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* $K(s,t) := \operatorname{Cov}(X_s, X_t)$.

We are left with answering question (II) from above: *how to calibrate the GP noise to fulfill DP?* We make the following important observations: (i) The Neyman-orthogonality of the loss again is not affected by the perturbation of the nuisance estimates. Hence, we again simply need to focus on the second stage of the meta-learner. (ii) If g lies in a reproducing kernel Hilbert space (RKHS), we can calibrate the GP 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 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 $\sup_{D\sim D', \mathbf{x}\in\mathcal{X}^d} ||\hat{f}_D(\mathbf{x}) - \hat{f}_{D'}(\mathbf{x})||_2$ to calibrate the Gaussian noise variable. Similarly, we now require knowledge about $\sup_{D\sim D'} ||\hat{f}_D - \hat{f}_{D'}||_{\mathcal{H}}$ where f specifies an RKHS regression to calibrate the Gaussian process. 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) in our setting.

Lemma 1. *Let* $\mathcal{H}$ *denote the RKHS induced by the Gaussian kernel* $K(x,x') = (\sqrt{2\pi}h)^{-q}\exp(-\|x-x'\|_2^2/(2h^2))$ *for* $x, x' \in \mathcal{X} \subseteq \mathbb{R}^q$*, and let* $\hat{f}_D$ *be the optimal solution to the RKHS regression*

$$\hat{f}_D(\cdot) = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} w(X_i)\cdot \ell(f(X_i), Y_i) + \lambda\|f\|_{\mathcal{H}}^2, \tag{18}$$

where $w(\cdot) > 0$ *is a weight function,* D *is a dataset with* $|D| = n$*, and* $\ell(\hat{y}, y)$ *is a convex and Lipschitz loss function in* $\hat{y}$ *with Lipschitz constant* L*. Then, we have*

$$||f_D - f_{D'}||_{\mathcal{H}} \leq \sup_{x\in\mathcal{X}}[w(x)]\,\frac{L}{\lambda n}\left(\sqrt{(2\pi)}h\right)^{-q}. \tag{19}$$

Proof. We prove Lemma 1 in Supplement F. □

We now use the above results and present our DP framework for CATE functions: (i) Stage 1: We estimate the $(\varepsilon/2, \delta/2)$-differentially private nuisance functions $\hat{\mu}_{\mathrm{DP}}$ and $\hat{\pi}_{\mathrm{DP}}$ through *any* parametric or non-parametric machine learning method and perturbation method. (ii) Stage 2: We perform a Gaussian kernel regression to minimize Eq. (3). (iii) We calibrate a suitably chosen Gaussian process based on Lemma 1 and add the resulting GP to the CATE function. The quasi-oracle efficiency of our framework directly follows from Theorem 2 and (Foster & Syrgkanis, 2019). We present the pseudo-code for DP-CATE in Supplement C.

Theorem 3 (DP-CATE for functional queries). *Let* $\hat{\mu}_{\mathrm{DP}}$ *and* $\hat{\pi}_{\mathrm{DP}}$ *denote the* $(\varepsilon/2, \delta/2)$*-differentially private nuisance estimators trained in stage 1 on* $\tilde{D}$*. Let* $z = (a, x, y)$ *be a data sample from dataset* D *with* $|D| = n$ *and* $x \in \mathcal{X} \subseteq \mathbb{R}^q$*. Let* $\mathcal{H}$ *denote the RKHS induced by the kernel* $K(x,x') = (\sqrt{2\pi}h)^{-q}\exp(-\|x-x'\|_2^2/2h^2)$*, and let* $\ell(\cdot)$ *be a convex and Lipschitz loss function with Lipschitz constant* L*. We define* $\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}})$ *as the second stage regression solving Eq. (3) via*

$$\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) = \arg\min_{g\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \rho(A_i, \hat{\pi}_{\mathrm{DP}}(X_i))\,\ell\big(g(X_i), \phi(Z_i, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(X_i)))\big) + \lambda\|g\|_{\mathcal{H}}^2. \tag{20}$$

Furthermore, let $U(\cdot) \in \mathcal{H}$ *be the sample path of a zero-centered Gaussian process with covariance function* $K(x,x')$*. Then,* (ε, δ)*-differential privacy is guaranteed by*

$$\hat{g}_{\mathrm{DP}}(\cdot;\hat{\eta}_{\mathrm{DP}}) := \hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) + \underbrace{\sup_{(a,x)\in\{0,1\}\times\mathcal{X}}\left[\rho(a, \hat{\pi}_{\mathrm{DP}}(x))\right]\frac{4L\sqrt{2\ln(2/\delta)}}{\left(\sqrt{2\pi}h\right)^{q}\lambda n\varepsilon}}_{r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}})}\cdot U(\cdot). \tag{21}$$

Proof. We prove Theorem 3 in Supplement F. □

Importantly, for both R- and DR-learner, $\sup_{(a,x)\in\{0,1\}\times\mathcal{X}}\left[\rho(a,\hat{\pi}_{\mathrm{DP}}(x))\right] \leq 1$. Also, 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.[7] We further require the second stage in the meta-learner to be a Gaussian kernel 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* machine learning model can be used for nuisance estimation. In Supplement C, we present an algorithm for releasing private outputs of the function $\hat{g}_{\mathrm{DP}}(\cdot;\hat{\eta}_{\mathrm{DP}})$.

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 via both a neural network (NN) and the Kernel regression estimator (KR). 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 G. We emphasize again that there are no suitable baselines for our task.

Performance metrics: As explained in Sec. 2, there are no flexible CATE meta-learners ensuring DP. Thus, 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 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 effects* (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 observed in real-world data. Thus, we follow common practice and evaluate DP-CATE on synthetic data, which allows us to access the ground-truth CATE and compute the PEHE (e.g., Kennedy, 2023a; Oprescu et al., 2019). We evaluate DP-CATE on 300 queries.

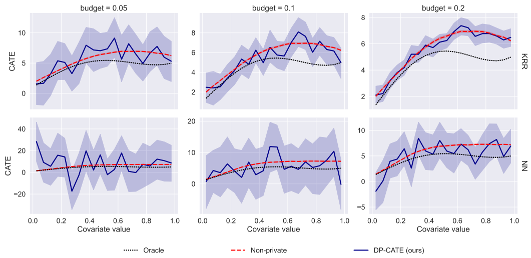


Figure 3: **Dataset 1** (finite queries). Predictions under different base learners and privacy budgets.

We consider two settings with different treatment effect complexities following Oprescu et al. (2019). • **Dataset 1** contains two observed confounders from which only one influences the CATE. This allows us to visualize the CATE function and the effect of privatization on the prediction for varying covariate values. • **Dataset 2** contains 30 observed confounders and multiple dimensions influence the CATE. Details are in Supplement G.

