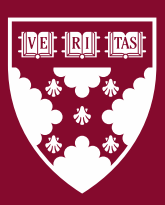
Working Paper 24-034

Debiasing Treatment Effect Estimation for Privacy-Protected Data: A Model Auditing and Calibration Approach

Ta-Wei Huang
Eva Ascarza



Harvard
Business
School

Debiasing Treatment Effect Estimation for Privacy-Protected Data: A Model Auditing and Calibration Approach

Ta-Wei Huang
Harvard Business School

Eva Ascarza
Harvard Business School

Working Paper 24-034

Funding for this research was provided in part by Harvard Business School.

Debiasing Treatment Effect Estimation for Privacy-Protected Data: A Model Auditing and Calibration Approach

Ta-Wei Huang
Harvard Business School, thuang@hbs.edu

Eva Ascarza
Harvard Business School, eascarza@hbs.edu

Data-driven targeted interventions have become a powerful tool for organizations to optimize business outcomes by utilizing individual-level data from experiments. A key element of this process is the estimation of Conditional Average Treatment Effects (CATE), which enables organizations to effectively identify differences in customer sensitivities to interventions. However, with the growing importance of data privacy, organizations are increasingly adopting Local Differential Privacy (LDP) — a privacy-preserving method that injects calibrated noise into individual records during the data collection process. Despite its privacy-protection benefits, we show that LDP can significantly compromise the predictive accuracy of CATE models and introduce biases, thereby undermining the effectiveness of targeted interventions. To overcome this challenge, we introduce a *model auditing and calibration* approach that improves CATE predictions while preserving privacy protections. Built on recent advancements in cross-fitting, gradient boosting, and multi-calibration, our method improves model accuracy by iteratively correcting errors in the CATE predictions without the need for data denoising. As a result, we can improve CATE predictions while maintaining the same level of privacy protection. Furthermore, we develop a novel *local learning with global optimization* approach to mitigate the bias introduced by LDP noise and overfitting during the error correction process. Our methodology, validated with simulation analyses and two real-world marketing experiments, demonstrates superior predictive accuracy and targeting performance compared to existing methods and alternative benchmarks. Our approach empowers organizations to deliver more precise targeted interventions while complying with privacy regulations and concerns.

Keywords: targeted intervention, conditional average treatment effect estimation, differential privacy, model calibration, gradient boosting

History: First version: September 18, 2023. This version: September 18, 2023.

1. Introduction

In the era of big data and advanced analytics, data-driven targeted interventions have become a powerful tool for organizations. Leveraging randomized controlled experiments and individual-level data, organizations can determine "who to target" based on observed individual characteristics, which significantly enhances the effectiveness and efficiency of their interventions. Fundamental

to these personalized interventions is the estimation of Conditional Average Treatment Effects (CATEs), which quantify the average difference in outcomes between treated and untreated individuals with similar characteristics (Athey and Wager 2021, Hitsch et al. 2023). With comprehensive covariate information and accurate outcome measures, CATE estimation enables precise and personalized decision-making. For example, in the field of marketing, CATE estimation helps companies optimize their promotional budgets by identifying individuals who are most likely to increase their purchases due to promotional efforts (e.g., Ascarza 2018, Lemmens and Gupta 2020, Simester et al. 2020, Ellickson et al. 2022). This data-driven strategy allows companiess to focus their resources on customers who respond most positively to interventions, thereby enhancing their return on investment.

However, as organizations gather user information and enhance their targeting abilities, concerns about the collection, storage, and safety of individual data are escalating. These concerns are shared by both consumers and regulators, highlighting the urgent need for advanced technologies that can ensure data confidentiality. While traditional privacy protection methods, such as anonymization and de-identification, have been employed by most organizations, recent research and real-world attacks have exposed their vulnerabilities, especially when faced with sophisticated adversaries with access to auxiliary information (e.g., Sweeney 1997, Narayanan and Shmatikov 2008, Acquisti and Gross 2009, Cohen 2022).

In response to these challenges, Differential Privacy (DP) (Dwork 2006) has emerged as a robust solution, adopted by leading companies and organizations. However, DP was originally designed to ensure privacy through a trusted central data curator, such as the US Census Bureau (Kenny et al. 2021), and did not account for situations where data curators might not be reliable (e.g., an organization suffering a data leak or misuse of its data). In these situations, it becomes necessary for individuals to protect their personal data before sharing it. To address this need, *Local Differential Privacy* (LDP) (Kasiviswanathan et al. 2011) has been proposed as an individual privacy protection mechanism that does not rely on a trusted curator. LDP enables data owners (e.g., users) to protect their data before it is shared with a central data curator (e.g., firms), thus ensuring privacy *locally*. This is achieved by introducing calibrated noise into individual data entries before they are stored in a database, making it challenging for attackers to extract precise information for re-identification, even with access to the raw data. The adoption of LDP by major technology firms such as Apple (Apple 2017) and Google (Erlingsson et al. 2014) underscores its importance in modern privacy protection and has stimulated extensive discussions among business practitioners (Yu and Smith 2018, Forbes 2022).

