Model Agnostic Differentially Private Causal Inference

Christian Lebeda[1,⋆] **Mathieu Even**[1,2,⋆], **Aurélien Bellet** [1], **Julie Josse** [1]

⋆ Equal contribution

[1]Inria, Université de Montpellier, INSERM, France

[2]Theremia

Abstract

Estimating causal effects from observational data is essential in fields such as medicine, economics and social sciences, where privacy concerns are paramount. We propose a general, model-agnostic framework for differentially private estimation of average treatment effects (ATE) that avoids strong structural assumptions on the data-generating process or the models used to estimate propensity scores and conditional outcomes. In contrast to prior work, which enforces differential privacy by directly privatizing these nuisance components and results in a privacy cost that scales with model complexity, our approach decouples nuisance estimation from privacy protection. This separation allows the use of flexible, state-of-the-art black-box models, while differential privacy is achieved by perturbing only predictions and aggregation steps within a fold-splitting scheme with ensemble techniques. We instantiate the framework for three classical estimators—the G-formula, inverse propensity weighting (IPW), and augmented IPW (AIPW)—and provide formal utility and privacy guarantees. Empirical results show that our methods maintain competitive performance under realistic privacy budgets. We further extend our framework to support meta-analysis of multiple private ATE estimates. Our results bridge a critical gap between causal inference and privacy-preserving data analysis.

1 Introduction

Quantifying treatment effects at the population level is a critical task with far-reaching implications for economics, policy, and public health. These estimates guide health interventions, shape policy decisions, and inform clinical recommendations. In modern evidence-based medicine, randomized controlled trials (RCTs) are considered the gold standard for estimating treatment effects—such as the average treatment effect (ATE) via the risk difference (RD)—because they effectively isolate causal effects from confounding factors [Imbens and Rubin, 2015]. However, RCTs present several limitations: they are costly and time-consuming, and their strict inclusion/exclusion criteria often result in study samples that are small and differ significantly from the broader population eligible for treatment. In contrast, observational data—gathered without deliberate intervention, as in disease registries, cohorts, biobanks, epidemiological studies, or electronic health records—offer a compelling alternative. These data sources are typically more comprehensive, less expensive to collect, and often provide a more accurate representation of real-world populations. Yet, estimating treatment effects from observational data is challenging due to the presence of *confounders*. Most existing approaches handle this by adjusting for *observed confounders* through regression of *nuisance* parameters [Chernozhukov et al., 2018, Wager, 2024].

In parallel to these statistical challenges, the sensitive nature of individual-level data used in causal inference studies raises significant privacy concerns. Medical, economic, and social science datasets

Preprint. Under review.

often contain personal and highly sensitive information, making it essential to prevent the reidentification of individuals and the reconstruction of private attributes from study outputs. Although the release of a single causal estimate may pose minimal privacy risks in isolation, datasets are frequently reused across various studies [e.g., the UK BioBank; Sudlow et al., 2015], which compounds the risk of disclosure. Indeed, it is well known that answering too many aggregate statistics—even as simple as counts or means—on the same dataset enables an attacker to reconstruct individual records with high accuracy [Dinur and Nissim, 2003]. Empirical attacks have confirmed this risk in domains ranging from genomics [Homer et al., 2008, Shringarpure and Bustamante, 2015] and census microdata [Dick et al., 2023] to location data [Pyrgelis et al., 2018]. These findings underscore the need for causal inference methods that provide formal privacy guarantees. Differential Privacy (DP) [Dwork et al., 2006, Dwork and Roth, 2014] is widely recognized as the gold standard for providing such guarantees: by adding noise proportional to the *sensitivity* of the computation—i.e., the maximum change in output due to a single individual's data—it ensures that each individual has only a limited influence on the output, thereby bounding what an adversary can infer. Crucially, DP also provides a formal mechanism to control the cumulative privacy loss across repeated analyses on the same dataset.

Existing approaches to differentially private causal inference are limited in scope. Some are designed for simplified settings—such as RCTs using private difference-in-means estimators [Chen et al., 2024, Ohnishi and Awan, 2025, Yao et al., 2024]—and therefore fall short of realizing the potential of large-scale observational datasets. Others depend on strong structural assumptions, including parametric forms for nuisance parameters [Lee et al., 2019, Ohnishi and Awan, 2024], which often do not hold in practice and lead to biased estimators, limiting their applicability to real-world data. Crucially, these methods enforce privacy by privatizing the nuisance models themselves, requiring differentially private training algorithms for the chosen model class and incurring privacy costs that scale with model complexity, which are unnecessary when only the final causal estimate is released.

Contributions. Motivated by these limitations, we propose a flexible, non-parametric, and model-agnostic framework for differentially private causal inference, designed to meet the practical needs of the broader causal inference community. We develop differentially private methods for estimating the ATE from observational data, without imposing structural assumptions, which is crucial to avoid bias from model misspecification. In contrast to prior methods, our approach is model-agnostic: any *non-private* estimator can be used for the nuisance parameters (e.g., propensity scores or outcome regressions), while privacy is enforced at the prediction and aggregation stages. The framework proceeds through five intuitive steps: *(i)* split the data into K folds; *(ii)* estimate nuisance parameters on each fold; *(iii)* aggregate the predictions of these K nuisance models to reduce sensitivity—and thus the amount of noise required for privacy; *(iv)* construct a private ATE estimator from these predictions and calibrated Gaussian noise; and *(v)* compute a differentially private variance estimate to produce a confidence interval. We then instantiate this framework to derive differentially private versions of three widely-used estimators: the plug-in *G-formula*, *Inverse Propensity Weighting* (IPW), and *Augmented Inverse Propensity Weighting* (AIPW) estimators. We provide a unified analysis of privacy and utility guarantees, leveraging the recent Gaussian DP framework [Dong et al., 2022] to derive tight privacy bounds, while quantifying the asymptotic variance of the estimators, shedding light on their privacy-utility trade-offs. Additionally, we extend our framework to perform meta analysis, in which several private ATE estimates are aggregated to effectively reduce variance. Finally, we demonstrate the effectiveness of our approach through experiments on synthetic data, highlighting its superiority over prior methods in practically relevant scenarios.

2 Preliminaries

2.1 Causal inference framework

We assume access to a dataset of n independent and identically distributed (i.i.d.) patients drawn from an underlying distribution $\mathcal{P}$. Each patient $i \in [n]$ is described by their *covariates* (feature vector) $X_i \in \mathcal{X}$, *treatment assignment* $A_i \in \{0, 1\}$ (we assume binary treatments), and *observed outcome* $Y_i \in \mathcal{Y} \subset \mathbb{R}$ (treatment response). Let $\mathcal{D} := (X_i, A_i, Y_i)_{i\in[n]} \sim \mathcal{P}$ denote the dataset, which may originate from either a randomized control trial (RCT) or an observational study. We adopt the potential outcome framework, which formalizes the notion of causal effects by positing the existence of two potential outcomes for each individual i, $Y_i(1), Y_i(0) \in \mathcal{Y}$, corresponding to the outcomes under treatment and control, respectively. Throughout this work, we assume the Stable

Unit Treatment Values Assumption (SUTVA), which states that each individual's observed outcome equals their potential outcome under the assigned treatment; that is, $Y_i = Y_i(A_i)$.

Assumption 1 (Stable Unit Treatment Value (SUTVA))**.** *The observed outcome of any sample is independent of treatment assignment of the other samples. Formally,* $Y = AY(1) + (1-A)Y(0)$.

We quantify treatment effects by estimating the average treatment effect with the risk difference.

Definition 1 (Average Treatment Effect (ATE))**.** *The average treatment effect (ATE) with the risk difference (RD) is defined as* $\tau := \mathbb{E}[Y(1) - Y(0)]$. *The expectation is taken over the distribution* $\mathcal{P}$.

The fundamental challenge of causal inference lies in the fact that we can only observe one of the potential outcomes $Y_i(0)$ and $Y_i(1)$ for each data point. Even though the joint distribution of $(Y_i(0), Y_i(1))$ will never be observed, we can still construct estimators of the ATE with the RD, provided that some identifiability assumption holds: there is no hidden confounder and any patient has a probability of being treated that is bounded away from 0 and 1.

Assumption 2 (Unconfoundedness)**.** *The potential outcomes and the treatment assignments are independent when conditioned on the covariates. Formally:* $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$.

Assumption 3 (Overlap)**.** *There exists* $\eta \in (0, 1/2]$ *such that for all* $x \in \mathcal{X}$*, we have* $\pi(x) \in [\eta, 1-\eta]$*, where* $\pi(x) := \mathbb{P}(A = 1|X = x)$ *is the propensity score.*

2.2 Non-private ATE estimators

We now present standard non-private ATE estimators under Assumptions 1 to 3, beginning with the associated nuisance functions. The propensity score $\pi(x) := \mathbb{P}(A = 1|X = x)$ is the probability of receiving treatment as a function of the covariates, while $\mu_a(x) := \mathbb{E}[Y|X = x, A = a]$ is the expected outcome as a function of covariates and treatment. The Plug-in G-Formula, IPW, and AIPW estimators we consider in this work can all be written as an average $\hat{\tau} := \frac{1}{n}\sum_{i=1}^{n} \Gamma_i$ of *scores* Γ_i for each patient $i \in [n]$ using (X_i, A_i, Y_i) and (non-parametric) estimators $\hat{\mu}_0, \hat{\mu}_1, \hat{\pi}$ of μ_0, μ_1, π:[1]

$$\text{G-formula: } \Gamma_i^{\mathrm{G}} = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\,, \quad \text{IPW: } \Gamma_i^{\mathrm{IPW}} = \frac{A_i}{\hat{\pi}(X_i)}Y_i - \frac{1-A_i}{1-\hat{\pi}(X_i)}Y_i \tag{1}$$

$$\text{AIPW: } \Gamma_i^{\mathrm{AIPW}} = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{A_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{1-A_i}{1-\hat{\pi}(X_i)}(Y_i - \hat{\mu}_0(X_i))\,. \tag{2}$$

We denote by $\hat{\tau}_{\mathrm{G}}$, $\hat{\tau}_{\mathrm{IPW}}$ and $\hat{\tau}_{\mathrm{AIPW}}$ the G-formula, IPW and AIPW estimators respectively.

In causal inference, the theoretical "utility" of an estimator is typically reported as its asymptotic variance, rather than through some finite sample bounds. For the three estimators above, the following asymptotic variance results (asymptotic unbiasedness and normality) hold [Wager, 2024] under Assumptions 1 to 3:

$$\sqrt{n}(\hat{\tau} - \mathbb{E}\left[Y_i(1) - Y_i(0)\right]) \to_{\mathbb{P}} \mathcal{N}(0, V^\star)\,, \text{with}$$

$$V_{\mathrm{G}}^\star := \mathrm{var}\left(\mu_1(X) - \mu_0(X)\right), \quad V_{\mathrm{IPW}}^\star := \mathbb{E}\Big[\frac{Y(1)^2}{\pi(X)}\Big] + \mathbb{E}\Big[\frac{Y(0)^2}{1-\pi(X)}\Big] - \tau^2\,, \tag{3}$$

$$V_{\mathrm{AIPW}}^\star := \mathbb{E}\Big[\frac{(Y(1) - \mathbb{E}\left[Y(1)|X\right])^2}{\pi(X)}\Big] + \mathbb{E}\Big[\frac{(Y(0) - \mathbb{E}\left[Y(0)|X\right])^2}{1-\pi(X)}\Big] + \mathrm{var}\left(\mu_1(X) - \mu_0(X)\right)\,, \tag{4}$$

provided that the nuisance parameters are pointwise consistent and $\mathbb{E}\big[(\hat{\mu}_a(i) - \mu_a(X_i))^2\big] = o_{\mathbb{P}}(n^{-1})$ (for G-Formula), $\mathbb{E}\big[(\hat{\pi}(X_i) - \pi(X_i))^2\big] = o_{\mathbb{P}}(n^{-1})$ (for IPW), or $\mathbb{E}\big[(\hat{\pi}(X_i) - \pi(X_i))^2\big]\mathbb{E}\big[(\hat{\mu}_a(X_i) - \mu_a(X_i))^2\big] = o_{\mathbb{P}}(n^{-1})$ (for AIPW). Note that under this assumption, AIPW estimated with cross-fitting [Chernozhukov et al., 2018] achieves the semi-parametric efficient variance. Importantly, it is also *doubly robust*, meaning it remains consistent if either the outcome model or the propensity score model is correctly specified—a key reason for its popularity.

In practice, the asymptotic variance $V^\star$ is used to give 95% coverage confidence intervals of the form $\big[\hat{\tau} - 1.96\sqrt{V^\star}, \hat{\tau} + 1.96\sqrt{V^\star}\big]$, which are asymptotically valid. Since $V^\star$ is usually unknown, it is often estimated as $\hat{V} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\Gamma_i - \hat{\tau})^2$ [Wager, 2024].