Results for finite queries: • **Dataset 1**: Fig. 3 shows the predictions for different base learners 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 estimator. (2) Our DP-CATE shows consistent patterns for different base learners. For example, the predictions under both KR and NN are almost identical, showing the flexibility and robustness of our framework. • **Dataset 2**: Fig. 5 shows the PEHE. Note that we directly compare DP-CATE for finite queries (combined with KR and NN) and DP-CATE for functional queries (kernel). Again, we find that our DP-CATE performs as expected: the PEHE decreases with increasing privacy budget and converges towards that of the non-private learner.

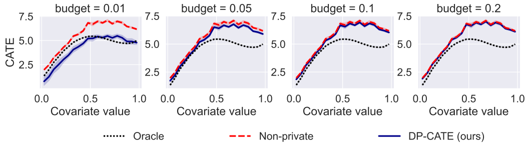


Figure 4: **Dataset 1** (functional queries). Predictions under different privacy budgets.

---

[7] For example, for the l_1 loss, the Lipschitz constant L equals 1; for the Huber loss, L equals the loss parameter δ; and, for the truncated l_2 loss, the constant equals the gradient at the truncation value.

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 robust across datasets and base learner specifications.

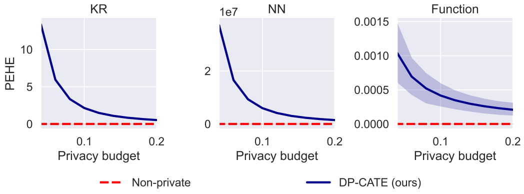


Figure 5: **Dataset 2**. Prediction errors under different privacy budgets of DP-CATE (finite) on KR and NN and DP-CATE (functional) over 10 runs. Plots centered at the PEHE of the original learner.

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 G.

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). 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., $d = 2659$). Overall, the loss in precision due to DP (i.e., when comparing DP-CATE to the non-private learner) is fairly small.

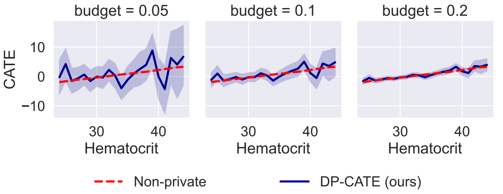


Figure 6: **MIMIC-III** (finite queries). Our DP-CATE generates private estimates of the effect of ventilation for different levels of hematocrit.

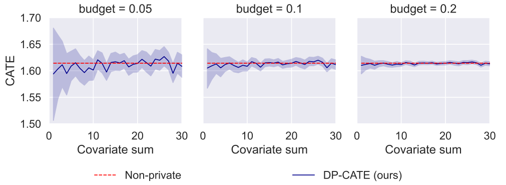


Figure 7: **TCGA** (finite queries). DP-CATE consistently estimates the constant treatment effect across the sum of all covariate values.

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 Neyman-orthogonal 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 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 focus on the popular R-learner due to its favorable theoretical properties. Nevertheless, our DP-CATE can be applied to any other Neyman-orthogonal meta-learner. In Supplement B, we thus provide an extension to the DR-learner. Therein, we also provide additional numerical experiments. At a technical level, the weight function for the DR-learner simplifies to $\lambda^{\mathrm{DR}}(\pi(x)) = 1$ in contrast to the weight function of the R-learner $\lambda^{\mathrm{R}}(\pi(x)) = \pi(x)\,(1 - \pi(x))$. However, $\lambda^{\mathrm{DR}}(\pi(x))$ is less suitable for DP because it needs a larger noise term during the perturbation (see Supplement B for a detailed discussion), which hinders downstream performance. Hence, we recommend employing DP-CATE with the R-learner in practice.

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.

ACKNOWLEDGEMENTS

We thank Dennis Frauen and Lars van der Laan for their helpful feedback on our manuscript. Our research was supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

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.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, New York, NY, 2004.

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. *arXiv preprint*, 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.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint*, 1901.09036v4, 2019.

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

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Fair off-policy learning from observational data. In *International Conference on Machine Learning (ICML)*, 2024b.

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*, 2408.14766, 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*, 2308.00957, 2024.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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*, 3:160035, 2016.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023a.

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

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

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

Yuchen Ma, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Counterfactual fairness for predictions using generative adversarial networks. *arXiv preprint*, 2310.17687v1, 2023.

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

James Martens. Deep learning via hessian-free optimization. In *International Conference on Machine Learning (ICML)*, 2010.

Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: A novel orthogonal learner. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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 (CLeaR)*, 2022.

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

Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe de Nicolao, and Lennart Ljung. *Regularized system identification: Learning dynamic models from data*. Springer eBook Collection. Springer, Cham, Switzerland, 2022.

Drago Plecko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learnin*, 17(3):304–589, 2024.

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*, 100(469):322–331, 2005.

Maresa Schröder, Dennis Frauen, and Stefan Feuerriegel. Causal fairness under unobserved confounding: a neural sensitivity framework. In *International Conference on Learning Representations (ICLR)*, 2024a.

Maresa Schröder, Dennis Frauen, Jonas Schweisthal, Konstantin Heß, Valentyn Melnychuk, and Stefan Feuerriegel. Conformal prediction for causal effects of continuous treatments. *arXiv preprint*, 2024b.

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.

Cuong Tran, My H. Dinh, and Ferdinando Fioretto. Differentially empirical risk minimization under the fairness lens. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

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

Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Differentially private stochastic gradient descent with low-noise. *Neurocomputing*, 585:127557, 2024.

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(1), 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*, 2401.14549, 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.

A ADDITIONAL BACKGROUND MATERIAL

A.1 DIFFERENTIAL PRIVACY

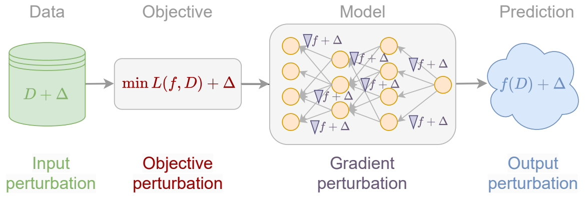


Figure 8: Privacy mechanisms in the machine learning workflow.

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 for the reasons explained below.

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, and (ii) it retains the power of existing CATE estimation methods for addressing the fundamental problem of causal inference. The other DP strategies discussed above fail to fulfill the requirements.

In contrast, 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 designed for 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 the choice of ML method for the first- and second-stage regressions, meaning that arbitrary choices are *not* supported. Furthermore, it has been shown that privatizing different parts of causal estimators separately 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 FUTURE RESEARCH DIRECTIONS FOR PRIVACY IN CAUSAL ML

Applying causal machine learning methods to real-world problems requires the methods to adhere to guidelines on ML safety and ethical ML. The privacy of predictions is only one aspect of the former. Of note, privacy has a direct effect on other ethical and safety-related aspects, such as the uncertainty or the fairness in the predictions. Hence, there are many impactful directions for future research.

Uncertainty in causal ML, including causality-specific types of uncertainty such as unobserved confounding, have been studied in the literature (e.g. Jesson et al., 2020; Melnychuk et al., 2024; Schröder et al., 2024a). However, so far, there is no method that quantifies the uncertainty added through the privatization of the causal model or the predictions. Developing such methods is an important step for future research.