While LDP provides strong privacy guarantees, it also presents significant challenges for CATE estimation and targeted interventions. The random noise injected by LDP mechanisms can obscure

individual characteristics and the outcome of interest, both of which are essential to understanding differences in individual sensitivity toward an intervention. As a result, the noise intended to enhance privacy can reduce the precision of CATE estimates and even introduce bias. This can undermine organizations' ability to implement effective targeted interventions, as the very measures taken to preserve privacy can inadvertently compromise the quality of the data.

The objectives of this research are twofold. First, we examine the impact of local differential privacy (LDP) on the performance of state-of-the-art conditional average treatment effect (CATE) models. Our theoretical analysis shows that when LDP is used to protect the covariates, even unbiased CATE estimators (when estimated using non-protected experiment data) can exhibit heterogeneous bias across individuals with different covariates. This bias is particularly pronounced in highly non-linear CATE models, which presents a significant challenge for advanced machine learning techniques for CATE estimation, as they are specifically designed to uncover complex relationships in the data. On the other hand, when LDP is applied to protect the outcome of interest, an unbiased CATE model remains unbiased, but its variance will scale with the variance of the injected noise. This increase in variance introduces instability into CATE models, leading to less effective targeting policies. These findings imply that organizations would not necessarily get an effective targeting policy with LDP-protected data due to the increased bias and variance in CATE models, even when applying state-of-the-art methods on a large experiment dataset.

Second, we introduce the *Model Auditing and Calibration* approach, a *post-processing* method designed to enhance the predictive performance of existing CATE models while maintaining the same privacy guarantees. This solution utilizes an iterative adjustment procedure, commonly known as *boosting*, to refine the initial model's predictions based on their errors. Our approach addresses three major challenges that arise when applying boosting to CATE estimation with data protected by LDP. Firstly, since the true CATE is not directly observable, we propose to use a valid proxy of CATEs, such as Robinson's transformation, to provide a learning target for the boosting algorithm. Secondly, correcting predictions using the proxy CATE can introduce additional bias, as the proxy CATE may be severely biased due to the influence of the LDP noise. To overcome this problem, we propose a *local learning* strategy that focuses on the subgroup most likely to yield a significant improvement in accuracy. This approach enhances the efficiency of error correction by specifically utilizing individuals who are informative for reducing bias. Lastly, adaptively learning from those subgroups can result in overfitting bias. Therefore, we propose a *global optimization* approach that minimizes the error over the entire population to set the adjustment scale for initial predictions. This strategy mitigates the overfitting bias that may be introduced by local learning, resulting in more accurate CATE estimation.

Through extensive analyses using both simulated and real-world datasets, we present robust evidence supporting the effectiveness of our proposed method when the data are protected by LDP. By constructing scenarios where the covariates and the outcome variable are protected by LDP, we showcase that our proposed solution significantly reduces bias and enhances the overall accuracy of CATE models, thereby leading to more effective targeting policies. Beyond its improved effectiveness, there are several compelling reasons for firms to implement our proposed solution. First and foremost, our method does not require firms to denoise the LDP-protected data, ensuring that the privacy guarantee of LDP remains uncompromised. Second, our approach is cost-efficient as it circumvents the need for collecting additional experiment data to boost accuracy. Last but not least, our solution can be readily implemented using standard machine learning models and existing software packages, enabling firms to swiftly and efficiently deploy the model.

Our research makes three key contributions to the existing literature. Substantively, we provide comprehensive evidence, including theoretical analyses, simulations, and empirical demonstrations, to understand the impact of LDP on CATE estimation and targeting. Our study is the first to address this issue, emphasizing the need for systematic evaluation across various data-driven marketing practices. Methodologically, we introduce a novel method that improves the accuracy of CATE models on data protected by LDP (and more generally, in the presence of noisy data). This method synergistically combines concepts from causal inference, boosting, model calibration, and multi-group fairness, and effectively reduces bias in CATE estimation when the experiment data is protected by LDP. From a managerial perspective, our research provides decision-makers with an effective pathway to address the accuracy loss when implementing privacy-protection mechanisms. Our solution offers a practical tool for marketers to enhance targeting performance while maintaining the same level of privacy guarantees. Furthermore, we showcase the potential of post-processing as a promising approach to expand the Pareto frontier beyond the privacy-accuracy trade-off. This highlights the practical relevance of our work and provides direction for future methodological research on various marketing challenges involving privacy-protected data.

This paper is structured as follows. Section 2 outlines the connections to existing literature. Section 3 provides a theoretical examination of the impact of LDP on popular CATE models. Section 4 introduces the proposed solution and discuss its benefits. We evaluate the empirical performance of the proposed solution using simulation studies in Section 5 and with two real-world datasets in Section 6. Finally, we conclude in Section 7 with recommendations for future research.

2. Research Context and Related Literature

Our research intersects the rapidly developing fields of privacy-preserving technologies and personalized interventions. To contextualize, consider an online learning platform seeking to enhance

course completion rates through targeted interventions. These interventions, such as personalized email reminders, are based on a variety of factors, such as demographic information (e.g., educational background, pronouns, age) and past course records (e.g., courses passed or failed in the past). To design an efficient targeting policy, the platform conducts a randomized controlled experiment, with some students receiving emails and others not. Concerned about potential data breaches, the platform anonymizes the experiment data to protect user identities. However, an attacker with access to these anonymized data and public LinkedIn profiles could potentially de-anonymize the data by matching the real identities on LinkedIn with the anonymized data, using demographic information and course certificates posted on LinkedIn. This could allow the attacker to access users' true identities and their comprehensive historical course records in the experiment data. Consequently, the platform may consider implementing local differential privacy (LDP) as an additional measure to enhance the privacy protection of its users. LDP is a privacy-preserving technology that adds noise to data to make it difficult to identify individuals. However, implementing LDP can also hinder the platform's ability to effectively target their emails, since the noise introduced to protect privacy may degrade the quality of the data used for personalization. This example highlights the challenges of designing personalized interventions that preserve user privacy.