[1] These non-parametric models are usually trained on a dataset independent of $(X_i, A_i, Y_i)_{i\in[n]}$. Cross-fitting [Chernozhukov et al., 2018, Athey and Wager, 2021] can also be used to obtain similar asymptotic results.

2.3 Differential Privacy

Informally, Differential Privacy (DP) [Dwork et al., 2006, Dwork and Roth, 2014] guarantees that the output distribution of a (randomized) algorithm does not differ significantly when a single individual's data is modified, thereby limiting what an adversary can infer. DP is widely considered the gold standard for privacy-preserving data analysis due to several key properties: *(i)* it provides strong, mathematically rigorous guarantees even against adversaries with arbitrary side information; *(ii)* it is *immune to post-processing*, meaning that any function applied to the output of a differentially private mechanism cannot degrade its privacy guarantees; and *(iii)* it supports precise accounting of the cumulative privacy loss through the principle of *composition*, allowing multiple analyses to be performed on the same data with controlled privacy degradation.

In this work, we express privacy guarantees in the Gaussian Differential Privacy (GDP) framework [Dong et al., 2022], which precisely characterizes the privacy loss incurred by adding Gaussian noise and tightly handles composition. GDP is increasingly regarded as best practice for reporting DP guarantees when applicable [Gomez et al., 2025], and converts losslessly to the classical (ε, δ)-DP definition (see Appendix A). GDP formulates privacy as a hypothesis testing problem: an adversary attempts to distinguish between the outputs of an algorithm $\mathcal{A}$ run on two *neighboring datasets*, denoted $\mathcal{D} \sim \mathcal{D}'$, which differ by the replacement of a single data point.[2] The inherent difficulty of this task is captured by the *trade-off function*, which describes the balance between type I (false positive) and type II (false negative) errors in the adversary's decision process.

Definition 2 (Trade-off function)**.** *Let P and Q be any two probability distributions defined on the same space. Let $0 \leqslant \phi \leqslant 1$ represent any rejection rule that outputs the probability of rejecting the null hypothesis $H_0 : P$ against $H_1 : Q$. The trade-off function $T(P, Q) : [0, 1] \to [0, 1]$ is defined as*

$$T(P, Q)(\alpha) := \inf\{1 - \mathbb{E}_Q[\phi] : \mathbb{E}_P[\phi] \leqslant \alpha\},$$

where the infimum is taken over all rejection rules.

We can now introduce the formal definition of GDP.

Definition 3 (ζ-Gaussian Differential Privacy [Dong et al., 2022])**.** *Let $\zeta \geqslant 0$. We say that a randomized algorithm $\mathcal{A} \colon \mathcal{X}^n \to \mathcal{R}$ satisfies ζ-GDP if for all $\mathcal{D} \sim \mathcal{D}'$ we have*

$$T(\mathcal{A}(\mathcal{D}), \mathcal{A}(\mathcal{D}')) \geqslant T(\mathcal{N}(0, 1), \mathcal{N}(\zeta, 1)) \ .$$

In other words, an algorithm $\mathcal{A}$ satisfies ζ-GDP if distinguishing between its outputs on two neighboring datasets is at least as hard as distinguishing between two unit-variance Gaussians: one with a fixed mean and the other with a mean shifted by ζ. Thus, the smaller ζ, the stronger the privacy guarantees. Our private ATE estimates will rely on the Gaussian mechanism to enforce privacy.

Lemma 1 (Gaussian mechanism [Dong et al., 2022])**.** *Let $f \colon \mathcal{X}^n \to \mathbb{R}$ be a function with sensitivity $\max_{\mathcal{D} \sim \mathcal{D}'} |f(\mathcal{D}) - f(\mathcal{D}')| \leqslant \Delta$. Then $\mathcal{A}(\mathcal{D}) := f(\mathcal{D}) + Z$, with $Z \sim \mathcal{N}(0, \Delta^2/\zeta^2)$, satisfies ζ-GDP.*

The Gaussian mechanism adds Gaussian noise calibrated to the *sensitivity* of the computation—that is, the maximum change in the output resulting from modifying a single data point.

3 Related work

Previous work on private ATE estimation falls into two main categories. The first relies on simple difference-in-means estimators that avoid estimating nuisance functions. Most of these methods are limited to RCTs [Chen et al., 2024, Javanmard et al., 2024, Yao et al., 2024], whereas our focus is on observational data. Koga et al. [2024] propose a matching-based estimator for observational settings, pairing each data point with another that has identical covariates but opposite treatment assignment. However, their method assumes discrete covariates and requires well-balanced treatment groups (because unmatched samples are discarded). We do not make such strong assumptions in this work.

The second category uses estimators that rely on nuisance functions like those considered in this paper. These methods follow a two-step procedure: (i) fit the nuisance functions using a differentially private algorithm, and (ii) plug them into the ATE estimator and add noise to privatize the result.

[2] That is, $\exists i \in [n]$ s.t. $\mathcal{D}_j = \mathcal{D}'_j$ for all $j \neq i$. This is sometimes referred to as *replace-one* or *bounded* DP.

Differential privacy follows by composition from the privacy of each step. Lee et al. [2019] proposed a private version of IPW using logistic regression fitted via Differentially Private Empirical Risk Minimization (DP-ERM). Ohnishi and Awan [2024] also consider logistic regression together with covariate balancing, which can reduce the bias of Lee et al. [2019] in some cases. Privacy is enforced with the 2-Norm Gradient Mechanism, reducing noise but making the approach intractable with more than a few features. Both works make the assumption of a linear propensity score and are limited to the IPW estimator, which often performs poorly in practice even in well-specified, non-private settings [Kang and Schafer, 2007]. Furthermore, making the propensity score model DP increases the privacy cost with the dimension [Bassily et al., 2014].[3] This added cost is unnecessary when only the final ATE estimate is of interest. In contrast, our work adds noise only to the final ATE estimate, supports arbitrary models (including non-parametric ones), and privatizes not only IPW, but also the G-Formula and doubly-robust AIPW estimators, which are widely used and effective in practice.

Concurrently with our work, Guha and Reiter [2025] proposed private IPW estimators using a data-splitting scheme that bears some resemblance to ours. However, their approach differs fundamentally: they estimate the propensity scores and ATE *independently on each data split*, then aggregate the results in a meta-analysis fashion. In contrast, we estimate nuisance functions on all but one split and compute the ATE on the held-out split—reducing variance and ensuring the independence structure needed for establishing asymptotic normality in non-parametric settings. Additionally, their method is limited to IPW with binary outcomes.

We note that recent work explored the challenging problem of estimating of the Conditional Average Treatment Effect (CATE) under DP [Nori et al., 2021, Niu et al., 2022, Schröder et al., 2025]. In this paper, however, we focus on the ATE, which is the primary estimand in many applied settings.

4 Differentially private estimators

In this section, we introduce our private estimators of the ATE. We first introduce our general framework, which we then instantiate to present differentially private G-Formula, IPW and AIPW estimators. Finally, we present a unified privacy and utility analysis of these estimators.

4.1 General framework

In the ATE estimators defined in Section 2.2, it is necessary to protect both the training data used to learn the nuisance functions and the data points on which the ATE is evaluated. Our goal in this work is to avoid making strong structural assumptions (such as parametric forms) about the data or the nuisance functions. However, this flexibility introduces a key challenge: without such assumptions, the sensitivity of these estimators—which determines the scale of the noise required for DP (see Lemma 1)—becomes inherently high. In particular, modifying a single data point can arbitrarily impact the fitted nuisance models, leading to a sensitivity of $O(1)$ that does not decrease with dataset size, making accurate private ATE estimation infeasible even with large datasets. A straightforward strategy would be to train *differentially private non-parametric models* for the nuisance functions π, μ_0, μ_1. However, this tends to be privacy-inefficient and restricts the use of powerful, state-of-the-art methods such as causal forests, limiting both flexibility and performance.

Instead, our approach partitions the dataset $\mathcal{D}$ into K folds and learns nuisance estimators independently on each fold. Crucially, since only the final ATE estimate is released—and not the nuisance models themselves—there is no need to make these models differentially private. Then, for each fold, we aggregate the predictions from the models trained on the $K-1$ other folds to form ensemble estimates. This strategy substantially reduces the overall sensitivity of the final computation, allowing us to use flexible, non-private nuisance models while still achieving strong privacy guarantees for the released ATE. Formally, our framework follows the following five steps.

1. **Data splitting.** Partition $\mathcal{D}$ into K folds $(\mathcal{I}_k)_{k\in[K]}$. For all $i \in [n]$, there exists a unique $k(i) \in [K]$ such that $i \in \mathcal{I}_{k(i)}$.
2. **Nuisance estimation.** Learn K sets of (potentially non-parametric) estimators $\left(\hat{\pi}^{(k)}, \hat{\mu}_1^{(k)}, \hat{\mu}_0^{(k)}\right)_{k\in[K]}$, where $\hat{\pi}^{(k)}, \hat{\mu}_1^{(k)}, \hat{\mu}_0^{(k)}$ are trained on $\mathcal{I}_k$.

[3] This dependence is hidden by assuming $\|X_i\|_2 \leqslant 1$ in Lee et al. [2019], Ohnishi and Awan [2024].

3. **Aggregation.** For all $i \in [n]$, compute the estimated nuisance parameters of patient i as:

$$\begin{aligned} \hat{\pi}_1(i) &= \Big(\tfrac{1}{K-1} \textstyle\sum_{k \in [K] \setminus \{k(i)\}} \tfrac{1}{\hat{\pi}^{(k)}(X_i)} \Big)^{-1}, \\ 1 - \hat{\pi}_0(i) &= \Big(\tfrac{1}{K-1} \textstyle\sum_{k \in [K] \setminus \{k(i)\}} \tfrac{1}{1-\hat{\pi}^{(k)}(X_i)} \Big)^{-1}, \\ \hat{\mu}_a(i) &= \tfrac{1}{K-1} \textstyle\sum_{k \in [K] \setminus \{k(i)\}} \hat{\mu}_a^{(k)}(X_i) \,. \end{aligned} \tag{5}$$

In other words, the propensity score (resp. the conditional outcome) of patient i is estimated using a harmonic (resp. arithmetic) mean over the $K-1$ models trained *excluding* patient i.[4] Intuitively, this design ensures that the sensitivity of the nuisance estimators $\hat{\mu}_a$, $\frac{1}{\hat{\pi}_1}$, $\frac{1}{1-\hat{\pi}_0}$ is $O(1/K)$, which decreases as the number of samples n and the number of folds K increase. Note that this implies separate propensity estimates for treated and control patients.

4. **Private ATE estimator.** For all $i \in [n]$, compute the score $\hat{\Gamma}_i$ of patient i as in Section 2.2 but using the "ensemble" nuisance models defined in Equation (5); For instance, $\hat{\Gamma}_i = \hat{\mu}_1(i) - \hat{\mu}_0(i) + \frac{A_i}{\hat{\pi}_1(i)}(Y_i - \hat{\mu}_1(i)) - \frac{1-A_i}{1-\hat{\pi}_0(i)}(Y_i - \hat{\mu}_0(i))$ for AIPW. Then, aggregate these scores and add Gaussian noise to get the private ATE estimate $\hat{\tau}_{\mathrm{DP}}$:

$$\hat{\tau}_{\mathrm{DP}} := \hat{\tau} + \mathcal{N}(0, \sigma_1^2) \,, \quad \text{with} \qquad \hat{\tau} := \tfrac{1}{n} \textstyle\sum_{i=1}^n \hat{\Gamma}_i \,. \tag{6}$$

5. **Private variance estimation and confidence interval.** Privately estimate the variance of the estimator as

$$\hat{V}_{\mathrm{DP}} := \left(\sqrt{\tfrac{1}{n(n-1)} \textstyle\sum_{i=1}^n \left(\hat{\Gamma}_i - \hat{\tau} \right)^2} + \mathcal{N}(0, \sigma_2^2) \right)^2 + \sigma_1^2 + 2.33\sigma_2^2 \,, \tag{7}$$

and construct a private asymptotically valid 95% coverage confidence interval:[5]

$$\left[\hat{\tau}_{\mathrm{DP}} - 2.05\sqrt{\hat{V}_{\mathrm{DP}}} \,,\, \hat{\tau}_{\mathrm{DP}} + 2.05\sqrt{\hat{V}_{\mathrm{DP}}} \right] . \tag{8}$$

Different choices of scores $\hat{\Gamma}_i$ yield our three private ATE estimators: $\hat{\tau}_{\mathrm{DP-G}}$, $\hat{\tau}_{\mathrm{DP-IPW}}$, and $\hat{\tau}_{\mathrm{DP-AIPW}}$.