Ethical causal ML has been mostly studied through the lens of causal fairness (e.g., Frauen et al., 2024b; Ma et al., 2023; Plecko & Bareinboim, 2024; Schröder et al., 2024b). However, the trade-off between fair and private causal ML is unexplored. Research outside the field of causal ML suggests opposing effects of privatization and fairness (e.g., Tran et al., 2021). Therefore, investigating this relationship is an interesting avenue for future research and may help in providing ethical, causal treatment effect estimation.

A.4 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 $Y(a)$, since 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 that are common in the literature (e.g., Curth & van der Schaar, 2021; Feuerriegel et al., 2024):

1. Consistency: The potential outcome $Y(a)$ equals the observed factual outcome Y when the individual was assigned treatment $A = a$.
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., there exists $\kappa > 0$ such that $\kappa < \pi(x) < 1 - \kappa$ for all $x \in \mathcal{X}$.
3. Unconfoundedness: Conditioned on the observed covariates, the treatment assignment is independent of the potential outcomes, i.e., $Y(0), Y(1) \perp\!\!\!\perp A \mid X$. 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

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mu(x, 1) - \mu(x, 0), \tag{22}$$

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, meta-learners are flexible and commonly employed in practice. CATE meta-learners can be classified into four different categories, depending on the ways they leverage the data: (1) one-step plug-in learners, (2) two-stage regression-adjusted learners, (3) two-stage propensity-weighted learners, and (4) two-stage Neyman-orthogonal learners (Curth & van der Schaar, 2021). Meta-learners for specialized tasks have also been proposed recently, such as partial identification or treatment effects over time (e.g., Frauen et al., 2024a; 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, i.e., $\hat{\mu}(x, 1)$ and $\hat{\mu}(x, 0)$. Then, the CATE is estimated directly as $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$.
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 by $\hat{\mu}$ in the first stage. A misspecification of $\hat{\mu}$ results in a biased CATE estimator, where the error of $\hat{\mu}$ propagates with the same order into the final estimator $\hat{\tau}$.
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 by $\hat{\pi}$. A misspecification of $\hat{\pi}$ results in a biased CATE estimator. Here, again, the error of $\hat{\pi}$ propagates with the same order into the final estimator $\hat{\tau}$.
4. *Two-stage Neyman-orthogonal learners:* Different from the above two-stage learners, these learners have a *lower error* if either the propensity function π or the outcome regressions μ are correctly specified. Specifically, the final estimation is first-order insensitive to small errors in the nuisance functions (known as quasi-oracle efficiency). This property is achieved through Neyman-orthogonal losses such as Eq. (3).

Hence, we focus on the two-stage Neyman-orthogonal learners 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 Neyman-orthogonal CATE meta-learner. Here, we now provide an extension to the DR-learner.

For the DR-learner, the weight function simplifies to $\lambda^{\mathrm{DR}}(\pi(x)) = 1$. Then, the d-dimensional (ε, δ)-differentially private CATE estimated through the DR-learner is given by

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) = \hat{g}_{D}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) + \gamma^{\mathrm{DR}}\big(T, D\big) \cdot c(\varepsilon, \delta, n) \cdot \mathbf{U}, \tag{23}$$

where $\gamma^{\mathrm{DR}}\big(T, D\big)$ is a sample version of the gross-error sensitivity of $\gamma^{\mathrm{DR}}\big(T, P\big)$, $c(\varepsilon, \delta, n) := 5\sqrt{2\ln(n)\ln\left(2/\delta\right)}\big/(\varepsilon n)$, and $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Specifically, $\gamma^{\mathrm{DR}}\big(T, P\big)$ is given by

$$\gamma^{\mathrm{DR}}\big(T, P\big) = \sup_{z \in \mathcal{Z}} \left\| h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}}) \left(\underbrace{\frac{(a - \hat{\pi}_{\mathrm{DP}}(x))(y - \hat{\mu}_{\mathrm{DP}}(x, a))}{\hat{\pi}_{\mathrm{DP}}(x)(1 - \hat{\pi}_{\mathrm{DP}}(x))} + \hat{\mu}_{\mathrm{DP}}(x, 1) - \hat{\mu}_{\mathrm{DP}}(x, 0)}_{\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))} - g^*(x; \hat{\eta}_{\mathrm{DP}}) \right) \right\|_2 . \tag{24}$$

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 (due to the 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.

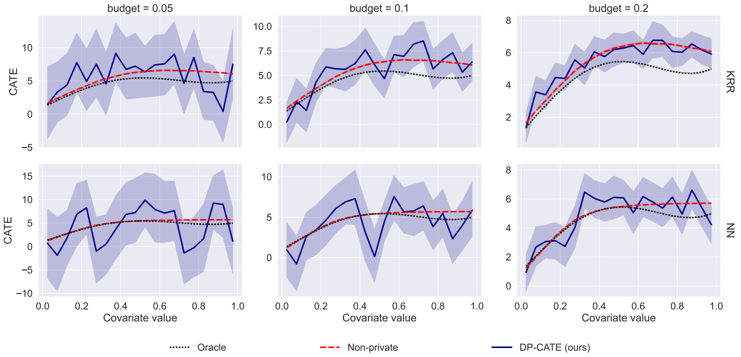


Figure 9: Evaluation of DP-CATE for finite queries the on DR-learner for different base-learner specifications on dataset 1.

C SIMILARITIES AND DIFFERENCES OF THE FINITE AND 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, which we call the "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, the functional approach is relevant when companies want to release a decision support system to guide treatment decisions of individual patients. Since such treatment decisions are made based on the entire CATE model, the functional approach is preferred. In contrast, the first approach (which we call "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 $U(\cdot)$ differs. We highlight the differences in the type of queries in the following:

- **Simultaneous finitely many queries:** When querying the function only once with a finite number of queries, sampling from a Gaussian process implies sampling from the prior distribution of the Gaussian 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 the finite-query approach are independent. Therefore, the functional approach might result in 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 $\hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{DP}})$ in Theorem 3 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 3 in an iterative way.

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

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

Iterative approach: If no query has been made to the private function yet, we can employ Algorithm 2 to provide a private CATE estimate $\hat{g}_{\mathrm{DP}}(x_1; \hat{\eta}_{\mathrm{DP}})$, 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 by 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 $x_1, \ldots, x_{i-1}$ and the privatized outputs $G_i = (\hat{g}_{\mathrm{DP}}(x_1; \hat{\eta}_{\mathrm{DP}}), \ldots, \hat{g}_{\mathrm{DP}}(x_{i-1}; \hat{\eta}_{\mathrm{DP}}))^T$.