The majority of LDP literature concentrates on developing new privacy-protection mechanisms and theoretically proving how these methods guarantee privacy (e.g., Kasiviswanathan et al. 2011, Erlingsson et al. 2014, Ding et al. 2017). Recently, a few studies, like Niu et al. (2022), have begun exploring methods for private CATE estimation within the centralized framework of differential privacy. Research has also investigated the trade-off between privacy and accuracy, either within an information-theoretic framework (Sarwate and Sankar 2014, Kalantari et al. 2018, Zhong and Bu 2022) or a statistical accuracy framework (Showkatbakhsh et al. 2018, Amin et al. 2019). Despite the broad adoption of LDP by major tech companies, the impact of LDP data protection on CATE estimation, along with its subsequent implications for targeted interventions, remain largely unexplored in current literature. This research takes a pioneering step towards understanding the influence of LDP on this important marketing practice.

Our research relates to the growing literature on targeting and CATE estimation. Prior studies in this area have predominantly focused on developing new CATE models (e.g., Wager and Athey 2018, Künzel et al. 2019, Nie and Wager 2021, Kennedy 2020b) and applying them in various marketing contexts, including customer retention strategies (Ascarza 2018, Lemmens and Gupta 2020, Yang et al. 2023), membership subscription optimization (Simester et al. 2020, Yoganarasimhan et al. 2022), pricing and promotions (Smith et al. 2021, Ellickson et al. 2022, Daljord et al. 2023, Huang and Ascarza 2023), and catalog mailing campaigns (Hitsch et al. 2023). A relatively recent

and increasingly important line of research focuses on the challenges posed by existing CATE models in low signal-to-noise environments, which are typical when organizations employ LDP to protect the outcome variable. For example, recent work by Huang and Ascarza (2023) has shown that the variance of popular CATE models increases with the amount of unexplained variation in the outcome variable. In response to this issue, they proposed a solution that leverages low-variance signals in the data to reduce the noise in the outcome variable, thereby creating notably more effective targeting policies. Our work builds upon these contributions in two significant ways. Firstly, we extend the theoretical investigation to situations where LDP affects not just the outcome variable, but also the individual covariates. Secondly, as the solution proposed by Huang and Ascarza (2023) is not applicable to LDP (given the absence of low-variance signals), we propose a different approach to enhance the accuracy of CATE estimation and the effectiveness of targeting policies. The novelty of our new method lies in its ability to achieve further accuracy in CATE predictions without the need for additional data, such as short-term signals. This enables our approach to function in a wider range of scenarios.

Substantively, our research aligns with the literature on *covariate measurement errors*, which arises in our context when organizations inject noise into individual covariates to protect privacy. Studies such as Chesher (1991) and Battistin and Chesher (2014) have demonstrated that measurement errors in covariates can introduce bias into the estimation of regression coefficients and treatment effects. Existing bias correction methods primarily adopt two strategies. The first strategy utilizes a small validation set containing both noisy and clean variables to recover the error distribution, which is then used to adjust parameter estimates derived from a larger dataset with only noisy variables (e.g., Bound et al. 1989, Hsiao 1989, Carroll et al. 1999, Wang and Sullivan Pepe 2000, Chen et al. 2005, Hu and Ridder 2012, Yang et al. 2022). However, this method is less than ideal in the context of privacy protection, as the collection of actual data from a small subset of individuals can expose them to privacy risks. The second common approach treats the true covariates as latent variables and recovers them using instrumental variables (e.g., Hausman et al. 1991, Newey 2001, Schennach 2007, Hu and Schennach 2008) or repeated measurements of noisy variables (e.g., Li 2002, Schennach 2004a,b, Agarwal and Singh 2021). However, these approaches require relatively strong assumptions, such as the availability of instrumental variables or repeated measurements, which may not be applicable in many real-world situations. Complementing the existing literature, we provide a comprehensive theoretical and empirical analysis of the impact of covariate measurement errors on CATE estimation. We also propose a solution that does not rely on the existence of clean data, instrumental variables, or repeated measurements.

Methodologically, our research finds its roots in the literature on *calibration*, a process aimed at aligning predicted scores from a machine learning model with observed outcomes (Lichtenstein

et al. 1977, Platt et al. 1999, Zadrozny and Elkan 2002, 2001, van der Laan et al. 2023). This concept has been extended to address multi-group fairness by ensuring that predicted probabilities are well-calibrated for all computationally identifiable subgroups (Hébert-Johnson et al. 2018, Burhanpurkar et al. 2021). Building on this idea, Kim et al. (2019) developed a boosting algorithm that iteratively refines predictions by focusing on specific subgroups where the classification model underperforms and stops when the desired accuracy level is reached for each subgroup. In our work, we extend this framework to CATE estimation and introduce novel techniques to identify key subgroups and prevent overfitting. As we demonstrate in this research, these advances are crucial for ensuring accurate CATE predictions.