Remark 1 (Data splitting and connections to existing techniques)**.** *Our data splitting scheme bears a resemblance to cross-fitting, which is commonly used to reduce overfitting when estimating nuisance parameters [Chernozhukov et al., 2018, Athey and Wager, 2021, Bach et al., 2024]. While the case of $K = 2$ folds aligns with standard cross-fitting, for $K > 2$, our approach departs from the classical setup. Unlike traditional cross-fitting, our primary motivation for data splitting is to reduce sensitivity and enable favorable privacy-utility trade-offs, rather than solely improving statistical efficiency [Bach et al., 2024]. Our approach also shares structural similarities with PATE (Private Aggregation of Teacher Ensembles) [Papernot et al., 2018], a differential privacy framework that splits data into disjoint subsets, each used to train a separate "teacher" model. These teachers vote on unlabeled examples, and their aggregated (noisy) votes are used to label a public dataset, which then trains a final "student" model. As in our framework, PATE ensures that individual data points influence only a single model. However, the goals, settings, and aggregation mechanisms differ substantially.*

4.2 Unified privacy and utility analysis

We now present a unified privacy and utility analysis of our private ATE estimators, deriving expressions for setting the noise multipliers σ_1^2 and σ_2^2 to achieve a target privacy level, and providing utility guarantees in terms of the resulting asymptotic variance.

Privacy analysis. For our privacy analysis, we assume that the outcomes and nuisance function estimators are bounded. A well-known negative result in DP states that privately estimating the mean of unbounded reals is impossible without distributional assumptions [Bun et al., 2015]. In practice, this boundedness assumption can be enforced through clipping. In Assumption 4, B_π is either $1/\eta$ if this quantity is known, an upper bound on $1/\eta$, or a positivity imposed by clipping if η is too small.

[4] Using the harmonic mean for propensity scores is natural, as the ATE estimators involve the *inverses* of these scores.

[5] We generalize this approach to more general confidence intervals in Appendix B.1, where numerical constants are replaced by plug-in parameters that use the Gaussian queue distribution.

Assumption 4 (Bounded outcomes and nuisance estimators)**.** *There exists* $B_\mu > 0$ *and* $0 < B_\pi < 1$ *s.t.* $\mathcal{Y} \subseteq [-B_\mu, B_\mu]$, $|\hat{\mu}_a^{(k)}(x)| \leqslant B_\mu$ *and* $\max\left(\frac{1}{\hat{\pi}^{(k)}(x)}, \frac{1}{1-\hat{\pi}^{(k)}(x)}\right) \leqslant B_\pi$ *for* $k \in [K]$, $x \in \mathcal{X}$, *and* $a \in \{0, 1\}$.

Theorem 1 (Privacy analysis)**.** *Assume that Assumption 4 hold. Let* $\zeta_1, \zeta_2 > 0$*. If we set the noise multiplier* σ_1^2 *to*

$$\sigma_1^2 = \frac{C}{\zeta_1^2}\left(\frac{1}{n} + \frac{1}{K-1}\right)^2, \tag{9}$$

where respectively for the cases where $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-G}}$, $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-IPW}}$ *and* $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-AIPW}}$*:*

$$C_{\mathrm{DP-G}} = 16B_\mu^2, \quad C_{\mathrm{IPW}} = 4B_\mu^2 B_\pi^2, \quad \textit{and} \quad C_{\mathrm{AIPW}} = 16B_\mu^2\left(1 + B_\pi\right)^2, \tag{10}$$

then releasing the private ATE estimator $\hat{\tau}_{\mathrm{DP}}$ *satisfies* ζ_1*-GDP. Furthermore, if we set* σ_2^2 *to:*

$$\sigma_2^2 = \frac{2C}{\zeta_2^2(n-1)}\left(\frac{1}{n} + \frac{1}{K-1} + \sqrt{\frac{1}{n} + \frac{1}{K-1}}\right)^2,$$

then releasing the confidence interval defined in Equation (8) satisfies $\sqrt{\zeta_1^2 + \zeta_2^2}$*-GDP.*

A few remarks about this theorem are in order. First, Equation (9) makes explicit how the privacy noise level σ_1^2 depends on the number of data points n, the number of folds K, and the bounds B_μ and B_π on the outcomes and imposed overlap, through the constant C. Our data-splitting and ensembling scheme ensure that the privacy cost (i.e., the variance σ_1^2 of the added noise) of achieving a given privacy level ζ_1 for the ATE estimate decreases with both n and K. Since in practice K is typically much smaller than n to preserve enough data in each fold for reliable nuisance estimation, the noise variance is of order $\frac{1}{K^2}$. Second, the additional noise σ_2^2 needed for the confidence interval is substantially smaller, scaling as $\frac{1}{nK^2}$, so the interval can be released with only a minor additional privacy cost. Importantly, unlike prior approaches that enforce privacy by directly privatizing the nuisance model estimators [Lee et al., 2019, Ohnishi and Awan, 2024], our privacy guarantees do not require any structural assumption on these models (beyond boundedness).

The constant C in Equation (10) is the only component of the privacy bound that depends on the specific choice of ATE estimator. This unified analysis underscores the flexibility of our approach, which could be extended to other estimators. In settings with large overlap η (small B_π)—resembling RCTs—the privacy cost introduced by the use of propensity models in IPW and AIPW estimators remains modest. However, in small-overlap regimes, which are challenging even in non-private causal inference, the required noise for privacy increases substantially for these estimators. Therefore, the G-formula estimator, whose privacy cost is unaffected by overlap, may thus be preferable in such scenarios.

Utility analysis. We now turn to the utility analysis of our private ATE estimators. Specifically, we provide an asymptotic bound on their variance and on the 95% coverage confidence intervals.

Theorem 2 (Utility analysis)**.** *Assume that Assumptions 1 to 4 hold. Assume that the (non-parametric) estimators* $\hat{\mu}_1^{(k)}, \hat{\mu}_0^{(k)}, \hat{\pi}^{(k)}$ *are all point-wise consistent almost everywhere and satisfy* $\mathbb{E}\left[\left(\hat{\mu}_a(i) - \mu_a(X_i)\right)^2\right] = o_{\mathbb{P}}(n^{-1})$ *if* $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-G}}$, $\mathbb{E}\left[\left(\hat{\pi}(i) - \pi(X_i)\right)^2\right] = o_{\mathbb{P}}(n^{-1})$ *if* $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-IPW}}$ *and* $\mathbb{E}\left[\left(\hat{\pi}(i) - \pi(X_i)\right)^2\right]\mathbb{E}\left[\left(\hat{\mu}_a(i) - \mu_a(X_i)\right)^2\right] = o_{\mathbb{P}}(n^{-1})$ *if* $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-AIPW}}$*. Then,* $\hat{\tau}_{\mathrm{DP-G}}$, $\hat{\tau}_{\mathrm{DP-IPW}}$ *and* $\hat{\tau}_{\mathrm{DP-AIPW}}$ *are asymptotically unbiased and the confidence interval defined in Equation (8) with* $\hat{V}_{\mathrm{DP}}$ *defined in Equation (7) is an asymptotically valid* 95% *coverage confidence interval, and we have that, for* σ_1^2, σ_2^2 *and* C *as in Theorem 1:*

$$\hat{V}_{\mathrm{DP}} = \frac{V^\star}{n} + \frac{C}{\zeta_1^2}\frac{1}{(K-1)^2} + \frac{4.66C}{\zeta_2^2}\frac{1}{n-1}\left(\frac{1}{K-1} + \sqrt{\frac{1}{K-1}}\right)^2 + o\left(n^{-1}\right),$$

where $V^\star = V_{\mathrm{G}}^\star$, $V^\star = V_{\mathrm{IPW}}^\star$ *or* $V^\star = V_{\mathrm{AIPW}}^\star$ *in the case of respectively* $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-G}}$, $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-IPW}}$ *and* $\hat{\tau}_{\mathrm{DP}} = \hat{\tau}_{\mathrm{DP-AIPW}}$*, for* $V_{\mathrm{G}}^\star, V_{\mathrm{IPW}}^\star, V_{\mathrm{AIPW}}^\star$ *defined in Equations (3) and (4).*

Proof. The proof follows from classical non-DP proofs that use crossfitting [Wager, 2024, e.g., Theorem 3.2], adapted to our setting. See Appendix B.1 for details on CIs asymptotic validity. □

This utility theorem shows that for large n and K of order $\sqrt{n}$, the additional variance introduced by privacy is of the same order as the oracle variance of the non-private estimator. Our utility result also does not depend on the dimension of the covariates or an imposed bound on the covariates, as opposed to previous approaches that paid the cost of privatizing nuisance models themselves. Since we here only privatize scalars, our error does not grow with the dimension.

Remark 2 (Model sensitivity)**.** *Our privacy analysis assumes worst-case model sensitivities, allowing a single data point in $\mathcal{I}_k$ to affect the output of $\left(\hat{\pi}^{(k)}, \hat{\mu}_1^{(k)}, \hat{\mu}_0^{(k)}\right)$ arbitrarily (within the bounds of Assumption 4). While this is conservative, it ensures broad applicability. Nonetheless, our analysis can directly incorporate smaller sensitivities—potentially as low as $O(1/|\mathcal{I}_k|)$—when using models and training algorithms for which such bounds are known [see e.g., Chaudhuri et al., 2011], thereby enhancing the privacy-utility trade-off.*

5 Meta-Analysis of Differentially Private ATEs

Meta-analyses are statistical methods for combining the results from $N \geqslant 2$ independent studies to increase statistical power [Hunter and Schmidt, 2004, Borenstein et al., 2021]. They are considered the pinnacle of evidence in clinical research hierarchies. In this section, we show how to combine several *private* ATE estimates via meta-analysis to achieve a lower-variance estimate.

We assume each study $j \in [N]$ releases an ATE estimate $\hat{\tau}_{\mathrm{DP}}^{(j)}$ and its estimated variance $\hat{V}_{\mathrm{DP}}^{(j)}$, computed from an independent sample of size n_j drawn from the same population, with $\zeta^{(j)}$-GDP guarantees. Constructing such estimates can be handled using the framework presented in Section 4. Note that each release can be conducted independently—in particular, we do not require the studies to use the same ATE estimator.

Given some weights $\lambda_1, \ldots, \lambda_N \geqslant 0$ with $\sum_{j=1}^N \lambda_j = 1$, the meta-analysis ATE estimate then writes as $\hat{\tau}_{\mathrm{DP-meta}} = \sum_{j=1}^N \lambda_j \hat{\tau}_{\mathrm{DP}}^{(j)}$, where $\lambda_1, \ldots, \lambda_N \geqslant 0$, which is asymptotically unbiased with an estimated variance $\hat{V}_{\mathrm{DP-meta}} = \sum_{j=1}^N \lambda_j^2 \hat{V}_{\mathrm{DP}}^{(j)}$. Importantly, as long as the weights are determined independently of private data, releasing $\hat{\tau}_{\mathrm{DP-meta}}$ does not affect the differential privacy guarantees, since it is a post-processing of already privatized estimates.

To minimize the variance of $\hat{\tau}_{\mathrm{DP-meta}}$, one should choose weights that are inversely proportional to the variance of the original estimates.

Proposition 1. *For $\lambda_j^* = \frac{\left(\hat{V}_{\mathrm{DP}}^{(j)}\right)^{-1/2}}{\sum_{k=1}^N \left(\hat{V}_{\mathrm{DP}}^{(k)}\right)^{-1/2}}$, we have $\hat{V}_{\mathrm{DP-meta}} = \frac{1}{N}\left(\frac{1}{N}\sum_{j=1}^N \left(\hat{V}_{\mathrm{DP}}^{(j)}\right)^{-1/2}\right)^{-2}$.*

As expected, this variance decreases as the number of studies increases, at a rate of $1/N$ when all variances are of the same order. Assuming our estimators are used, and applying the asymptotic variance results from Theorem 2, the optimal weights are, to a first-order approximation, $\lambda_j^* \propto \left(\frac{V_j^\star}{n_j} + \frac{C_j}{(\zeta_1^{(j)}(K_j-1))^2}\right)^{-1/2} + o(1/n_j) + o(1/K_j^2)$, where $V_j^\star$ is the oracle variance, K_j the number of folds, and C_j the estimator-specific constant (see Theorem 1) for study j. Here, the classical oracle variance term appears—assigning smaller weights to studies with higher variance—augmented by an additional term reflecting the impact of privacy: studies with weaker privacy constraints receive larger weights in the meta-analysis.