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

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

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

D THEORETICAL EVALUATION OF PROPOSED ALGORITHMS

D.1 EXCESS RISK

Let $\hat{g}_{\mathrm{DP}}(\cdot;\hat{\eta}_{\mathrm{DP}})$ and $\hat{g}_{\mathrm{D}}(\cdot;\hat{\eta}_{\mathrm{DP}})$ denote the private and non-private CATE estimates, respectively. Also, let $g^*(\cdot;\hat{\eta}_{\mathrm{DP}})$ denote the second-stage population minimizer. We follow the existing literature (e.g., Bassily et al., 2014) and measure success of our method through the worst-case expected *excess empirical risk* with respect to dataset D, defined as

$$\xi_D(\hat{g}_{\mathrm{DP}},\hat{\eta}_{\mathrm{DP}}) := \mathbb{E}_A\Big[\big\|\hat{g}_{\mathrm{DP}}(\cdot;\hat{\eta}_{\mathrm{DP}}) - g^*(\cdot;\hat{\eta}_{\mathrm{DP}})\big\|_{L_2}^2 - \big\|\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) - g^*(\cdot;\hat{\eta}_{\mathrm{DP}})\big\|_{L_2}^2\Big], \tag{25}$$

where the expectation is taken over the randomness in the second-stage privatization algorithm.

First, observe that, for a suitable calibration factor $r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}})$ and noise variable U specified in Theorems 1 and 3, we can bound the general formula of the excess risk by

$$\xi_D(\hat{g}_{\mathrm{DP}},\hat{\eta}_{\mathrm{DP}}) = \mathbb{E}_U\Big[\big\|\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) + r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}})\cdot U - g^*(\cdot;\hat{\eta}_{\mathrm{DP}})\big\|_{L_2}^2 \tag{26}$$

$$- \big\|\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}}) - g^*(\cdot;\hat{\eta}_{\mathrm{DP}})\big\|_{L_2}^2\Big] \tag{27}$$

$$\leq r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}})^2 \cdot \mathbb{E}_U[\|U\|_{L_2}^2] \tag{28}$$

Finitely many queries: From Theorem 1, we have $U = \mathbf{U} \sim \mathcal{N}(0,\boldsymbol{I}_d)$ and, by setting $d = 1$, we yield

$$r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}}) = \gamma(T,D)\cdot\frac{5\sqrt{2\ln(n)\ln(2/\delta)}}{\varepsilon\, n}, \tag{29}$$

where n is the sample size of the training data D. As $\mathbb{E}_U[\|U\|_{L_2}^2] = \mathbb{E}_U[U^2] = 1$, the excess risk can be bounded by

$$\xi_D(\hat{g}_{\mathrm{DP}},\hat{\eta}_{\mathrm{DP}}) \leq \gamma(T,D)\cdot\frac{50d\ln(n)\ln(2/\delta)}{(\varepsilon\, n)^2} = O(n^{-2}\ln(n)) = o(n^{-1/2}). \tag{30}$$

Functional queries: From Theorem 3, we have

$$r(\varepsilon,\delta,\hat{g}_D,\hat{\eta}_{\mathrm{DP}}) = \sup_{(a,x)\in\{0,1\}\times\mathcal{X}}\big[\rho(a,\hat{\pi}_{\mathrm{DP}}(x))\big]\,\frac{4L\sqrt{2\ln(2/\delta)}}{\big(\sqrt{2\pi}h\big)^q\lambda n\varepsilon}, \tag{31}$$

where h denotes the kernel bandwidth and q the dimension of the covariates. Instead of perturbing a random variable U as above, Theorem 3 perturbs the functional output by a sample path $U = U(\cdot) \in \mathcal{H}$ of a zero-centered Gaussian process with covariance function $K(x,x') = (\sqrt{2\pi}h)^{-q}\exp(-\|x-x'\|_2^2/(2h^2))$. First, we note that

$$\mathbb{E}_U[\mathbb{E}_X(U(X)^2)] = \mathbb{E}_X[\mathbb{E}_U(U(X)^2)] = \mathbb{E}_X[K(X,X)] = (\sqrt{2\pi}h)^{-q}. \tag{32}$$

Therefore, we get

$$\xi_D(\hat{g}_{\mathrm{DP}},\hat{\eta}_{\mathrm{DP}}) \leq \Big(\sup_{(a,x)\in\{0,1\}\times\mathcal{X}}\big[\rho(a,\hat{\pi}_{\mathrm{DP}}(x))\big]\Big)^2\,\frac{32L^2\ln(2/\delta)}{(\sqrt{2\pi}h)^{3q}(\lambda n\varepsilon)^2} = O(n^{-2}) = o(n^{-1/2}). \tag{33}$$

E SUPPORTING LEMMAS AND DEFINITIONS

Our proofs are based on important properties of differentially private algorithms, which we introduce below. For proofs and further details, see (Avella-Medina, 2021; Dwork & Roth, 2014; Nissim et al., 2007).

Lemma 2 (Post-processing theorem). *Let the mechanism $\mathcal{M}$ satisfy (ε, δ)-DP. For any function f, it then holds that $f(\mathcal{M})$ satisfies (ε, δ)-DP as well.*

Lemma 3 (Sequential composition property). *Let the set of privacy mechanisms $\mathcal{M}_j$, $j = 1, \ldots, k$ satisfy $(\varepsilon_j, \delta_j)$-DP. When applying the mechanisms $\mathcal{M}_j$ on the same dataset, the resulting overall mechanism (i.e., the concatenation of all the $\mathcal{M}_j$) guarantees $(\sum_{j=1}^{k} \varepsilon_j, \sum_{j=1}^{k} \delta_j)$-DP.*

Definition 5 (ξ-smooth and local sensitivities). *Let $\hat{f}_D$ be a learned function on dataset D with $|D| = n$ and let $\mathbf{x} \in \mathcal{X}^d$. Then, local sensitivity of $\hat{f}$ is defined as*

$$LS(\hat{f}, D) := \sup_{D' \sim D,\, \mathbf{x} \in \mathcal{X}^d} \|\hat{f}_D(\mathbf{x}) - \hat{f}_{D'}(\mathbf{x})\|_2. \tag{34}$$

where the supremum is taken wrt. D' (unlike Δ_2, where the supremum is taken wrt. to both D and D'). Furthermore, a ξ-smooth sensitivity of $\hat{f}$ is defined as

$$SS_\xi(\hat{f}, D) := \sup_{D' \in \mathcal{Z}^n} \Big[\exp(-\xi d_{\mathrm{H}}(D, D^{'}))\, LS(\hat{f}, D^{'}) \Big], \tag{35}$$

where $\mathcal{Z}^n$ denotes the data domain.

Lemma 4 (Differential privacy via ξ-smooth sensitivity (Avella-Medina, 2021; Nissim et al., 2007)). *Let $\hat{f}_D$ be a learned function on dataset D with $|D| = n$ and let $\mathbf{x} \in \mathcal{X}^d$. Then, $\hat{f}_{\mathrm{DP}}(\mathbf{x})$ fulfills (ε, δ)-differential privacy:*

$$\hat{f}_{\mathrm{DP}}(\mathbf{x}) = \hat{f}_D(\mathbf{x}) + \frac{5\sqrt{2\log(2/\delta)}}{\varepsilon} SS_\xi(\hat{f}, D) \cdot \mathbf{U}, \tag{36}$$

where $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\xi = \frac{\varepsilon}{4(d+2\log(2/\delta))}$.