3. Problem: Impact of LDP on CATE Estimation

We begin by studying the impact of LDP on CATE estimation. We investigate two realistic scenarios: (i) when the covariates are protected by LDP and (ii) when the outcome of interest is protected by LDP. Before doing so, we first characterize the state-of-the-art CATE models, such as T-learners and Causal Forests, as these are the class of models that we focus on in our theoretical investigation.

3.1. Setup for CATE Estimation

Consider a decision maker aiming to improve a specific outcome (Y_i) through an intervention with two treatment conditions ($W_i \in \{0, 1\}$). The decision maker hypothesizes that the intervention could potentially alter the value of Y_i for each individual i, but the impact of this intervention may vary among individuals with different characteristics. This heterogeneous impact is characterized as the conditional average treatment effect (CATE), which measures the difference in expected outcomes among individuals based on their treatment assignment and pre-treatment characteristics. Mathematically, the CATE is defined as follows:

$$\tau(\mathbf{X}_i) \equiv \mathbb{E}[Y_i(1)|\mathbf{X}_i] - \mathbb{E}[Y_i(0)|\mathbf{X}_i],$$

where $Y_i(W_i)$ is the potential outcome (Rubin 1974, Holland 1986) of individual i's response given the treatment condition W_i, and $\mathbf{X}_i$ denotes a set of pre-treatment characteristics that are believed to moderate the treatment effect.

We assume that the decision maker conducts a randomized control experiment[1] on a subset of the individuals to estimate CATEs. We refer to the resulting outcomes (denoted as $\mathcal{Y}$), treatment assignments (denoted as $\mathcal{W}$), and covariates (denoted as $\mathcal{X}$) as the "experiment data" (denoted as $D = \{\mathcal{Y}, \mathcal{W}, \mathcal{X}\}$). Within this context, the experiment data needs to satisfy the following assumptions:

[1] For simplicity, we assume complete randomization in our analysis. However, our findings can be easily extended to scenarios where the treatment assignment depends on the covariates $\mathbf{X}_i$.

ASSUMPTION 1. ***(Identification of CATE)***

1. *[Complete Randomization] The treatment assignment within the experiment set is independent of their potential outcomes, i.e.,* $Y_i(1), Y_i(0) \perp\!\!\!\perp W_i$.
2. *[Overlap] The probability of an individual receiving (or not receiving) a treatment should always be positive. This can be represented as follows:* $0 < \mathbb{P}[W_i = 1|\mathbf{X}_i] < 1$ *for any possible* $\mathbf{X}_i$.
3. *[No Interference] The potential outcomes of each individual are independent of the treatments received by other individuals, i.e.,* $Y_i(1), Y_i(0) \perp\!\!\!\perp W_j,\ \forall j \neq i$.

We examine several popular CATE models commonly used in practice, such as Causal Forests (Wager and Athey 2018) and T-learners (Künzel et al. 2019). These models can be formally characterized under the following assumption:

ASSUMPTION 2. ***(Class of CATE Estimators)***

Let D *denote the experiment data available for CATE estimation. For a given individual with covariates* $\mathbf{x}_{\text{new}}$*, the predicted CATE can be expressed as the difference between two associated outcome models, i.e.,* $\widehat{\tau}(\mathbf{x}_{\text{new}}) = \widehat{\mu}^1(\mathbf{x}_{\text{new}}) - \widehat{\mu}^0(\mathbf{x}_{\text{new}})$*. The outcome models can be expressed as a weighted average of indivdual outcomes, given by:*

$$\widehat{\mu}^w(\mathbf{x}_{\text{new}}) = \sum_{i \in \mathcal{I}^o : W_i = w} \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) Y_i,$$

where $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$ *is the weight function,* $\mathcal{D}^\ell$ *denotes the information in* D *that is utilized to determine the weight, and* $\mathcal{I}^o$ *includes individuals in* D *that are used to generate outcome predictions. Furthermore, we assume that the CATE model is estimated using* honest estimation *(Wager and Athey 2018), whereby* $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$ *is derived without using the outcome information of individuals in* $\mathcal{I}^o$*. In this case, the weight function either depends only on the covariate information, or it is estimated using a separate dataset that does not include any information about individuals in* $\mathcal{I}^o$.

To understand this mathematical description intuitively, consider the case of a T-learner that employs nearest neighbors for outcome modeling. This T-learner calculates CATEs by contrasting predictions from two separate n-nearest neighbor models: one trained on the treatment group and the other on the control group. In this setting, the predicted CATEs can be formulated as follows:

$$\widehat{\tau}(\mathbf{x}_{\text{new}}) = \sum_{i:\ W_i=1} \underbrace{\frac{\mathbb{1}[\mathbf{X}_i \in N_n^1(\mathbf{x}_{\text{new}})]}{n}}_{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)} Y_i - \sum_{i:\ W_i=0} \underbrace{\frac{\mathbb{1}[\mathbf{X}_i \in N_n^0(\mathbf{x}_{\text{new}})]}{n}}_{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)} Y_i,$$

where $N_n^w(\mathbf{x}_{\text{new}})$ denotes the set of n-nearest neighbors in the experiment data with treatment assignment $W_i = w$ for the new individual. Here, $\mathcal{I}^o$ includes all individuals in the experiment data, while $\mathcal{D}^\ell$ consists of the covariates for every individual in D. Note that the weight functions in this context adhere to the honest estimation assumption since they only use covariate information.