6 Experiments

We present a set of experiments that estimate the ATE on synthetic data. We compare our private estimators with the prior approaches proposed by Lee et al. [2019], Ohnishi and Awan [2024], which we refer to as LPGM19 and OA24, respectively. We also include an experiment with binary outcomes in Appendix F.2, in which we compare against the method from Guha and Reiter [2025], denoted GR25. To demonstrate the flexibility of our approach, we consider multiple synthetic data generation setups. In all cases, covariates are sampled from a standard d-dimensional Gaussian, i.e., $X_i \sim \mathcal{N}(0, I_d)$. We keep data low-dimensional ($d \leqslant 10$) as OA24 becomes intractable for larger dimensions or small overlaps due to exponential explosion of the computation times. LPGM19 and

OA24 assume that the ℓ_2 norm of covariates is bounded by 1. To satisfy this assumption, we rescale the covariates such that their norms are bounded by 1 for at least 99% of the samples, and we apply clipping for any remaining samples that exceed the bound. This setup provides a best-case scenario for their methods, where an approximate norm bound is enforced. We set $B_\mu = 1$ for all experiments and repeat each setup 100 times. Privacy budget used in GDP and $\varepsilon-$DP are respectively 1.5 and 7.05 if not specified otherwise ($1.5-$GDP implies $(7.05, 10^{-5})-$DP). The details of all experiments can be found in the appendix.

Well-specified setting. We first consider a setting where the estimators are well-specified, with a logistic regression model for π (in accordance with the assumption made by LPGM19 and OA24), and a linear regression model for μ_0, μ_1. Previous work has primarily focused on IPW-based methods, which are known to suffer from high variance when the overlap is limited, due to large weights assigned to certain samples, even in non-private settings [Kang and Schafer, 2007]. We illustrate this issue by generating data with low overlap ($\eta \approx 0.004$). The results are shown in the first plot of Figure 1, where we used $K = 200$ for our estimators. Setting $B_\pi = 1/0.0035$ (i.e., close to the true overlap), all IPW-based estimators exhibit large variance. On the other hand, clipping propensity scores more aggressively to $B_\pi = 1/0.05$ or $B_\pi = 1/0.2$ reduces variance but introduces bias. In contrast to IPW-based estimators, our G-formula estimator performs very well, as the noise added for privacy does not depend on the overlap. Similarly, AIPW remains unbiased with aggressive clipping, as it automatically relies heavily on its well-performing outcome regression part. In Appendix F.2, we report results for an additional well-specified setting with good overlap and binary outcomes. As expected, LPGM19, OA24, and GR25 perform well in this scenario, which aligns precisely with the assumptions they were designed to exploit. Nonetheless, several of our estimators achieve comparable performance.

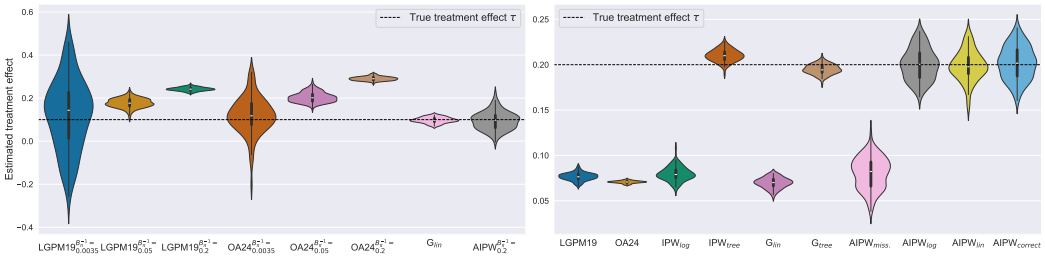

Figure 1: The left plot depicts a well-specified setting with low overlap. The right plot shows a setting where misspecified estimators have large bias.

Misspecified setting. We now turn to the setting where estimators can be misspecified. Here, propensity scores and outcome responses are modeled by decision trees. We use $K = 500$ for our estimators. The second plot of Figure 1 shows that previous methods, which rely on parametric (linear) nuisance models, are thus highly biased. In contrast, when instantiated with decision tree models, our three estimators do not suffer from such biases. We also see that, thanks to its double-robustness property, our AIPW estimator performs well even if one of the models is misspecified.

Effect of parameter K. Our approach requires selecting the number of folds K. In Appendix F.2, we analyze the effect of this parameter and examine the sensitivity of our methods to its choice.

Conclusion

In this work, we provided a general *plug-in and model-agnostic approach* to estimate the average treatment effect under differential privacy constraints. Our approach departs from previous works, that either relied on RCT settings or parametric assumptions on nuisance models. As demonstrated by our experiments, the flexibility of our methodology enables strong performance across a variety of settings. Our work paves the way towards broader adoption of DP in causal inference, but also opens several promising directions for future research, including: *(i)* extending our framework to the private estimation of heterogeneous treatment effects, and *(ii)* developing methods for generalizing causal estimates across populations under differential privacy constraints—for instance, enabling private transportability of ATEs when the source and target populations differ.

7 Acknowledgements

We thank Ahmed Boughdiri and Tudor Cebere for helpful comments. We thank Yuki Ohnishi for sharing the implementation of their technique. The work of Christian Lebeda and Aurélien Bellet is supported by grant ANR-20-CE23-0015 (Project PRIDE) and the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR. Mathieu Even and Julie Josse acknowledge fundings by Theremia.

References

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Philipp Bach, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. In *Causal Learning and Reasoning*, pages 1065–1117. PMLR, 2024.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 403–412. PMLR, 2018.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *FOCS*, 2014.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In *FOCS*, pages 634–649. IEEE Computer Society, 2015.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Wei-Ning Chen, Graham Cormode, Akash Bharadwaj, Peter Romov, and Ayfer Özgür. Federated experiment design under distributed differential privacy. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 2458–2466. PMLR, 2024. URL https://proceedings.mlr.press/v238/chen24c.html.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Damien Desfontaines. Converters between differential privacy variants. https://desfontain.es/blog/converters-differential-privacy.html, 06 2024. Ted is writing things (personal blog).

Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. Confidence-ranked reconstruction of census microdata from published statistics. *Proceedings of the National Academy of Sciences*, 120(8), 2023.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, 2003.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

Juan Felipe Gomez, Bogdan Kulynych, Georgios Kaissis, Jamie Hayes, Borja Balle, and Antti Honkela. (ε, δ) considered harmful: Best practices for reporting differential privacy guarantees, 2025. URL https://arxiv.org/abs/2503.10945.

Sharmistha Guha and Jerome P. Reiter. Differentially private estimation of weighted average treatment effects for binary outcomes. *Computational Statistics & Data Analysis*, 207:108145, 2025.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLOS Genetics*, 4(8):1–9, 2008.

John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Adel Javanmard, Vahab Mirrokni, and Jean Pouget-Abadie. Causal inference with differentially private (clustered) outcomes, 2024. URL https://arxiv.org/abs/2308.00957.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), November 2007. ISSN 0883-4237. doi: 10.1214/07-sts227. URL http://dx.doi.org/10.1214/07-STS227.

Tatsuki Koga, Kamalika Chaudhuri, and David Page. Differentially private multi-site treatment effect estimation. In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2024, Toronto, ON, Canada, April 9-11, 2024*, pages 472–489. IEEE, 2024. doi: 10.1109/SATML59370.2024.00030. URL https://doi.org/10.1109/SaTML59370.2024.00030.

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

Ilya Mironov. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, CSF '17, pages 263–275, Washington, DC, USA, 2017. IEEE Computer Society.

Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *1st Conference on Causal Learning and Reasoning, CLeaR 2022, Sequoia Conference Center, Eureka, CA, USA, 11-13 April, 2022*, volume 177 of *Proceedings of Machine Learning Research*, pages 618–633. PMLR, 2022. URL https://proceedings.mlr.press/v177/niu22a.html.

Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. Accuracy, interpretability, and differential privacy via explainable boosting. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8227–8237. PMLR, 2021.

Yuki Ohnishi and Jordan Awan. Differentially private covariate balancing causal inference. *CoRR*, abs/2410.14789, 2024. doi: 10.48550/ARXIV.2410.14789. URL https://doi.org/10.48550/arXiv.2410.14789.

Yuki Ohnishi and Jordan Awan. Locally private causal inference for randomized experiments. *Journal of Machine Learning Research*, 26(14):1–40, 2025.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *ICLR*, 2018.

Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. In *25th Annual Network and Distributed System Security Symposium, NDSS*, 2018.

Maresa Schröder, Valentyn Melnychuk, and Stefan Feuerriegel. Differentially private learners for heterogeneous treatment effects. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1z3SOCwst9.

Suyash S. Shringarpure and Carlos D. Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 97(5):631–646, 2015.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 2015. doi: 10.1371/journal.pmed.1001779.

Stefan Wager. Causal inference: A statistical learning approach, 2024.

Leon Yao, Paul Yiming Li, and Jiannan Lu. Privacy-preserving quantile treatment effect estimation for randomized controlled trials, 2024. URL https://arxiv.org/abs/2401.14549.

A Conversion from ζ-GDP to (ε, δ)-DP

For completeness, we discuss here the relationship between ζ-GDP and the classical (ε, δ)-DP definition, which we recall below. For further discussion on its interpretation, we refer the reader to [Dwork and Roth, 2014].

Definition 4 ((ε, δ)-DP [Dwork and Roth, 2014])**.** *A randomized algorithm $\mathcal{A}$ is (ε, δ)-DP if for any pair of neighboring datasets $\mathcal{D} \sim \mathcal{D}'$, and any event E,*

$$\mathbb{P}[\mathcal{A}(\mathcal{D}) \in E] \leqslant e^{\varepsilon}\mathbb{P}[\mathcal{A}(\mathcal{D}') \in E] + \delta.$$

where the probability is taken over the randomness of $\mathcal{A}$.

Unlike GDP, (ε, δ)-DP does not tightly handle composition. Yet, (ε, δ)-DP offers more intuitive control over the privacy leakage (especially when δ is taken to be negligible) and is often referenced in legal, policy, and industry contexts as the standard formal privacy definition. It can thus be useful to convert GDP guarantees to (ε, δ)-DP guarantees. Desfontaines [2024] provides a helpful online tool for converting specific parameters.

Lemma 2 (Conversion from ζ-GDP to (ε, δ)-DP [Balle and Wang, 2018, Dong et al., 2022])**.** *Let $\mathcal{A}$ be an algorithm satisfying $\zeta-$GDP. Then $\mathcal{A}$ also satisfies $(\varepsilon, \delta)-$DP for any $\varepsilon > 0$ where*

$$\delta = \Phi\left(-\frac{\varepsilon}{\zeta} + \frac{\zeta}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon}{\zeta} - \frac{\zeta}{2}\right) \quad ,$$

where Φ is the CDF of the standard Gaussian distribution.

In our experiments, we report the privacy guarantees in both ζ-GDP and the corresponding (ε, δ)-DP guarantee obtained with Lemma 2 for a fixed value of δ.

We note that it is also possible to convert GDP guarantees to other popular variants of DP, such as Rényi DP [Mironov, 2017, Corollary 3] and zCDP [Bun and Steinke, 2016, Proposition 1.6].

B Unified analysis

B.1 General confidence intervals

In the confidence interval step of our method, we could generalize and instead privately estimate the variance of the estimator as, for $p \in (0, 1)$ and $p_1 \in (0, p/2)$[6]:

$$\hat{V}_{\mathrm{DP}} := \left(\sqrt{\tfrac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\Gamma}_i - \hat{\tau}\right)^2} + \mathcal{N}(0, \sigma_2^2)\right)^2 + \sigma_1^2 + \Phi^{-1}(1-p_1)\sigma_2^2\,, \tag{11}$$

where $\Phi(t) = \mathbb{P}\left(\mathcal{N}(0,1) \geqslant t\right)$ for $t \in \mathbb{R}$, and construct a private asymptotically valid $p-$coverage confidence interval:

$$\mathrm{CI}_p := \left[\hat{\tau}_{\mathrm{DP}} - \Phi^{-1}(1 - p/2 + p_1/2)\hat{V}_{\mathrm{DP}}^{1/2}\,,\ \hat{\tau}_{\mathrm{DP}} + \Phi^{-1}(1 - p/2 + p_1/2)\hat{V}_{\mathrm{DP}}^{1/2}\right]. \tag{12}$$

Indeed, we first have that, for $\hat{V} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\Gamma}_i - \hat{\tau}\right)^2$:

$$\begin{aligned}
&\mathbb{P}\left(\left(\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\Gamma}_i - \hat{\tau}\right)^2} + \mathcal{N}(0, \sigma_2^2)\right)^2 + \Phi^{-1}(1-p_1)\sigma_2^2 \geqslant \hat{V}\right) \\
&\geqslant \mathbb{P}\left(\mathcal{N}(0, \sigma_2^2)^2 \leqslant \Phi^{-1}(1-p_1)^2\sigma_2^2\right) \\
&\geqslant \mathbb{P}\left(\mathcal{N}(0,1) \leqslant \Phi^{-1}(1-p_1)\right) \\
&\geqslant 1 - p_1\,,
\end{aligned}$$

by definition of Φ. Thus, with probability at least $1 - p_1$, we have that $\hat{V}_{\mathrm{DP}} \geqslant \hat{V} + \sigma_1^2$. Then, *in an oracle setting* where the estimated variance $\hat{V}$ is the actual variance $V^\star$ of the oracle estimator up to