Finally, we show that in our case, the ξ-smooth sensitivity can be upper-bounded by the (appropriately scaled) gross-error sensitivity.

Lemma 5. (Avella-Medina, 2021) *Let $\hat{f}_D$ be a learned function on dataset D with $|D| = n$ and let $\mathbf{x} \in \mathcal{X}^d$. Furthermore, let $\gamma(T, D)$ denote the sample gross error sensitivity of the parameter of the interest $T(D) = \hat{f}_D(\mathbf{x})$. Under minimal regularity and boundedness conditions, for the ξ-smooth sensitivity $SS_\xi(\hat{f}, D)$, with $\xi = \frac{\varepsilon}{4(d+2\log(2/\delta))}$ it holds that*

$$SS_\xi(\hat{f}, D) \leq \frac{\sqrt{\log(n)}}{n} \gamma(T, D). \tag{37}$$

F PROOFS

F.1 PROOFS OF THEOREM 1, COROLLARIES 1, 2, THEOREM 2

Theorem 1 (DP-CATE for finite queries). *Let $z := (a, x, y)$ define a data sample following the joint distribution $\mathcal{Z}$ and $\hat{\eta}_{\mathrm{DP}} = (\hat{\mu}_{\mathrm{DP}}, \hat{\pi}_{\mathrm{DP}})$ the nuisance functions estimated on dataset $\tilde{D}$ in a $(\varepsilon/2, \delta/2)$-differentially private manner of choice.*[8] *Furthermore, let D be the training dataset for the second-stage model with $|D| = n$. For $z = (a, x, y) \in \mathcal{Z}$, we define*

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) := \hat{g}_D(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) + \underbrace{\gamma\big(T, D\big) \cdot c(\varepsilon, \delta, n)}_{r(\varepsilon, \delta, \hat{g}_D, \hat{\eta}_{\mathrm{DP}})} \cdot \mathbf{U}, \qquad T(P) = g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}), \tag{10}$$

$$\gamma\big(T, P\big) = \sup_{z \in \mathcal{Z}} \left\| h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}})\, \rho(a, \hat{\pi}_{\mathrm{DP}}(x)) \big(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(x))) - g^*(x; \hat{\eta}_{\mathrm{DP}})\big) \right\|_2, \tag{11}$$

where $\gamma\big(T, D\big)$ is a sample gross-error sensitivity with $T(D)$ instead of $T(P)$ in Eq. 9, $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $c(\varepsilon, \delta, n) := 5\sqrt{2\ln(n)\ln(2/\delta)}\big/(\varepsilon n)$, and where $h(g^*, \mathbf{x}, z, \hat{\eta}_{\mathrm{DP}}) \in \mathbb{R}^d$ and $g^*(\cdot; \hat{\eta}_{\mathrm{DP}})$ depend on the machine learning model employed for the second stage regression. Then, $\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ is (ε, δ)-differentially private.

Proof. First, observe that by Lemma 3, the first stage nuisance estimation $\hat{\eta}_{\mathrm{DP}}$ overall fulfills (ε, δ)-DP. By Lemma 2, the privacy of the nuisances is not affected by the second-stage regression. Note that we estimate both stages on disjoint subsets of the data, i.e., $\bar{D} = \tilde{D} \cup D$.

If we had access to the ξ-smooth sensitivity $SS_\xi(g, D)$ of the meta-learner with $\xi = \frac{\varepsilon}{4(d + 2\log(2/\delta))}$, the estimator $\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ with

$$\hat{g}_{\mathrm{DP}}(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) = \hat{g}_D(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}) + \frac{5\sqrt{2\log(2/\delta)}}{\varepsilon} SS_\xi(\hat{g}, D) \cdot \mathbf{U}, \tag{38}$$

where $\mathbf{U} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $\mathbf{x} \in \mathbb{R}^d$, would fulfill (ε, δ)-differential privacy by Lemma 4 and the parallel composition property of DP.

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 $r(\varepsilon, \delta, \hat{g}_D, \hat{\eta}_{\mathrm{DP}})$. By Lemma 5, we find such an upper bound through

$$SS_\xi(\hat{g}, D) \leq \frac{\sqrt{\log(n)}}{n} \gamma(T, D) = \frac{\sqrt{\log(n)}}{n} \sup_{z \in \mathcal{Z}} \|\mathrm{IF}(z, T; D)\|_2, \tag{39}$$

where $\gamma(T, D)$ denotes the sample gross-error sensitivity with the parameter of interest T defined as $T(P) = g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ (population version) or $T(D) = \hat{g}_D(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ (sample version).

Now, the influence function of $T(P) = g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ where $g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) = \arg\min_{g \in \mathcal{G}} R_P(g, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))$ is as follows:

$$\begin{aligned}
\mathrm{IF}(z, T; P) &= \frac{\mathrm{d}}{\mathrm{d}t} \left[\big(\arg\min_{g \in \mathcal{G}} R_{(1-t)P + t\delta_z}(g, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))\big)(\mathbf{x}) \right] \Bigg|_{t=0} &(40)\\
&= \frac{\mathrm{d}}{\mathrm{d}t} \left[\big(\arg\min_{g \in \mathcal{G}} \big\{ (1-t) R_P(g, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) + t\rho(a, \hat{\pi}_{\mathrm{DP}}(x))(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}) - g(x))^2 \big\}\big)(\mathbf{x}) \right] \Bigg|_{t=0}. &(41)
\end{aligned}$$

Here, the $\arg\min$ is achieved at $g_t^*(\cdot; \hat{\eta}_{\mathrm{DP}})$, and, for smooth models $\mathcal{G}$, the following holds:

$$\frac{\mathrm{d}}{\mathrm{d}g} \big\{ (1-t) R_P(g, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) + t\rho(a, \hat{\pi}_{\mathrm{DP}}(x))(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}) - g(x))^2 \big\} \Big|_{g = g_t^*(\cdot; \hat{\eta}_{\mathrm{DP}})} = 0 \tag{42}$$

$$\Longleftrightarrow \underbrace{(1-t) \frac{\mathrm{d}}{\mathrm{d}g} \big\{ R_P(g, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) \big\} - 2t\rho(a, \hat{\pi}_{\mathrm{DP}}(x))(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) - g(x)) \frac{\mathrm{d}g(x)}{\mathrm{d}g}}_{F(g,t)} \Big|_{g = g_t^*(\cdot; \hat{\eta}_{\mathrm{DP}})} = 0. \tag{43}$$

---

[8] As we need to query the nuisance functions multiple times, the privatization method needs to be suitable for functions, such as gradient perturbation through DP-SGD (Abadi et al., 2016).