As demonstrated in Huang and Ascarza (2023), the class of estimators described in Assumption 2 includes a wide range of CATE models, including T-learners and S-learners with varying outcome models (Künzel et al. 2019) and Causal Forest (Wager and Athey 2018). In our theoretical proofs, we extend our consideration to other CATE models that incorporate adjustment functions to reduce observed heterogeneity in outcomes, such as the R-learner (Nie and Wager 2021) and DR-learner (Kennedy 2020a). While these models are not discussed in detail here to maintain simplicity, the primary conclusions remain consistent for them.

3.2. Impact of LDP on CATE Estimation

3.2.1. Differentially-private Covariates. We first examine the scenario where the covariates are protected by LDP. This corresponds to cases where the data collector (e.g., a firm) is interested in protecting individual data such as demographic information, location, or past interactions with the firm that are useful for determining the targeting rule. Our analysis specifically focuses on mechanisms featuring an additive noise structure, which includes most commonly-employed LDP methods such as the Laplace mechanism (Dwork et al. 2006, Kasiviswanathan et al. 2011) and response randomization (Warner 1965, Erlingsson et al. 2014).[2]

Let's imagine the data collector observes individual covariates that are protected by LDP. For an individual i in the experiment data, the noisy covariate is denoted as $\widetilde{\mathbf{X}}_i = \mathbf{X}_i + \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ is the vector of random noises injected to the covariates. For each element $\eta_{i,p}$ in $\boldsymbol{\eta}_i$ (corresponding to the noise added to p-th covariate of individual i), we assume its k-th moment $\sigma_k \equiv \mathbb{E}[\eta_{i,p}^k]$ to be non-negative for $k \in \mathbb{N}$, which holds for the Laplace and randomized response mechanisms. Notably, larger moments indicate that the noises have a more pronounced effect on the data, thus enhancing stronger privacy protection. For simplicity, we assume that the noise distribution is consistent across all covariates. We refer to σ_k as the "maximum (k-th order) privacy level" that is applied to protect all the covariates. Similarly, we assume the data collector only has access to the noisy covariates of a "new" individual for CATE predictions, denoted as $\widetilde{\mathbf{x}}_{\text{new}} = \mathbf{x}_{\text{new}} + \boldsymbol{\eta}_{\text{new}}$, and each noise component adheres to the same distribution. Despite this simplified scenario, our theoretical

[2] The Laplace mechanism and response randomization are two widely-used techniques to achieve LDP. The Laplace mechanism perturbs the actual value of a numeric variable with a Laplace noise. The amount of noise is determined by the scale parameter of the Laplace distribution, which controls the level of privacy protection. On the other hand, response randomization is used for discrete variables. It involves a coin flip to decide whether to disclose the true value or report a randomly chosen alternative value, and the probability of reporting the truth controls the privacy level in this technique. These two techniques have been proven to satisfy the definition of LDP and serve as fundamental building blocks in the design of more advanced LDP mechanisms. Note that we can reformulate the response randomization technique as the injection of additive noise in the following way. Consider a binary variable X. The randomized response of X with flipping probability f can be expressed as $\tilde{X} = X + \mathbb{1}(X=1)\eta_1 + \mathbb{1}(X=0)\eta_0$, where $-\eta_1 \sim Bernoulli(f)$ and $\eta_0 \sim Bernoulli(f)$. A similar argument can be applied to the dummy transformation of a discrete variable with multiple possible values.

findings can be readily extended to accommodate situations with varying noise distributions across different covariates.

Our first objective is to characterize the factors that potentially affect the *bias* in the CATE models introduced by the injected noise in LDP covariates. Conceptually, introducing noise into covariates results in two main sources of errors: (1) When predicting CATE for any new individual, the data collector must rely on the perturbed covariates ($\widetilde{\mathbf{x}}_{\text{new}}$), instead of the actual covariates ($\mathbf{x}_{\text{new}}$), which can lead to significant prediction errors. (2) When estimating the CATE model, the noise injected into the experiment data will introduce measurement error biases to the weight functions (i.e., $\widehat{\ell}^0$ and $\widehat{\ell}^1$). More formally, we can characterize these two sources of bias with the following theorem:

THEOREM 1. ***(Bias Analysis: Differentially-Private Covariates)***
Suppose that the CATE model is an unbiased estimator of the true CATE function in the absence of LDP protection. Further, assume that both $\widehat{\tau}$ and τ satisfy the smoothness conditions detailed in Electronic Companion EC.1.1, allowing for the use of Taylor approximation. In the scenario described above, the potential bias in the predicted CATE can be written as

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}) - \tau(\mathbf{x}_{\text{new}})\right] \approx \underbrace{\Delta_\tau(\mathbf{x}_{\text{new}})}_{\text{Bias driven by } \boldsymbol{\eta}_{\text{new}}} + \underbrace{\sum_{i\in\mathcal{I}^o:\ W_i=1} \Delta_i^1(\mathbf{x}_{\text{new}}) - \sum_{i\in\mathcal{I}^o:\ W_i=0} \Delta_i^0(\mathbf{x}_{\text{new}})}_{\text{Bias driven by noises injected into the experiment set}}.$$