[6] the CI for Equation (8) is obtained with $p = 0.05$ and $p_1 = 0.01$

$o(1/n)$ and under asymptotic $\sqrt{n}-$consistency assumptions, under the central limit theorem, with $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\Gamma}_i$ and $c = \Phi^{-1}(1 - p/2 + p_1/2)$:

$$
\begin{aligned}
\mathbb{P}\left(\tau \in \mathrm{CI}_p\right) &= \mathbb{P}\left(\hat{\tau} + \mathcal{N}(0, \sigma_1^2) \in \left[\tau - c\sqrt{\hat{V}_{\mathrm{DP}}}, \tau + c\sqrt{\hat{V}_{\mathrm{DP}}}\right]\right) \\
&\geqslant \mathbb{P}\left(\hat{\tau} + \mathcal{N}(0, \sigma_1^2) \in \left[\tau - c\sqrt{\hat{V} + \sigma_1^2}, \tau + c\sqrt{\hat{V} + \sigma_1^2}\right] \middle| \hat{V}_{\mathrm{DP}} \geqslant \hat{V} + \sigma_1^2\right) - p_1 \\
&\approx \mathbb{P}\left(\mathcal{N}(0, V^\star) + \mathcal{N}(0, \sigma_1^2) \in \left[\tau - c\sqrt{V^\star + \sigma_1^2}, \tau + c\sqrt{V^\star + \sigma_1^2}\middle| \hat{V}_{\mathrm{DP}} \geqslant \hat{V} + \sigma_1^2\right]\right) - p_1 \\
&= \mathbb{P}\left(\mathcal{N}(0, \sqrt{V^\star + \sigma_1^2}) \in \left[\tau - c\sqrt{V^\star + \sigma_1^2}, \tau + c\sqrt{V^\star + \sigma_1^2}\middle| \hat{V}_{\mathrm{DP}} \geqslant \hat{V} + \sigma_1^2\right]\right) - p_1 \\
&= 1 - 2\mathbb{P}\left(\mathcal{N}(0,1) \geqslant c\right) - p_1 \\
&= 1 - 2p/2 + 2p_1/2 - p_1 \\
&= 1 - p\,,
\end{aligned}
$$

where the first inequality uses the probabilistic bound above, and the $\approx$ holds up to a $o(1)$ under $\sqrt{n}-$consistency and exact oracle variance up to $o(1/n)$. These assumptions are classically implied by those of Theorem 2 in the case of our estimators [Wager, 2024].

B.2 More general aggregation steps

Remark 3. *The **aggregation step** can be made more general:*

$$
\begin{aligned}
\hat{\pi}(i) = &\Phi_\pi\left(\left\{\hat{\pi}^{(k)}\right\}, k(i)\right), \quad \hat{\mu}_a(i) = \Phi_\mu\left(\left\{\hat{\mu}_a^{(k)}\right\}, k(i)\right), \\
&1 - \tilde{\pi}(i) = \Phi_{1-\pi}\left(\left\{1 - \hat{\pi}^{(k)}\right\}, k(i)\right),
\end{aligned}
\tag{13}
$$

Examples of functions Φ used in Equation (13) are, for some models $(\hat{m}^{(k)})_{k\in[K]}$ (that could represent either $\hat{\pi}^{(k)}, \hat{\mu}_a^{(k)}$ or $1 - \hat{\pi}^{(k)}$):

1. ***Sampling.*** *For all $i \in [n]$, let*

$$
\Phi_\ell\left(\left\{\hat{m}^{(k)}\right\}, k(i)\right) = \hat{m}^{(\ell(i))}\,, \tag{14}
$$

   for $\ell(i) \in [K] \setminus \{k(i)\}$. For instance, $\ell(i) \sim \mathcal{U}([K] \setminus \{k(i)\})$ can be chosen.

2. ***Mean.*** *For $\mathcal{K} \subset [K]$, for all $i \in [n]$, let*

$$
\Phi_{\mathcal{K}}\left(\left\{\hat{m}^{(k)}\right\}, k(i)\right) = \frac{1}{\#\mathcal{K} - \mathbb{1}_{\{k(i)\in\mathcal{K}\}}} \sum_{k\in\mathcal{K}\setminus\{k(i)\}} \hat{m}^{(\ell)}\,. \tag{15}
$$

3. ***Harmonic mean.*** *For $\mathcal{K} \subset [K]$, if $m^{(k)}$ have positive values, for all $i \in [n]$, let*

$$
\tilde{\Phi}_{\mathcal{K}}\left(\left\{\hat{m}^{(k)}\right\}, k(i)\right) = \left(\frac{1}{\#\mathcal{K} - \mathbb{1}_{\{k(i)\in\mathcal{K}\}}} \sum_{k\in\mathcal{K}\setminus\{k(i)\}} \frac{1}{\hat{m}^{(\ell)}}\right)^{-1}. \tag{16}
$$

Other alternatives could be used, such as a median estimator (for its robustness properties), or more evolved ensemble methods to combine the models $m^{(k)}$.

*In this paper and in all our results, propensity scores use the harmonic mean aggregator while conditional outcomes use the mean aggregator. Motivations for this choice of aggregators stem from Lemma 3. Note that the sampling approach could be also used, for both propensity scores **and** conditional outcomes. We provide sensitivity analyses for this approach in the proofs in the appendices below.*

B.3 Unified sensitivity analysis

Proposition 2 (Sensitivity analysis)**.** *Let*

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\Gamma}_i \,,$$

where $\hat{\Gamma}_i = \Psi\left((X_i, A_i, Y_i), k(i), \left\{\hat{\pi}^{(k)}\right\}, \left\{\hat{\mu}_a^{(k)}\right\}, \left\{\tilde{\pi}^{(k)}\right\}\right)$*. Assume that for all* $i \in [n]$*, for all datasets* $\mathcal{D} \sim_i \mathcal{D}'$*, we have that:*

1. $|\hat{\Gamma}_i - \hat{\Gamma}'_i| \leqslant \Delta_{=}$,
2. *For all* $j \in \mathcal{I}_{k(i)} \setminus \{i\}$, $\hat{\Gamma}_j = \hat{\Gamma}'_j$,
3. $\frac{1}{n}\sum_{j \notin \mathcal{I}_{k(i)}} |\hat{\Gamma}_i - \hat{\Gamma}'_i| \leqslant \frac{\Delta_{\neq}}{K}$,

where $(\hat{\Gamma}'_j)$ *corresponds to the quantities when models are trained and evaluated on* $\mathcal{D}'$*. Then, we have that:*

$$\sup_{\mathcal{D}\sim\mathcal{D}'} |\hat{\tau} - \hat{\tau}'| \leqslant \frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K} \,,$$

so that for all $\sigma > 0$*, the estimator* $\hat{\tau}_{\mathrm{DP}} := \hat{\tau} + \mathcal{N}(0, \sigma^2)$ *is* $\zeta-$*GDP with* $\zeta^2 = \frac{\left(\frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K}\right)^2}{\sigma^2}$*.*

Under the same assumptions and if furthemore $\sup_i |\hat{\Gamma}_i| \leqslant M$*:*

$$\sup_{\mathcal{D}\sim\mathcal{D}'} |\sqrt{\hat{V}} - \sqrt{\hat{V}'}| \leqslant \sqrt{\frac{2}{(n-1)}}\left(2\sqrt{M}\left(\frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K}\right)^{1/2} + \frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K}\right) \,,$$

where $\hat{V} := \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\Gamma}_i - \hat{\tau}\right)^2$*. Thus, for* $\gamma_1, \gamma_2 \sim \mathcal{N}(0,1)$ *independent and* $\sigma_1, \sigma_2 > 0$*, writing* $\hat{V}_{\mathrm{DP}} = (\sqrt{\hat{V}} + \sigma_2\gamma_2)^2$ *and* $\hat{\tau}_{\mathrm{DP}} := \hat{\tau} + \mathcal{N}(0, \sigma^2)$*, we have that outputting the confidence interval*

$$\mathcal{A}(\mathcal{D}) = \left[\hat{\tau}_{\mathrm{DP}} - 1,96\sqrt{\hat{V}_{\mathrm{DP}}}, +, \hat{\tau}_{\mathrm{DP}} + 1,96\sqrt{\hat{V}_{\mathrm{DP}}}\right] \,,$$

is $\zeta-$*GDP, with* $\zeta^2 = \frac{\left(\frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K}\right)^2}{\sigma_1^2} + \frac{4\left(\frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K}\right)^2}{\sigma_2^2}\frac{n}{n-1}$*.*

Proof. We begin by proving that under the listed assumptions, we have:

$$\sup_{\mathcal{D}\sim\mathcal{D}'} |\hat{\tau} - \hat{\tau}'| \leqslant \frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K} \,.$$

Let $\mathcal{D} \sim_i \mathcal{D}'$. We have:

$$\begin{aligned}
|\hat{\tau} - \hat{\tau}'| &= \left|\frac{1}{n}\sum_{j=1}^{n}\hat{\Gamma}_j - \hat{\Gamma}'_j\right| \\
&\leqslant \frac{1}{n}\sum_{j=1}^{n}\left|\hat{\Gamma}_j - \hat{\Gamma}'_j\right| \\
&\leqslant \frac{\left|\hat{\Gamma}_i - \hat{\Gamma}'_i\right|}{n} + \frac{1}{n}\sum_{j\in\mathcal{I}_{k(i)}\setminus\{i\}}^{n}\left|\hat{\Gamma}_j - \hat{\Gamma}'_j\right| + \frac{1}{n}\sum_{j\notin\mathcal{I}_{k(i)}}^{n}\left|\hat{\Gamma}_j - \hat{\Gamma}'_j\right| \\
&\leqslant \frac{\Delta_{=}}{n} + \frac{1}{n}\sum_{j\in\mathcal{I}_{k(i)}\setminus\{i\}}^{n} 0 + \frac{\Delta_{\neq}}{K} \\
&\leqslant \frac{\Delta_{=}}{n} + \frac{\Delta_{\neq}}{K} \,,
\end{aligned}$$

by a direct application of our assumptions. The DP guarantees are then given by applying Lemma 1.

We now show that

$$\sup_{\mathcal{D}\sim\mathcal{D}'} |\sqrt{\hat{V}} - \sqrt{\hat{V}'}| \leqslant \frac{2}{n-1}\left(\frac{\Delta_=}{n} + \frac{\Delta_\neq}{K}\right).$$

We have, for $\mathcal{D} \sim_i \mathcal{D}'$:

$$\begin{aligned}
\left|\sqrt{\hat{V}} - \sqrt{\hat{V}'}\right| &= \sqrt{\frac{1}{n(n-1)}}\left|\sqrt{\sum_{i=1}^n \left(\hat{\Gamma}_i - \hat{\tau}\right)^2} - \sqrt{\sum_{i=1}^n \left(\hat{\Gamma}'_i - \hat{\tau}'\right)^2}\right| \\
&\leqslant \sqrt{\frac{1}{n(n-1)}}\sqrt{\sum_{i=1}^n \left(\hat{\Gamma}_i - \hat{\tau} - \left\{\hat{\Gamma}'_i - \hat{\tau}'\right\}\right)^2} \\
&\leqslant \sqrt{\frac{1}{n(n-1)}}\sqrt{2\sum_{i=1}^n \left(\hat{\Gamma}_i - \hat{\Gamma}'_i\right)^2 + 2\sum_{i=1}^n (\hat{\tau} - \hat{\tau}')^2} \\
&\leqslant \sqrt{\frac{2}{n(n-1)}}\left(2\sqrt{M}\sqrt{\sum_{i=1}^n \left|\hat{\Gamma}_i - \hat{\Gamma}'_i\right|} + \sqrt{n}|\hat{\tau} - \hat{\tau}'|\right) \\
&\leqslant \sqrt{\frac{2}{n(n-1)}}\left(2\sqrt{M}\sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^n \left|\hat{\Gamma}_i - \hat{\Gamma}'_i\right|} + \sqrt{n}|\hat{\tau} - \hat{\tau}'|\right) \\
&\leqslant \sqrt{\frac{2}{(n-1)}}\left(2\sqrt{M}\left(\frac{\Delta_=}{n} + \frac{\Delta_\neq}{K}\right)^{1/2} + \frac{\Delta_=}{n} + \frac{\Delta_\neq}{K}\right),
\end{aligned}$$

using the same arguments as above. The DP guarantees are then given by applying Lemma 1.