Therefore, by the implicit function theorem:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big[g_t^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})\Big]\Big|_{t=0} = \left(\frac{\mathrm{d}F(g,t)}{\mathrm{d}g}\Big|_{t=0;\, g=g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})}\right)^{-1}\frac{\mathrm{d}F(g,t)}{\mathrm{d}t}\Big|_{t=0;\, g=g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})} \tag{44}$$

$$= \left(\frac{\mathrm{d}^2}{\mathrm{d}^2 g}\big\{R_P(g,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}}))\big\}\Big|_{g=g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})}\right)^{-1}\Bigg(\underbrace{-\frac{\mathrm{d}}{\mathrm{d}g}\big\{R_P(g,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}}))\big\}\Big|_{g=g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})}}_{=0}$$

$$- 2\rho(a,\hat{\pi}_{\mathrm{DP}}(x))(\phi(z,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}})) - g^*(x;\hat{\eta}_{\mathrm{DP}}))\cdot\frac{\mathrm{d}g(x)}{\mathrm{d}g}\Big|_{g=g^*(x;\hat{\eta}_{\mathrm{DP}})}\Bigg). \tag{45}$$

Now, by regrouping and setting $h(g^*,\mathbf{x},x,\hat{\eta}_{\mathrm{DP}})$ as

$$h(g^*,\mathbf{x},x,\hat{\eta}_{\mathrm{DP}}) = 2\left(\frac{\mathrm{d}^2}{\mathrm{d}^2 g}\big\{R_P(g,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}}))\big\}\Big|_{g=g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})}\right)^{-1}\cdot\frac{\mathrm{d}g(x)}{\mathrm{d}g}\Big|_{g=g^*(x;\hat{\eta}_{\mathrm{DP}})}, \tag{46}$$

we obtain the desired influence function:

$$\mathrm{IF}(z,T;P) = -h(g^*,\mathbf{x},x,\eta)\cdot\rho(a,\hat{\pi}_{\mathrm{DP}}(x))(\phi(z,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}})) - g^*(x;\hat{\eta}_{\mathrm{DP}})). \tag{47}$$

Here, $h(g^*,\mathbf{x},x,\hat{\eta}_{\mathrm{DP}})\in\mathbb{R}^d$ depends on the machine learning model employed for the second stage regression (see Corollaries 1 and 2). Finally, taking the supremum over the data space $\mathcal{Z}$ and evaluating $\mathrm{IF}(z,T,P)$ at the empirical distribution D with trained functions $\hat{\eta}_{\mathrm{DP}}$ and $\hat{g}_D(\cdot;\hat{\eta}_{\mathrm{DP}})$ states the desired result. $\square$

Corollary 1 (Parametric second-stage regression)**.** *If the second-stage regression is a smooth parametric model, namely $\mathcal{G}=\{g(x;\theta):\mathcal{X}\mapsto\mathbb{R},\theta\in\Theta\subseteq\mathbb{R}^p\}$, then in Theorem 1*

$$g^*(\cdot;\hat{\eta}_{\mathrm{DP}}) = g(\cdot;\theta^*) \quad \textit{and} \quad h(g^*,\mathbf{x},x,\hat{\eta}_{\mathrm{DP}}) = 2\,J_\theta\big[g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})\big]\cdot H_\theta^{-1}\cdot\nabla_\theta\big[g^*(x;\hat{\eta}_{\mathrm{DP}})\big], \tag{13}$$

where $\theta^*=\arg\min_{\theta\in\Theta} R_P(g,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}}))$; $J_\theta\in\mathbb{R}^{d\times p}$ is a Jacobian matrix wrt. θ; $H_\theta=\nabla_\theta^2\big[R_P(g^*,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}}))\big]\in\mathbb{R}^{p\times p}$ is a Hessian matrix; and ∇_θ is a gradient.

Proof. First, we note, that θ^* is an example of M-estimator. Therefore, $\mathrm{IF}(z,T;P)$ yields an influence function of the M-estimator. Specifically, it can be shown that

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}^2 g}\big\{R_P(g,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}}))\big\}\Big|_{g=g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})}\right)^{-1} = J_\theta\big[g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})\big]\cdot H_\theta^{-1}\ \text{and} \tag{48}$$

$$\frac{\mathrm{d}g(x)}{\mathrm{d}g}\Big|_{g=g^*(x;\hat{\eta}_{\mathrm{DP}})} = \nabla_\theta\big[g^*(x;\hat{\eta}_{\mathrm{DP}})\big]. \tag{49}$$

Therefore, we yield the desired result

$$h(g^*,\mathbf{x},x,\hat{\eta}_{\mathrm{DP}}) = 2\,J_\theta\big[g^*(\mathbf{x};\hat{\eta}_{\mathrm{DP}})\big]\cdot H_\theta^{-1}\cdot\nabla_\theta\big[g^*(x;\hat{\eta}_{\mathrm{DP}})\big]. \tag{50}$$

$\square$

Remark 1. *Importantly, for the sample gross error sensitivity, $\gamma(T,D)=\sup_{z\in\mathcal{Z}}\|\mathrm{IF}(z,T;D)\|_2$, to be finite, we need additional regularity conditions. Specifically, we denote the score function of Eq. (3) by*

$$\psi(z,g,\hat{\eta}_{\mathrm{DP}}) := \rho(a,\pi(x))\,(\phi(z,\hat{\eta}_{\mathrm{DP}},\lambda(\hat{\pi}_{\mathrm{DP}})) - g(x)). \tag{51}$$

First, note that $\psi(z,g,\hat{\eta}_{\mathrm{DP}})$ is differentiable w.r.t. $g(x)$ for all $z\in\mathcal{Z}$. Since we assume that our data stems from a bounded domain, there exist constants K and L such that $\psi(z,g,\hat{\eta}_{\mathrm{DP}})$ and the derivative $\psi^{'}(z,g,\hat{\eta}_{\mathrm{DP}})$ are uniformly bounded in $\mathcal{Z}$ by K and L. Additionally, we assume that $\|\nabla_\theta\hat{g}_D(x;\hat{\eta}_{\mathrm{DP}})\|_2\leq\tilde{K}$. The constant $\tilde{K}$ can be regulated through, e.g., gradient clipping. Furthermore, the Hessian H_θ of the loss w.r.t. θ is positive semi-definite by assumption.

If H_θ is further restricted to be positive definite, we can employ Lemmas 1 and 2 from Avella-Medina (2021) to show the desired upper bound of the smooth sensitivity for a sufficiently large sample size n.[9] *Specifically, the minimum sample size depends inversely on the privacy budget ε and δ and grows in the bounds $L,K,\tilde{K}$ and the ratio of the maximum and minimum eigenvalue of H_θ.*

Note: If the Hessian does not fulfill the above criteria and has negative eigenvalues, one can employ the "damping" trick (Martens, 2010) by approximating H_θ by $\tilde{H}_\theta := H_\theta+\alpha\mathbf{I}$ for $\alpha>0$. The parameter α defines the conservativeness of the approximation.