The bias resulting from noise injected into the covariates of the new individual is given by

$$\Delta_\tau(\mathbf{x}_{\text{new}}) = \underbrace{\sum_{k=1}^{K} \frac{1}{k!}\sigma_k \mathit{trace}\left(\partial^k_{\mathbf{x}_{\text{new}}}\tau(\mathbf{x}_{\text{new}})\right)}_{\approx \tau(\widetilde{\mathbf{x}}_{\text{new}}) - \tau(\mathbf{x}_{\text{new}})}.$$

The bias introduced by noises in the experimental set is given by

$$\Delta_i^w(\mathbf{x}_{\text{new}}) = \underbrace{\sum_{k=1}^{K} \frac{1}{k!}\sigma_k \mathit{trace}\left(\mathbb{E}\left[\partial^k_{\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right)}_{\equiv \Delta_i^{\ell w}(\mathbf{x}_{\text{new}}),\ \text{the expected impact on the weight}} \mathbb{E}\left[Y_i\right],$$

where $\mathcal{X}^\ell$ denotes the covariate information in $\mathcal{D}^\ell$.

Proof: See Electronic Companion EC.1.1.

Theorem 1 offers critical insights into the bias resulting from LDP protection. Firstly, the bias caused by the noise added to $\mathbf{x}_{\text{new}}$ can be characterized by $\Delta_\tau(\mathbf{x}_{\text{new}})$, which measures the disparity between the true CATE values on $\widetilde{\mathbf{x}}_{\text{new}}$ and $\mathbf{x}_{\text{new}}$. This term depends on (i) the magnitude of the noise, captured by σ_k, and (ii) the sensitivity of the true CATE function to minor fluctuations around $\mathbf{x}_{\text{new}}$, captured by trace $\left(\partial^k_{\mathbf{x}_{\text{new}}}\tau(\mathbf{x}_{\text{new}})\right)$. When the underlying structure of the true CATE

is highly nonlinear, the k-th order derivative could be non-zero for many values of k. This can result in a large magnitude of bias, especially when the privacy level (σ_k) is high. Furthermore, this bias element is independent of the CATE model in use, as it only depends on the true CATE function. As a result, it is an *irreducible bias*, meaning that it cannot be eliminated by adjusting the CATE model. The only way to reduce this bias is to remove the noise in the covariates of the new individual.

Secondly, the bias arising from the noise within the experiment data, $\Delta_i^w(\mathbf{x}_{\text{new}})$, is influenced by two primary factors: (i) the scale of noise introduced into the data, i.e., σ_k, and (ii) the expected sensitivities of the weight functions toward noises in $\mathcal{D}^\ell$, denoted as $\Delta_i^{\ell^w}(\mathbf{x}_{\text{new}})$. For example, consider the previously mentioned T-learner with nearest-neighbor method. In that case, the bias magnitude depends on the extent to which the nearest neighbors of $\mathbf{x}_{\text{new}}$ within the experiment data may shift due to the presence of noise. Significant changes would result in a greater bias. Notably, when a machine learning model exhibits high complexity (e.g., a small value of n for the nearest-neighbor estimator), the average sensitivity toward the noise will increase. This observation suggests that sophisticated machine learning models might encounter significant bias when the covariates are protected by LDP, particularly due to their inherent high non-linearity.

Next, we investigate the effect of LDP covariate noise injection on the variance of predicted CATEs. As with our prior analysis, our goal is to identify and characterize the factors that influence this variance. This leads us to the following theorem:

THEOREM 2. ***(Variance Analysis: Differentially-Private Covariates)***
Under the same setting and assumptions in Theorem 1, we can approximate the variance of the predicted CATE for a specific $\mathbf{x}_{\text{new}}$ *as follows:*

$$\begin{aligned}\text{Var}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}})\right] \;\approx\; \text{Var}\left[\widehat{\tau}(\mathbf{x}_{\text{new}})\right] + \underbrace{\sum_{k=1}^{K}\frac{1}{(k!)^2}\sigma_k^2\text{Var}\left[\Delta_{\widehat{\tau}}^k(\mathbf{x}_{\text{new}})\right] + \sum_{k=1}^{K}\frac{1}{k!}\text{Var}\left[\eta_{i,p}^k\right]\mathbb{E}\left[\Gamma_{\widehat{\tau}}^k(\mathbf{x}_{\text{new}})\right]}_{(A)} + \\ \underbrace{2\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\nu_k(\mathbf{x}_{\text{new}})}_{(B)} + \underbrace{\sum_{k_1\neq k_2}\frac{1}{k_1!}\frac{1}{k_2!}\sigma_{k_1}\sigma_{k_2}\zeta_{k_1,k_2}(\mathbf{x}_{\text{new}})}_{(C)}.\end{aligned} \tag{1}$$

Here, $\Delta_{\widehat{\tau}}^k(\mathbf{x}_{\text{new}}) = \text{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}})\right)$ *captures the k-th order influence of covariate noise on the predicted CATE, and* $\Gamma_{\widehat{\tau}}^k(\mathbf{x}_{\text{new}}) = \text{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right)$ *represents the squared perturbation of the predicted CATE caused by LDP noise.*[3] *The term* $\nu_k(\mathbf{x}_{\text{new}}) = \text{Cov}\left[\widehat{\tau}(\mathbf{x}_{\text{new}}), \Delta_{\widehat{\tau}}^k\right]$ *captures the covariance between the predicted CATE value (estimated on non-private data) and*