□

B.4 Sensitivity of our aggregators

Lemma 3. *For models $\{m^{(k)}\}$ trained on dataset $\mathcal{D}$ as in Section 4.1 and $\hat{m} : [n] \to \mathbb{R}$ defined as $\hat{m}(i) = \Phi\left(\{m^{(k)}\}, k(i)\right)$ as in Equation (5). Let $\mathcal{D} \sim_i \mathcal{D}'$ be two adjacent datasets, and denote as $\hat{m}'$ the global model corresponding to training on $\mathcal{D}'$. We have the following properties for the examples detailed in Remark 3.*

1. ***Sampling.*** *Assume that $m^{(k)}$ lie in a space of diameter $M > 0$. If $\Phi = \Phi_\ell$, for $\ell : [n] \to [K]$ such that $\ell(i) \neq k(i)$, then we have, for all $j \in [n] \setminus \mathcal{I}_{k(i)}$:*

$$|\hat{m}(j) - \hat{m}'(j)| \leqslant M\mathbb{1}_{\{j\in\mathcal{J}_{k(i)}\}},$$

where $\mathcal{J}_k = \{i : \ell(i) = k\}$, and $\hat{m}(j) = \hat{m}'(j)$ for $j \in \mathcal{I}_{k(i)}$.

2. ***Mean.*** *Assume that $m^{(k)}$ lie in a space of diameter $M > 0$. If $\Phi = \Phi_\mathcal{K}$ for $\mathcal{K} \subset [K]$, then we have, for all $j \in [n] \setminus \mathcal{I}_{k(i)}$:*

$$|\hat{m}(j) - \hat{m}'(j)| \leqslant \frac{M}{\#\mathcal{K} - \mathbb{1}_{\{k(i)\in\mathcal{K}\}}},$$

and $\hat{m}(j) = \hat{m}'(j)$ for $j \in \mathcal{I}_{k(i)}$.

3. ***Harmonic mean (1).*** *Assume that $m^{(k)} > 0$ and that $1/m^{(k)}$ are uniformly bounded by a constant $M > 0$. If $\Phi = \tilde{\Phi}_\mathcal{K}$ for $\mathcal{K} \subset [K]$, then we have, for all $j \in [n] \setminus \mathcal{I}_{k(i)}$:*

$$\left|\frac{1}{\hat{m}(j)} - \frac{1}{\hat{m}'(j)}\right| \leqslant \frac{M}{\#\mathcal{K} - \mathbb{1}_{\{k(i)\in\mathcal{K}\}}},$$

and $\hat{m}(j) = \hat{m}'(j)$ for $j \in \mathcal{I}_{k(i)}$.

4. ***Harmonic mean (2).*** *Assume that $0 < m^{(k)}$ and $1/m^{(k)}$ are uniformly bounded by a constant $M > 0$. If $\Phi = \tilde{\Phi}_{\mathcal{K}}$ for $\mathcal{K} \subset [K]$, then we have, for all $j \in [n] \setminus \mathcal{I}_{k(i)}$:*

$$|\hat{m}(j) - \hat{m}'(j)| \leqslant \frac{M^3}{\#\mathcal{K} - \mathbb{1}_{\{k(i) \in \mathcal{K}\}}},$$

and $\hat{m}(j) = \hat{m}'(j)$ for $j \in \mathcal{I}_{k(i)}$.

The harmonic mean in its second form above is thus way less competitive and thus should only be used for propensity scores.

Proof. The proof of the first, second and third points are straightforward. For the harmonic mean, let $j \in [n] \setminus \mathcal{I}_{k(i)}$. We have that

$$\hat{m}(j) = \frac{1}{\frac{1}{K}\sum_{k \in \mathcal{K} \setminus \{k(i)\}} \hat{m}^{(k)}(X_j)^{-1} + \frac{\hat{m}^{(k(i))}(X_j)^{-1}}{K}},$$

and

$$\hat{m}'(j) = \frac{1}{\frac{1}{K}\sum_{k \in \mathcal{K} \setminus \{k(i)\}} \hat{m}^{(k)}(X_j)^{-1} + \frac{\hat{m}'^{(k(i))}(X_j)^{-1}}{K}}.$$

Let $C = \frac{1}{K}\sum_{k \in \mathcal{K} \setminus \{k(i)\}} \hat{m}^{(k)}(X_j)^{-1} \geqslant \frac{1}{M}$, and note that $f : x > 0 \mapsto \frac{1}{C + \frac{x}{K}}$ is convex. Thus, for $0 < x_1 < x_2 \leqslant M$,

$$\begin{aligned}
0 \leqslant f(x_1) - f(x_2) &\leqslant f'(x_1)(x_1 - x_2) \\
&\leqslant |f'(0)|(x_2 - x_1) \\
&= \frac{x_2 - x_1}{KC^2} \\
&\leqslant \frac{M^2(x_2 - x_1)}{K} \\
&\leqslant \frac{M^3}{K}.
\end{aligned}$$

□

C Analysis of the DP-G-Formula estimator

Lemma 4 (G-Formula)**.** *For $\ell : [n] \to [K]$ such that for all $i \in [n]$ we have $\ell(i) \in [K] \setminus \{k(i)\}$, let*

$$\Phi_\mu \in \left\{\Phi_{[K]}, \Phi_\ell\right\},$$

as defined in Equation (14). Let $\mathcal{J}_k := \{i \in [n] : \ell(i) = k\}$ for $k \in [K]$. Then, for the DP-G-Formula estimator, the assumptions of Proposition 2 are verified with:

$$\Delta_= = 4B_\mu,$$

and

$$\Delta_{\neq} = 4B_\mu \times \frac{K}{K-1}.$$

if $\Phi_\mu = \Phi_{[K]}$, or

$$\Delta_{\neq} = 4B_\mu \times \frac{K \sup_{k \in [K]} \#\mathcal{J}_k}{n},$$

if $\Phi_\mu = \Phi_\ell$.

Proof. Let $\mathcal{D} \sim \mathcal{D}'$ and assume without loss of generality that $(X_i, A_i, Y_i) \neq (X_i, A_i, Y_i)$ for $i = 1$ and that $k(1) = 1$ (i.e., $1 \in \mathcal{I}_1$). Denote by $\hat{\mu}_t^{'(k)}, \hat{\mu}'_t, \hat{\pi}', \hat{\pi}^{'(k)}$ the models learned on $\mathcal{D}'$. Since $\mathcal{D}$ and $\mathcal{D}'$ differ only on the datapoint $i = 1$ and that $1 \notin \mathcal{I}_k$ for $k \geqslant 2$, we have that $\hat{\mu}_a^{'(k)} = \hat{\mu}_a^{(k)}$ for $k \geqslant 2$.

First, for all $j \in \mathcal{I}_1 \setminus \{1\}$, we have $(X_i, A_i, T_i) = (X'_i, A'_i, T'_i)$ and $\hat{\mu}_a(i) = \hat{\mu}'_a(i)$, so that $\hat{\Gamma}_i = \hat{\Gamma}'_i$. Then, for $i = 1$:

$$\begin{aligned}|\hat{\Gamma}_1 - \hat{\Gamma}'_1| &\leqslant |\hat{\mu}_1(1) - \hat{\mu}_0(1) - \hat{\mu}'_1(1) + \hat{\mu}'_0(1)| \\ &\leqslant |\hat{\mu}_1(1)| + |\hat{\mu}_0(1)| + |\hat{\mu}'_1(1)| + |\hat{\mu}'_0(1)| \\ &\leqslant 4B\,,\end{aligned}$$

where we used Assumption 4. This expression is independent of $i = 1$, so we can conclude that

$$\Delta_= = 4B_\mu\,.$$

We then find a bound for $\Delta_\neq$.

Case 1: $\Phi_\mu = \Phi_\ell$. Let $\mathcal{J}_k = \{i \in [n] \text{ s.t. } \ell(i) = k\}$. We have $\mathcal{J}_1 \subset [n] \setminus \mathcal{I}_1$. For all $j \notin \mathcal{J}_1$, we have $\hat{\mu}_a(j) = \hat{\mu}'_a(j)$. For $i \in \mathcal{J}_1$, using Lemma 3 and Assumption 4,

$$|\hat{\Gamma}_i - \hat{\Gamma}'_i| \leqslant 4B_\mu\,.$$

Thus, by summing over all $j \notin \mathcal{I}_1$,

$$\Delta_\neq = 4B_\mu \times \frac{K \sup_{k\in[K]} \#\mathcal{J}_k}{n}\,.$$

Case 2: $\Phi_\mu = \Phi_{[K]}$. Using Lemma 3 and Assumption 4, for all $i \notin \mathcal{I}_1$, we have:

$$|\hat{\Gamma}_i - \hat{\Gamma}'_i| \leqslant \frac{4B_\mu}{K-1}\,.$$

By summing, we thus obtain:

$$\Delta_\neq = \frac{4B_\mu K}{K-1}\,.$$

□

D Analysis of the DP-IPW estimator

Lemma 5 (IPW). *For $\ell : [n] \to [K]$ such that for all $i \in [n]$ we have $\ell(i) \in [K] \setminus \{k(i)\}$, let*

$$\Phi_\pi,\ \Phi_{1-\pi} \in \left\{\tilde{\Phi}_{[K]}, \Phi_\ell\right\},$$

as defined in Equation (14). Let $\mathcal{J}_k := \{i \in [n] : \ell(i) = k\}$ for $k \in [K]$. Then, for the DP-IPW estimator, the assumptions of Proposition 2 are verified with:

$$\Delta_= = 2B_\mu B_\pi\,,$$

and

$$\Delta_\neq = B_\mu B_\pi \times \frac{K}{K-1}\,.$$

if $\Phi_\pi,\ \Phi_{1-\pi} = \tilde{\Phi}_{[K]}$, or

$$\Delta_\neq = B_\mu B_\pi \times \frac{K \sup_{k\in[K]} \#\mathcal{J}_k}{n}\,,$$

if $\Phi_\pi,\ \Phi_{1-\pi} = \Phi_\ell$.

Proof. Let $\mathcal{D} \sim \mathcal{D}'$ and assume without loss of generality that $(X_i, A_i, Y_i) \neq (X_i, A_i, Y_i)$ for $i = 1$ and that $k(1) = 1$ (i.e., $1 \in \mathcal{I}_1$). Denote by $\hat{\mu}_t^{'(k)}, \hat{\mu}'_t, \hat{\pi}', \hat{\pi}^{'(k)}$ the models learned on $\mathcal{D}'$. Since $\mathcal{D}$ and $\mathcal{D}'$ differ only on the datapoint $i = 1$ and that $1 \notin \mathcal{I}_k$ for $k \geqslant 2$, we have that $\hat{\pi}^{'(k)} = \hat{\pi}^{(k)}$ for $k \geqslant 2$.

First, for all $j \in \mathcal{I}_1 \setminus \{1\}$, we have $(X_i, A_i, T_i) = (X'_i, A'_i, T'_i)$ and $\hat{\pi}(i) = \hat{\pi}'(i)$ and $\tilde{\pi}(i) = \tilde{\pi}'(i)$, so that $\hat{\Gamma}_i = \hat{\Gamma}'_i$.

Then, for $i = 1$, since $\hat{\pi}(1) = \hat{\pi}'(1)$ and $\tilde{\pi}(1) = \tilde{\pi}'(1)$:

$$|\hat{\Gamma}_1 - \hat{\Gamma}'_1| = |\frac{A_1 Y_1}{\hat{\pi}(1)} - \frac{A'_1 Y'_1}{\hat{\pi}'(1)} - \frac{(1-A_1)Y_1}{1-\tilde{\pi}(1)} + \frac{(1-A'_1)Y'_1}{1-\tilde{\pi}'(1)}|$$
$$\leqslant B_\mu(A_1 + 1 - A_1 + A'_1 + 1 - A'_1)B_\pi \,,$$

where we used Assumption 4. This expression is independent of $i = 1$, so we can conclude that

$$\Delta_= = 2B_\mu B_\pi \,.$$

We then find a bound for $\Delta_\neq$.