---

[9] For details, we refer to Avella-Medina (2021).

Corollary 2 (Non-parametric second-stage regression)**.** *If the second-stage regression is a kernel ridge regression with $\Lambda(g) = \lambda\|g\|^2_{\mathcal{H}}$, where $\mathcal{G} = \mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) induced by a kernel $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$, then in Theorem 1*

$$g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) = \big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} S(\cdot) \quad \textit{and} \quad h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}}) = \big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} K(\cdot, x)(\mathbf{x}), \tag{14}$$

where $\mathbb{L}_\rho : \mathcal{H} \mapsto \mathcal{H}$ is a weighted covariance operator $\mathbb{L}_\rho f(\cdot) = \mathbb{E}\big[\rho(A, \hat{\pi}_{\mathrm{DP}}(X))\, K(\cdot, X) f(X)\big]$; $\lambda\mathbb{I}f(\cdot) = \lambda f(\cdot)$ is a scaling operator; $S \in \mathcal{H}$ is a cross-covariance functional $S(\cdot) = \mathbb{E}\big[\rho(A, \hat{\pi}_{\mathrm{DP}}(X))\, K(\cdot, X)\, Y\big]$.

Proof. By the representer theorem (population version), the optimal solution of the weighted kernel ridge regression is given by $g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) = \big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} S(\cdot)$, where $\mathbb{L}_\rho f(\cdot) = \mathbb{E}\big[\rho(A, \hat{\pi}_{\mathrm{DP}}(X))\, K(\cdot, X) f(X)\big]$ and $S(\cdot) = \mathbb{E}\big[\rho(A, \hat{\pi}_{\mathrm{DP}}(X))\, K(\cdot, X)\, \phi(Z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))\big]$. For a description, see (e.g., Berlinet & Thomas-Agnan, 2004; Pillonetto et al., 2022).

We now turn to deriving the influence function of $g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})$ at $z = (a_0, x_0, y_0)$; $\mathrm{IF}(z, g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}); P)$; to then receive the explicit form of $h(g^*, \mathbf{x}, x, \hat{\eta}_{\mathrm{DP}})$.

Consider the point-mass perturbation of the distribution P in the direction of z denoted as $P_\varepsilon = (1-\varepsilon)P + \varepsilon\delta_z$. Then, the solution to the kernel ridge regression at P_ε is given by

$$g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})_{P_\varepsilon} = \big(\mathbb{L}_{P_\varepsilon} + \lambda\mathbb{I}\big)^{-1} S_{P_\varepsilon}(\mathbf{x}), \tag{52}$$

where $\mathbb{L}_{P_\varepsilon} = (1-\varepsilon)L_\rho + \varepsilon L_{\delta_z}$ and $S_{P_\varepsilon} = (1-\varepsilon)S + \varepsilon S_{\delta_z}$ with $L_{\delta_z} f(\mathbf{x}) = \rho(a_0, \hat{\pi}_{\mathrm{DP}}(x_0))\, K(\mathbf{x}, x_0) f(z)$ and $S_{\delta_z}(\mathbf{x}) = \rho(a_0, \hat{\pi}_{\mathrm{DP}}(x_0))\, K(\mathbf{x}, x_0)\, \phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}))$.

Then we get

$$\begin{aligned}
\mathrm{IF}(z, g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}); P) &= \frac{\partial}{\partial\varepsilon} g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}})_{P_\varepsilon}|_{\varepsilon=0} && (53)\\
&= -\big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} L_{\delta_z}(\mathbf{x}) \big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} S(\mathbf{x}) + \big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} S_{\delta_z}(\mathbf{x}) && (54)\\
&= -\rho(a_0, \hat{\pi}_{\mathrm{DP}}(x_0)) \cdot \Big(\big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} K(\cdot, x_0)\Big)(\mathbf{x}) \cdot g^*(x_0; \hat{\eta}_{\mathrm{DP}}) && (55)\\
&\quad + \rho(a_0, \hat{\pi}_{\mathrm{DP}}(x_0)) \cdot \Big(\big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} K(\cdot, x_0)\Big)(\mathbf{x}) \cdot \phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) && (56)
\end{aligned}$$

With $h(g^*, \mathbf{x}, x_0, \hat{\eta}_{\mathrm{DP}}) = \Big(\big(\mathbb{L}_\rho + \lambda\mathbb{I}\big)^{-1} K(\cdot, x_0)\Big)(\mathbf{x})$ we get

$$\mathrm{IF}(z, g^*(\mathbf{x}; \hat{\eta}_{\mathrm{DP}}); P) = h(g^*, \mathbf{x}, x_0, \hat{\eta}_{\mathrm{DP}}) \cdot \rho(a_0, \hat{\pi}_{\mathrm{DP}})\big(\phi(z, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})) - g^*(x_0; \hat{\eta}_{\mathrm{DP}})\big)\big), \tag{57}$$

which shows the desired result.

□

Theorem 2 (Quasi-oracle efficiency of DP-CATE)**.** *The privatization of the nuisance functions does not affect the quasi-oracle efficiency of the final CATE estimator. Specifically, for machine learning models f with sufficiently fast convergence rate in $o_P(n^{-1/4})$, privatization through output or gradient perturbation still retains the stated rate, rendering the final estimate efficient:*

$$\begin{aligned}
\|g^*(\cdot; \eta) - \hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{DP}})\|^2_{L_2} \lesssim\; & R_P\big(g^*(\cdot; \hat{\eta}_{\mathrm{DP}}), \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})\big) - R_P\big(g^*(\cdot; \eta), \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}})\big) && (15)\\
& + R_2(\hat{\eta}_{\mathrm{DP}}, \eta) + O_P(n^{-2}) + \|g^*(\cdot; \hat{\eta}_{\mathrm{DP}}) - \hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{D}})\|^2_{L_2}, && (16)
\end{aligned}$$

where

$$\|\hat{\eta}_{\mathrm{DP}} - \eta\|^2_{L_2} \leq \|\hat{\eta}_{\mathrm{DP}} - \hat{\eta}_D\|^2_{L_2} + \|\hat{\eta}_D - \eta\|^2_{L_2} \tag{17}$$

Proof. To show the quasi-oracle efficiency of DP-CATE, we need to show that $\hat{\mu}_{\mathrm{DP}}$ and $\hat{\pi}_{\mathrm{DP}}$ are each consistent with convergence rate in $o_P(n^{-\frac{1}{4}})$, where we a, where we assume that the non-private estimators $\mu_{D'}$ and $\pi_{D'}$ achieve this rate.

We focus on the basic gradient perturbation method, namely DP-SGD (Bassily et al., 2014; Abadi et al., 2016). Let m denote the number of model parameters. From the literature, (e.g., Chen et al., 2020), we know that the convergence rate for (ε, δ)-DP-SGD for convex and Lipschitz losses holds

$$r^{DP-SGD} \in O(\frac{\sqrt{m}}{n\varepsilon}), \tag{58}$$

where n denotes the sample size.