[3] $\circ 2$ denotes the Hadamard's (i.e., element-wise) square operator.

the k-th order change of predicted CATEs due to injected noises. Lastly, $\zeta_{k_1,k_2}(\mathbf{x}_{\text{new}}) = \text{Cov}\left[\text{trace}\left(\partial^{k_1}\widehat{\tau}(\mathbf{x}_{\text{new}})\right), \text{trace}\left(\partial^{k_2}\widehat{\tau}(\mathbf{x}_{\text{new}})\right)\right]$ *measures how much the different orders of predicted CATE differences are correlated with each other.*

Proof: See Electronic Companion EC.1.2.

Theorem 2 demonstrates the mixed effects of LDP covariate noise on the variance of the predicted CATE. In Equation (1), term (A) captures the increase in variance due to the additional randomness introduced by LDP protection. Since all the elements in (A) are strictly positive, this term is increasing with the magnitude of injected noise, specifically with respect to σ_k and $\text{Var}\left[\eta_{i,p}^k\right]$. Term (B) reflects the relationship between the predicted CATE estimated using non-private data (had it been available) and the influence of the injected noise on these predictions. If the injected noise tends to increase the predictions for individuals already showing a positive CATE (such that the covariance in (B) is positive), estimating CATE on injected noise will make even more extreme predictions. Consequently, LPD will further increase the variance of the predicted CATE. On the other hand, if this covariance is negative, the added noise would draw the predictions towards zero, reducing the variance of the predicted CATE. Term (C) denotes the co-movement across k_1-th and k_2-th order shifts in predicted CATEs. If these shifts tend to move in the same direction (i.e., having positive covariance), the variance is increased with larger σ_k. Conversely, if the k_1-th and k_2-th order fluctuations are in opposite directions (i.e., having negative covariance), they will offset each other, leading to no increase in variance.

Together, Theorems 1 and 2 highlight the inherent challenges in quantifying (and therefore correcting) the bias and variance posed by LDP protection in real-world applications. Existing methods for correcting covariate measurement error bias without auxiliary data quantify and remove the bias by leveraging information about the noise distribution and making parametric assumptions about the relationship between the desired outcome and the covariates (Battistin and Chesher 2014). However, this solution is not suitable for our context for several reasons. Firstly, the extent of the bias and variance is linked to the higher-order derivatives of the CATE models. Estimating these derivatives poses its own set of challenges, especially when dealing with complex machine learning models that lack straightforward analytical solutions. Moreover, the impact of injected noises on bias and variance vary based on the specific value of $\mathbf{x}_{\text{new}}$. Accurately estimating these impacts may be as difficult, if not more so, than the estimation of the CATE. Lastly, even if we derive some formulas for these higher-order derivatives for a specific CATE model, such a method would not be generalizable to other CATE models. These observations underscore the pressing need for a data-driven approach that can both identify individuals with potentially large prediction errors and correct those errors for *any CATE model*.

3.2.2. Differentially-private Outcome. We now turn to the scenario in which the outcome variable is protected by LDP. This applies to cases where the data curator intends to change individual behaviors that could be exploited by potential attackers for re-identification, such as test scores in classes, ratings of movies on social media, or clicks on a specific advertisement. Suppose that the focal firm only has access to the LDP-protected outcome $\widetilde{Y}_i(W_i) \equiv Y_i(W_i) + \eta_i$, where η_i denotes the injected noise drawn from the same distribution for all i.

In this setting, estimating CATEs is challenging due to the increased noise in the outcome variable. This noise is directly linked to the variance of η_i, which will lead to unstable CATE models and less accurate predictions. We formally characterize this result in the following theorem:

THEOREM 3. ***(Bias-Variance Analysis: Differentially-Private Outcome)*** *Consider a CATE model ($\widehat{\tau}$) that satisfies Assumption 2 and is an unbiased estimator of the true CATE function when there is no LDP protection. Let $\widehat{\tau}_{\widetilde{Y}}$ be a CATE model estimated on a the experiment data with the LDP-protected outcome $\widetilde{Y}_i = Y_i + \eta_i$. Then,*

1. **(Bias)** *The CATE model $\widehat{\tau}_{\widetilde{Y}}$ is still an unbiased estimator for the true CATE, i.e., $\mathbb{E}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\text{new}})] = \tau(\mathbf{x}_{\text{new}})$.*
2. **(Variance)** *The variance of the predicted CATE, $\text{Var}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\text{new}})]$, is bounded within a range that scales with the variance of the introduced noise, $\text{Var}[\eta_i]$. More formally, there exist constants C_1, C_2 such that $C_1\text{Var}[\eta_i] \leqslant \text{Var}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\text{new}})] \leqslant C_2\text{Var}[\eta_i]$ when $\text{Var}[\eta_i] > C$ for some C. Mathematically, this relationship can be expressed as $\text{Var}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\text{new}})] = \Theta(\text{Var}[\eta_i])$.*

Proof: See Electronic Companion EC.1.3 and EC.1.4.