Case 1: $\Phi_\mu = \Phi_\ell$. Let $\mathcal{J}_k = \{i \in [n] \text{ s.t. } \ell(i) = k\}$. We have $\mathcal{J}_1 \subset [n] \setminus \mathcal{I}_1$. For all $j \notin \mathcal{J}_1$, we have $\hat{\pi}(j) = \hat{\pi}'(j)$ and $\tilde{\pi}(j) = \tilde{\pi}'(j)$. Thus, for $i \in \mathcal{J}_1$, using Lemma 3 and Assumption 4, since $A_i = A'_i$ and $Y_i = Y'_i$:

$$|\hat{\Gamma}_i - \hat{\Gamma}'_i| = |A_i Y_i(1/\hat{\pi}(i) - 1/\hat{\pi}'(i)) + (1-A_i)Y_i(1/(1-\hat{\pi}(i)) - 1/(1-\hat{\pi}'(i))| \leqslant B_\mu B_\pi \,.$$

Thus, by summing over all $j \notin \mathcal{I}_1$,

$$\Delta_\neq = B_\mu B_\pi \times \frac{K \sup_{k\in[K]} \#\mathcal{J}_k}{n} \,.$$

Case 2: $\Phi_\mu \in \left\{\Phi_{[K]}, \tilde{\Phi}_{[K]}\right\}$. Using Lemma 3 and Assumption 4, for all $i \notin \mathcal{I}_1$, we have:

$$|\hat{\Gamma}_i - \hat{\Gamma}'_i| \leqslant B_\mu B_\pi \times \frac{1}{K-1} \,.$$

By summing, we thus obtain:

$$\Delta_\neq = B_\mu B_\pi \times \frac{K}{K-1} \,.$$

□

E Analysis of the DP-AIPW estimator

Lemma 6 (AIPW - Sampling)**.** *For $\ell : [n] \to [K]$ such that for all $i \in [n]$ we have $\ell(i) \in [K]\setminus\{k(i)\}$, let*

$$\Phi_\pi = \Phi_{1-\pi} = \Phi_\mu = \Phi_\ell \,,$$

as defined in Equation (14). Let $\mathcal{J}_k := \{i \in [n] : \ell(i) = k\}$ for $k \in [K]$. Then, for the DP-AIPW estimator, the assumptions of Proposition 2 are verified with:

$$\Delta_= = 2B_\mu(2 + B_\pi) \,,$$

and

$$\Delta_\neq = (4B_\mu + 3B_\mu B_\pi) \times \frac{K \sup_{k\in[K]} \#\mathcal{J}_k}{n} \,.$$

Proof. Let $\mathcal{D} \sim \mathcal{D}'$ and assume without loss of generality that $(X_i, A_i, Y_i) \neq (X_i, A_i, Y_i)$ for $i = 1$ and that $k(1) = 1$ (i.e., $1 \in \mathcal{I}_1$). Denote by $\hat{\mu}_t^{'(k)}, \hat{\mu}'_t, \hat{\pi}', \hat{\pi}^{'(k)}$ the models learned on $\mathcal{D}'$. Since $\mathcal{D}$ and $\mathcal{D}'$ differ only on the datapoint $i = 1$ and that $1 \notin \mathcal{I}_k$ for $k \geqslant 2$, we have that $\hat{\mu}_a^{'(k)} = \hat{\mu}_a^{(k)}$ and $\hat{\pi}^{'(k)} = \hat{\pi}^{(k)}$ for $k \geqslant 2$.

First, for all $j \in \mathcal{I}_1 \setminus \{1\}$, we have $(X_i, A_i, T_i) = (X'_i, A'_i, T'_i)$ and $\hat{\mu}_a(i) = \hat{\mu}'_a(i)$, so that $\hat{\Gamma}_i = \hat{\Gamma}'_i$.

Then, for $i = 1$:

$$\begin{aligned}
&|\hat{\Gamma}_1 - \hat{\Gamma}'_1| \leqslant |\hat{\mu}_1(1) - \hat{\mu}_0(1) - \hat{\mu}'_1(1) + \hat{\mu}'_0(1)| \\
&+ \left|\frac{A_i}{\hat{\pi}(i)}(Y_i - \hat{\mu}_1(i)) + \frac{1-A_i}{1-\hat{\pi}(i)}(Y_i - \hat{\mu}_0(i)) - \left\{\frac{A_i}{\hat{\pi}'(i)}(Y_i - \hat{\mu}'_1(i)) + \frac{1-A_i}{1-\hat{\pi}'(i)}(Y_i - \hat{\mu}'_0(i))\right\}\right| \\
&\leqslant |\hat{\mu}_1(1)| + |\hat{\mu}_0(1)| + |\hat{\mu}'_1(1)| + |\hat{\mu}'_0(1)| \\
&+ \left|\frac{A_1}{\hat{\pi}(1)}(Y_1 - \hat{\mu}_1(1))\right| + \left|\frac{1-A_1}{1-\hat{\pi}(1)}(Y_1 - \hat{\mu}_0(1))\right| + \left|\frac{A'_1}{\hat{\pi}'(1)}(Y'_1 - \hat{\mu}'_1(1))\right| + \left|\frac{1-A'_1}{1-\hat{\pi}'(1)}(Y'_1 - \hat{\mu}'_0(1))\right| \\
&\leqslant 4B_\mu + 2A_1 B_\mu B_\pi + 2(1-A_1)B_\mu B_\pi + 2A'_1 B_\mu B_\pi + 2(1-A'_1)B_\mu B_\pi \\
&\leqslant 2B_\mu(2 + B_\pi) \,,
\end{aligned}$$

where we used Assumption 4 to bound $1/\hat{\pi}'(1), 1/\hat{\pi}(1), 1/(1-\hat{\pi}'(1)), 1/(1-\hat{\pi}(1))$. This expression is independent of $i=1$, so we can conclude that

$$\Delta_{=} = 2B_\mu(2+B_\pi)\,.$$

We then find a bound for $\Delta_{\neq}$. Let $\mathcal{J}_k = \{i \in [n] \text{ s.t. } \ell(i) = k\}$. We have $\mathcal{J}_1 \subset [n] \setminus \mathcal{I}_1$. For all $j \notin \mathcal{J}_1$, we have $\hat{\pi}(j) = \hat{\pi}'(j)$, $\tilde{\pi}(j) = \tilde{\pi}'(j)$, and $\hat{\mu}_a(j) = \hat{\mu}'_a(j)$. For $i \in \mathcal{J}_1$,

$$\begin{aligned}
&|\hat{\Gamma}_i - \hat{\Gamma}'_i| \\
&= \left| \frac{A_i}{\hat{\pi}(i)}(Y_i - \hat{\mu}_1(i)) + \frac{1-A_i}{1-\hat{\pi}(i)}(Y_i - \hat{\mu}_0(i)) - \left\{ \frac{A_i}{\hat{\pi}'(i)}(Y_i - \hat{\mu}_1(i)) + \frac{1-A_i}{1-\hat{\pi}'(i)}(Y_i - \hat{\mu}_0(i)) \right\} \right| \\
&\leqslant A_i \left| \frac{Y_i - \hat{\mu}_1(i)}{\hat{\pi}(i)} - \frac{Y_i - \hat{\mu}_1(i)}{\hat{\pi}'(i)} \right| + (1-A_i) \left| \frac{Y_i - \hat{\mu}_0(i)}{1-\hat{\pi}(i)} - \frac{Y_i - \hat{\mu}_0(i)}{1-\hat{\pi}'(i)} \right| \\
&\leqslant 2B_\mu A_i \left| \frac{1}{\hat{\pi}(i)} - \frac{1}{\hat{\pi}'(i)} \right| + 2B_\mu(1-A_i) \left| \frac{\hat{\mu}_0(i)}{\hat{\pi}(i)} - \frac{\hat{\mu}'_0(i)}{\hat{\pi}'(i)} \right| \\
&= 2B_\mu B_\pi\,.
\end{aligned}$$

Then, for $j \in \mathcal{J}'_1$, we have $\hat{\pi}(i) = \hat{\pi}'(i)$, and thus:

$$\begin{aligned}
&|\hat{\Gamma}_i - \hat{\Gamma}'_i| \\
&= \Bigg| \hat{\mu}_1(i) - \hat{\mu}_0(i) + \frac{A_i}{\hat{\pi}(i)}(Y_i - \hat{\mu}_1(i)) + \frac{1-A_i}{1-\hat{\pi}(i)}(Y_i - \hat{\mu}_0(i)) \\
&\quad - \left\{ \hat{\mu}'_1(i) - \hat{\mu}'_0(i) + \frac{A_i}{\hat{\pi}'(i)}(Y_i - \hat{\mu}'_1(i)) + \frac{1-A_i}{1-\hat{\pi}'(i)}(Y_i - \hat{\mu}'_0(i)) \right\} \Bigg| \\
&\leqslant 4B_\mu + A_i|Y_i| \left| \frac{1}{\hat{\pi}(i)} - \frac{1}{\hat{\pi}'(i)} \right| + A_i \left| \frac{\hat{\mu}_1(i)}{\hat{\pi}(i)} - \frac{\hat{\mu}'_1(i)}{\hat{\pi}'(i)} \right| + (1-A_i)|Y_i| \left| \frac{1}{\hat{\pi}(i)} - \frac{1}{\hat{\pi}'(i)} \right| + (1-A_i) \left| \frac{\hat{\mu}_0(i)}{\hat{\pi}(i)} - \frac{\hat{\mu}'_0(i)}{\hat{\pi}'(i)} \right| \\
&\leqslant 4B_\mu + (B_\mu B_\pi + 2B_\mu B_\pi)(A_i + 1 - A_i) \\
&= 4B_\mu + 3B_\mu B_\pi\,.
\end{aligned}$$

Thus, by summing over all $j \notin \mathcal{I}_1$,

$$\Delta_{\neq} = (4B_\mu + 3B_\mu B_\pi) \times \frac{K \sup_{k \in [K]} \#\mathcal{J}_k}{n}\,.$$

□

Lemma 7 (AIPW - Complete Means)**.** *Let*

$$\Phi_\pi, \Phi_{1-\pi}, \Phi_\mu = (\tilde{\Phi}_{[K]}, \tilde{\Phi}_{[K]}, \Phi_{[K]})\,,$$

as defined in Equations (15) and (16): the ensembling is done via mean or harmonic mean over all samples (except those in the fold over which the sample models are evaluated on is in). Then, for the DP-AIPW estimator, the assumptions of Proposition 2 are verified with:

$$\Delta_{=} = 4B_\mu(1+B_\pi)\,,$$

and

$$\Delta_{\neq} = \frac{4B_\mu K}{K-1}\,(1+B_\pi)\,.$$

Proof. Let $\mathcal{D} \sim \mathcal{D}'$ and assume without loss of generality that $(X_i, A_i, Y_i) \neq (X_i, A_i, Y_i)$ for $i=1$ and that $k(1)=1$ (i.e., $1 \in \mathcal{I}_1$). Denote by $\hat{\mu}_t^{'(k)}, \hat{\mu}'_t, \hat{\pi}', \hat{\pi}^{'(k)}$ the models learned on $\mathcal{D}'$. Since $\mathcal{D}$ and $\mathcal{D}'$ differ only on the datapoint $i=1$ and that $1 \notin \mathcal{I}_k$ for $k \geqslant 2$, we have that $\hat{\mu}_a^{'(k)} = \hat{\mu}_a^{(k)}$ and $\hat{\pi}^{'(k)} = \hat{\pi}^{(k)}$ for $k \geqslant 2$.

First, for all $j \in \mathcal{I}_1 \setminus \{1\}$, we have $(X_i, A_i, T_i) = (X'_i, A'_i, T'_i)$ and $\hat{\mu}_a(i) = \hat{\mu}'_a(i)$ (they are computed using only models trained on folds $k \geqslant 2$ that are thus unchanged), so that $\hat{\Gamma}_i = \hat{\Gamma}'_i$.

Then, for $i = 1$, we also have that the models are unchanged (trained on folds $k \geqslant 2$), so that:

$$
\begin{aligned}
&|\hat{\Gamma}_1 - \hat{\Gamma}'_1| \leqslant |\hat{\mu}_1(1) - \hat{\mu}_0(1) - \hat{\mu}'_1(1) + \hat{\mu}'_0(1)| \\
&+ \left| \frac{A_i}{\hat{\pi}(i)}(Y_i - \hat{\mu}_1(i)) + \frac{1-A_i}{1-\hat{\pi}(i)}(Y_i - \hat{\mu}_0(i)) - \left\{ \frac{A_i}{\hat{\pi}'(i)}(Y_i - \hat{\mu}'_1(i)) + \frac{1-A_i}{1-\hat{\pi}'(i)}(Y_i - \hat{\mu}'_0(i)) \right\} \right| \\
&\leqslant |\hat{\mu}_1(1)| + |\hat{\mu}_0(1)| + |\hat{\mu}'_1(1)| + |\hat{\mu}'_0(1)| \\
&+ \left| \frac{A_1}{\hat{\pi}(1)}(Y_1 - \hat{\mu}_1(1)) \right| + \left| \frac{1-A_1}{1-\hat{\pi}(1)}(Y_1 - \hat{\mu}_0(1)) \right| + \left| \frac{A'_1}{\hat{\pi}'(1)}(Y'_1 - \hat{\mu}'_1(1)) \right| + \left| \frac{1-A'_1}{1-\hat{\pi}'(1)}(Y'_1 - \hat{\mu}'_0(1)) \right| \\
&\leqslant 4B_\mu + 2A_1 B_\mu B_\pi + 2(1-A_1)B_\mu B_\pi + 2A'_1 B_\mu B_\pi + 2(1-A'_1)B_\mu B_\pi \\
&\leqslant 4B_\mu(1 + B_\pi)\,,
\end{aligned}
$$

where we used Assumption 4 to bound $1/\hat{\pi}'(1), 1/\hat{\pi}(1), 1/(1-\hat{\pi}'(1)), 1/(1-\hat{\pi}(1))$/ This expression is independent of $i = 1$, so we can conclude that

$$\Delta_= = 4B_\mu(1 + B_\pi)\,.$$

Now using Lemma 3 and Assumption 4, we have that for $i \in [n] \setminus \mathcal{I}_1$,

$$
\begin{gathered}
|\hat{\pi}(i)^{-1} - \hat{\pi}'(i)^{-1}| \leqslant B_\pi/(K-1)\,, \quad |(1-\tilde{\pi}(i))^{-1} - (1-\tilde{\pi}'(i))^{-1}| \leqslant B_\pi/(K-1)\,, \\
|\hat{\mu}_a(i) - \hat{\mu}'_a(i)| \leqslant 2B_\mu/(K-1)\,.
\end{gathered}
$$