Recall that in our setup we only allocate the budget $(\frac{\varepsilon}{2}, \frac{\delta}{2})$ to the privatization of each $\hat{\pi}_{D'}$ and $\hat{\mu}_{D'}$. We yield the following result by straightforward calculation: For π and μ with $m \leq \sqrt{n}^3 \varepsilon^2$, $r^{DP-SGD} \in p(n^{-\frac{1}{4}})$.

For other losses (e.g., non-Lipschitz, non-convex) as well as advanced versions of DP-SGD, an upper bound of $O(\frac{1}{\sqrt{n}} + \frac{\sqrt{m \log(1/\delta)}}{n\varepsilon})$ has been derived in the literature. For an overview, see (Wang et al., 2024). In this case, we achieve

$$r^{DP-SGD} \in o_P(n^{-\frac{1}{4}}) \iff m \log(1/\delta) \leq n\varepsilon^2 (n^{1/4} - 1)^2. \tag{59}$$

Overall, we thus face a trade-off between the model size in terms of the number of parameters, privacy budget, and sample size. For appropriately chosen factors, the first-stage nuisances achieve rate $o_P(n^{-\frac{1}{4}})$, rendering DP-CATE oracle-efficient.

$\square$

F.2 PROOFS OF LEMMA 1, THEOREM 3

Lemma 1. *Let $\mathcal{H}$ denote the RKHS induced by the Gaussian kernel $K(x, x') = (\sqrt{2\pi}h)^{-q} \exp(-\|x - x'\|_2^2/(2h^2))$ for $x, x' \in \mathcal{X} \subseteq \mathbb{R}^q$, and let $\hat{f}_D$ be the optimal solution to the RKHS regression*

$$\hat{f}_D(\cdot) = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} w(X_i) \cdot \ell(f(X_i), Y_i) + \lambda \|f\|_{\mathcal{H}}^2, \tag{18}$$

where $w(\cdot) > 0$ is a weight function, D is a dataset with $|D| = n$, and $\ell(\hat{y}, y)$ is a convex and Lipschitz loss function in $\hat{y}$ with Lipschitz constant L. Then, we have

$$||f_D - f_{D'}||_{\mathcal{H}} \leq \sup_{x \in \mathcal{X}} [w(x)] \, \frac{L}{\lambda\, n} \left(\sqrt{(2\pi)}h\right)^{-q}. \tag{19}$$

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

$$||\hat{f}_D - \hat{f}_{D'}||_{\mathcal{H}} \leq \frac{L}{\lambda n} \sqrt{\sup_x K(x, x)}. \tag{60}$$

With our observation above, the result follows. $\square$

Theorem 3 (DP-CATE for functional queries)**.** *Let $\hat{\mu}_{\mathrm{DP}}$ and $\hat{\pi}_{\mathrm{DP}}$ denote the $(\varepsilon/2, \delta/2)$-differentially private nuisance estimators trained in stage 1 on $\tilde{D}$. Let $z = (a, x, y)$ be a data sample from dataset D with $|D| = n$ and $x \in \mathcal{X} \subseteq \mathbb{R}^q$. Let $\mathcal{H}$ denote the RKHS induced by the kernel $K(x, x') = (\sqrt{2\pi}h)^{-q} \exp(-\|x - x'\|_2^2/2h^2)$, and let $\ell(\cdot)$ be a convex and Lipschitz loss function with Lipschitz constant L. We define $\hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}})$ as the second stage regression solving Eq. (3) via*

$$\hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}}) = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \rho(A_i, \hat{\pi}_{\mathrm{DP}}(X_i))\, \ell\big(g(X_i), \phi(Z_i, \hat{\eta}_{\mathrm{DP}}, \lambda(\hat{\pi}_{\mathrm{DP}}(X_i)))\big) + \lambda \|g\|_{\mathcal{H}}^2. \tag{20}$$

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

$$\hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{DP}}) := \hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}}) + \underbrace{\sup_{(a,x) \in \{0,1\} \times \mathcal{X}} \big[\rho(a, \hat{\pi}_{\mathrm{DP}}(x))\big] \frac{4L\sqrt{2 \ln{(2/\delta)}}}{\big(\sqrt{2\pi}h\big)^q \lambda n \varepsilon}}_{r(\varepsilon, \delta, \hat{g}_D, \hat{\eta}_{\mathrm{DP}})} \cdot U(\cdot). \tag{21}$$

Proof. Let U be the sample path of a zero-centered Gaussian process with covariance function $K(x, y)$. For our proof, we make use of Corollary 9 from Hall et al. (2013): for $\hat{f} \in \mathcal{H}$, where $\mathcal{H}$ is the RKHS corresponding to the kernel K, the release of

$$\tilde{f}_D = \hat{f}_D + \frac{\Delta c(\delta)}{\varepsilon} \cdot U(\cdot) \tag{61}$$

is (ε, δ)-differentially private for

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

and

$$\Delta \geq \sup_{D \sim D'} \|\hat{f}_D - \hat{f}_{D'}\|_{\mathcal{H}}. \tag{63}$$

Therefore, for $\Delta^* := \sup_{D \sim D'} \|\hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}}) - \hat{g}_{D'}(\cdot; \hat{\eta}_{\mathrm{DP}})\|_{\mathcal{H}}$,

$$\hat{g}_{\mathrm{DP}}(\cdot; \hat{\eta}_{\mathrm{DP}}) = \hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}}) + \frac{\Delta^* \sqrt{2 \log(2/\delta)}}{\varepsilon} \cdot U(\cdot) \tag{64}$$

is (ε, δ)-differentially private. Note that here, again, we implicitly make use of Lemma 2 and Lemma 3 in the same manner as in the proof of Theorem 1.

Finally, from Lemma 1, we know that

$$\sup_{D \sim D'} \|\hat{g}_D(\cdot; \hat{\eta}_{\mathrm{DP}}) - \hat{g}_{D'}(\cdot; \hat{\eta}_{\mathrm{DP}})\|_{\mathcal{H}} \leq \frac{L}{\lambda n} (\sqrt{(2\pi)} h)^{-q}. \tag{65}$$

Thus, the desired result follows. □

G EXPERIMENTS

G.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.

G.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.

G.3 IMPLEMENTATION DETAILS

Our experiments are implemented in Python. We provide our code in our GitHub repository: `https://github.com/m-schroder/DP-CATE`.

Our DP-CATE is model-agnostic and highly flexible. Therefore, we implement multiple versions of the R- and the DR-leaner with varying base learner instantiations. 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 with the finite-query DP-CATE, we implement the pseudo-outcome regression in the second stage as (a) a kernel ridge regression model with a Gaussian kernel and default parameter specifications (KR) and (b) a neural network (NN) with two hidden layers of width 32 with tanh activation function trained in the same manner as the nuisance models. In the experiments for 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 l_1 loss or can be upper-bounded. In our settings, we employ the l_2 loss on a bounded domain. Therefore, although the l_2 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.