Thus, for $i \notin \mathcal{I}_1$,

$$
\begin{aligned}
|\hat{\mu}_1(i) - \hat{\mu}_0(i) - \{\hat{\mu}'_1(i) - \hat{\mu}'_0(i)\}| &\leqslant |\hat{\mu}_1(i) - \hat{\mu}'_1(i)| + |\hat{\mu}_0(i) - \hat{\mu}'_0(i)| \\
&\leqslant \frac{4B_\mu}{K-1}\,,
\end{aligned}
$$

and

$$
\begin{aligned}
\left| \frac{A_i}{\hat{\pi}(i)}(Y_i - \hat{\mu}_1(i)) - \frac{A_i}{\hat{\pi}'(i)}(Y_i - \hat{\mu}'_1(i)) \right| &\leqslant A_i \left| \frac{\hat{\mu}(i) - \hat{\mu}'(i)}{\hat{\pi}(i)} \right| + A_i \left| (Y_i - \hat{\mu}'(i)) \left(\frac{1}{\hat{\pi}(i)} - \frac{1}{\hat{\pi}'(i)} \right) \right| \\
&\leqslant \frac{4B_\mu B_\pi A_i}{K-1}\,.
\end{aligned}
$$

Similarly,

$$\left| \frac{1-A_i}{1-\hat{\pi}(i)}(Y_i - \hat{\mu}_0(i)) - \frac{1-A_i}{1-\hat{\pi}'(i)}(Y_i - \hat{\mu}'_0(i)) \right| \leqslant \frac{4B_\mu B_\pi(1-A_i)}{K-1}\,.$$

Thus,

$$|\hat{\Gamma}_i - \hat{\Gamma}'_i| \leqslant \frac{4B_\mu}{K-1}(1 + B_\pi)\,.$$

We can thus choose:

$$\Delta_{\neq} = \frac{4B_\mu K}{K-1}(1 + B_\pi)\,.$$

□

F Experiments

F.1 Experimental details

We give details on the setup used in the experiments of Section 4.

In Figure 1 the subscript for our estimators denote the estimator for the nuisance functions. IPW_{log} and IPW_{tree} estimates the propensity score with a logistic model and decision tree, respectively. G_{lin} and G_{tree} estimate the potential outcomes respectively with a linear regression model or a decision tree. $AIPW_{miss.}$ is misspecified for both nuisance functions and uses a logistic model for the propensity score and linear regression for the potential outcomes. $AIPW_{correct}$ is correctly specified by using a decision tree for both nuisance estimators. $AIPW_{log}$ and $AIPW_{lin}$ are partially misspecified by using a logistic model for the propensity score or a linear regression for the potential outcomes, respectively. In Figure 2, $AIPW_{log}$ uses a logistic model both for the propensity score and the potential outcomes.

Well-specified setting with low overlap. We generate data with $n = 5000$ and $d = 1$, where the propensity score is defined as $\pi(X_i) = \text{clip}_{[0.004,0.996]}(\text{expit}(-0.2+6X_{i,1}))$. The observed outcome follows a linear regression model where $Y_i = -0.05 + 0.225X_{i,1} + 0.1A_i + e$, with $e \sim \mathcal{N}(0, 0.01)$.

Misspecified setting. Here, propensity scores and outcome responses do not follow linear models but are instead based on splitting the feature space into four different regions for both propensities and outcome responses. These regions are modeled by two different decision trees with overlapping regions, where samples that are more likely to receive treatment have worse outcomes—a setup that could occur in medical settings, where practitioners might prescribe particular drugs based on 2 symptoms. We consider a simple setup with $d = 2$, $n = 250000$, $B_\pi = 1/0.2$, $B_\mu = 1$. The outcome is set as $Y_i = \gamma(X_i) + 0.2A_i + e$, where $e \sim \mathcal{N}(0, 0.025)$. The propensity score and $\gamma(X_i)$ are as follows:

$$\pi(X_i) = \begin{cases} 0.75 & \text{if } X_{i,1} > 0.1 \text{ and } X_{i,2} > 0. \\ 0.6 & \text{if } X_{i,1} \leqslant 0.1 \text{ and } X_{i,2} > 0. \\ 0.25 & \text{if } X_{i,1} < -0.05 \text{ and } X_{i,2} \leqslant 0. \\ 0.5 & \text{if } X_{i,1} \geqslant -0.05 \text{ and } X_{i,2} \leqslant 0. \end{cases} \quad \gamma(X_i) = \begin{cases} -0.7 & \text{if } X_{i,1} > 0 \text{ and } X_{i,2} > 0. \\ 0.1 & \text{if } X_{i,1} > 0 \text{ and } X_{i,2} \leqslant 0. \\ -0.4 & \text{if } X_{i,1} \leqslant 0 \text{ and } X_{i,2} > 0.05. \\ 0.6 & \text{if } X_{i,1} \leqslant 0 \text{ and } X_{i,2} \leqslant 0.05. \end{cases}$$

F.2 Additional results

Well-specified setting with good overlap. For this experiment, we again consider a well-specified setting for Lee et al. [2019], Ohnishi and Awan [2024], Guha and Reiter [2025]. The outcomes $Y_i \in \{0, 1\}$ are binary, and both the propensity score and potential outcomes follow logistic regression models. This setup is essentially the same as one of the well-specified experiments in Ohnishi and Awan [2024]. We set $n = 50000$, $d = 10$ and $B_\pi = 1/0.1$. The propensity score is $\pi(x) = \text{clip}_{[0.1,0.9]}(\text{expit}(0.1 + X_i^T\beta_\pi))$, where $\beta_\pi = (-0.15, 0.225, -0.15, -0.2, 0.1, 0.05, -0.075, 0.225, -0.15, -0.2)$. Here β_π is chosen to give us a range of propensity scores while only few samples require clipping. That is, we have large overlap. The outcome of a sample is generated by $\text{Bernoulli}(\text{expit}(-0.05 + \text{X}_\text{i}^\text{T}\beta_\mu + 0.42585\text{A}_\text{i}))$, where $\beta_\mu = (0.175, 0.1, -0.125, 0.075, -0.1, 0.2, -0.2, 0.175, -0.1, 0.2)$. This gives an ATE of 0.1.

For our estimators, we set 500 and run IPW, G-formula and AIPW with logistic regression for the nuisance estimators. We also run G-formula using a simple linear model. We run the experiment with two sets of privacy parameters. This experiment is an ideal setup for previous work, so we expect them to perform well here. This is confirmed by the results in Figure 2. However, our G-formula estimator achieves similar performance and outperforms our other estimators, as it requires less noise thanks to its lower sensitivity.

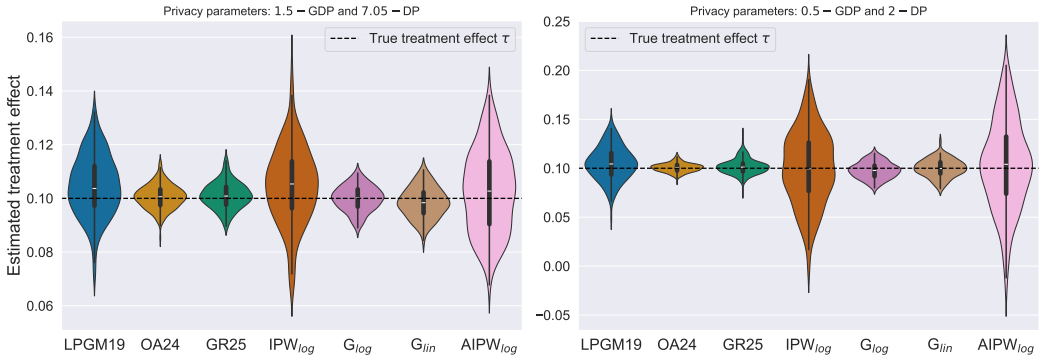

Figure 2: Well-specified setting with good overlap: propensity scores and outcome are linear models.

Effect of parameter K. Our approach requires selecting the number of folds K. By increasing K, we reduce sensitivity, allowing us to achieve differential privacy with less noise as shown in section 4. However, if we set K too high, the nuisance estimators may not have enough samples to train well. Therefore, we want to use the largest value of K that maintains good nuisance estimation performance. This choice depends on both the choice of nuisance estimator and the dataset size. Here, we evaluate the effect of K on our estimators for datasets with $n = 20000$,

$d = 20$, and $B_\pi = 1/0.1$. The propensity score follows a logistic regression model where $\pi(X_i) = \text{clip}_{[0.1,0.9]}(\text{expit}(0.1 + X_i^T \beta_\pi))$ with $\beta_\pi = (-0.17, -0.06, 0.05, 0.14, 0.12, -0.195, -0.205, 0.07, 0.18, 0.14, -0.14, 0.05, 0.01, -0.16, -0.18, -0.1, 0.2, 0.03, -0.16, -0.1)$. The outcome follows a linear regression model such that $Y_i = -0.08 + X_i^T \beta_Y + 0.15 A_i + e$, where $e \sim \mathcal{N}(0, 0.0025)$ and $\beta_Y = (-0.0385, -0.0111, -0.105, -0.0344, 0.1405, 0.0550, 0.0344, -0.0908, -0.0023, -0.0243, -0.0076, -0.0416, 0.0193, -0.0846, 0.0582, 0.0824, 0.0184, 0.0064, -0.0895, 0.0241)$.

The results are shown in Figure 3 (G-Formula) and fig. 4 (IPW and AIPW).

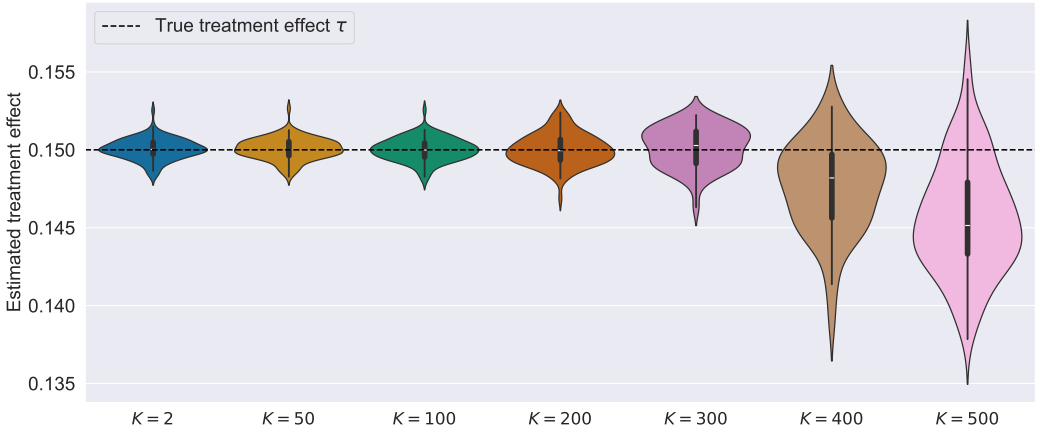

Figure 3: Effect of changing K for G-formula linear regression. The aggregated estimators perform well even as K increases. Although the nuisance estimators are each trained on smaller folds which increases variance this is balanced by aggregating more estimators. However, at $K = 400$ the performance of the aggregator estimator degrades as each fold contains only 50 samples.

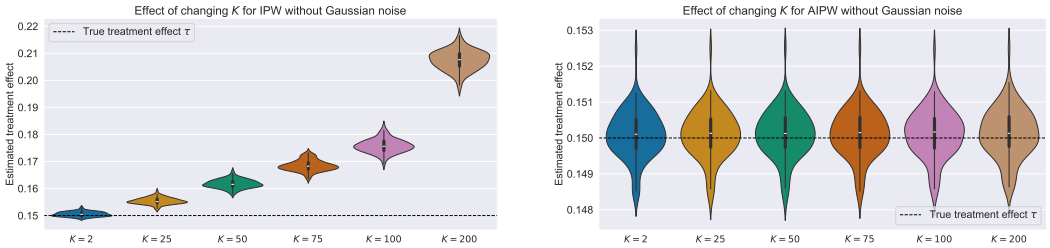

Figure 4: Effect of changing K for IPW and AIPW. We see that in this setup the accuracy of IPW quickly decreases. In contrast, AIPW benefits from the performance of G-formula and remains stable for moderate values of K.