Theorem 3.1 demonstrates that applying LDP to the outcome variable does not affect the consistency property of consistent CATE models. However, in practice, even consistent methods (e.g., Causal Forest) can be biased in small-sample scenarios (Wager and Athey 2018, Athey et al. 2019). We investigate such cases via simulation analyses (Section 5) and find that, when the CATE model is not unbiased (due to small sample size), the bias gets amplified when the outcome variable is protected by LDP. Specifically, the magnitude of this bias is increasing with the privacy level of the injected noise.

Furthermore, even if the CATE model remains unbiased with a small sample, Theorem 3.2 shows that the noise introduced by LDP will increase the variance of the CATE predictions, as both the upper bound and lower bound of $\text{Var}\left[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\text{new}})\right]$ are monotonically increasing in $\text{Var}[\eta_i]$ when the variance of the injected noise is non-trivial. This increase in variance of the predicted CATEs leads to higher mistargeting probability, as discussed in Huang and Ascarza (2023).

3.2.3. Summary of Theoretical Analysis. In summary, our investigation into the impact of LDP on CATE estimation reveals complex trade-offs between privacy protection and predictive accuracy. When covariates are protected by LDP, the introduction of noise can generate significant biases in the estimated CATEs, leading to potentially misleading targeting decisions. On the other hand, when the outcome variable is protected by LDP, unbiased CATE models are subject to increased variance while they can maintain their unbiasedness. This additional noise introduces instability into CATE models, resulting in suboptimal targeting policies.

Importantly, these complex trade-offs between privacy protection and predictive accuracy are moderated by several factors. Most notably, the bias and variance of the CATE predictions are individual-specific (i.e., depend on $\mathbf{x}_{\text{new}}$), and can vary substantially across different CATE models. These findings motivate us to develop a generic data-driven method designed to improve the precision of CATE predictions when data is under LDP protection. This approach is designed to accommodate complex true CATE functions and be compatible with *any* CATE model.

4. Solution: Debiasing by Model Auditing and Calibration

We now describe our proposed solution, which is designed to mitigate the prediction errors resulting from the noise injected by the LDP mechanism for *any* CATE models.

4.1. Solution Concept

Our solution employs the concept of *post-processing* in machine learning, which adjusts the raw model predictions to better fit the specific needs of the problem at hand. This technique has been utilized in a range of contexts, including improving model interpretability and ensuring valid inference (Chernozhukov et al. 2018b), optimizing model performance (Friedman 2001), and enhancing model transportability (Kim et al. 2022). In our case, post-processing involves refining the predictions of a CATE model that has been trained on data protected by LDP, with the aim of improving its accuracy.

Specifically, the proposed solution builds on the *gradient boosting* algorithm (Friedman 2001). Our iterative algorithm that starts with an initial CATE model. At each iteration, the algorithm calculates the residual errors, which are the differences between the model's predictions and the proxy values of the true CATEs. A model called the *calibrator* is then trained to predict these residual errors using the covariates. The predicted residual errors are then used to adjust the initial predictions. The algorithm further optimizes the degree of adjustment by minimizing the squared errors between the updated predictions and the proxy CATEs. By integrating the calibrator with the initial model, a new ensemble CATE model is formed. This process is repeated, with each subsequent model focusing on correcting the prediction errors made by its predecessor. As a result, a progressive improvement in the accuracy of the CATE predictions is achieved.

As we empirically demonstrate in Sections 5 and 6, our approach can significantly reduce the prediction error caused by the injected noise in LDP data. Unlike alternative approaches that focus on data cleansing (i.e., pre-processing the data to remove noise) or developing new (and more robust) CATE models, our post-processing approach offers several significant advantages:

1. *It maintains the desired privacy guarantee.*[4] Traditional approaches for correcting measurement error attempt to denoise the data and recover the true variable. However, this approach increases the privacy risk, as it increases the chances for an attacker to identify matches between the experiment data and their auxiliary information. In contrast, a post-processing approach preserves the privacy of the data by leaving the LDP noise in place. This minimizes the potential for re-identification of individuals.
2. *It does not depend on additional data collection or assumptions.* Many existing strategies for bias correction in the presence of covariate measurement error rely on either (i) leveraging auxiliary information such as a small noise-free sample (e.g., Yang et al. 2022) or instrumental variables (e.g., Hu and Schennach 2008), or (ii) assuming that the noisy covariates are repeated measurements of some low-dimensional characteristics and removing noise by constructing latent variables that approximate those characteristics (e.g., Agarwal and Singh 2021). However, the effectiveness of these approaches rests on relatively strong assumptions and the collection of clean data, which is often infeasible, especially in the context of privacy protection.
3. *It is model-agnostic.* As we demonstrate in our empirical sections, our proposed solution enhances the performance of a wide range of CATE models, including popular models such as meta-learners (Künzel et al. 2019), R-learner (Nie and Wager 2021), and Causal Forest (Wager and Athey 2018). This generalizability is particularly appealing for organizations, as it enables them to employ any off-the-shelf CATE model for the initial predictions and enhance their accuracy and targeting performance using our post-processing algorithm.

4.2. Challenges in Algorithm Development

While gradient boosting has shown remarkable success in supervised learning problems (e.g., Chen and Guestrin 2016), its application to CATE estimation, especially in situations with substantial noise, presents several challenges.

Challenge 1: The inability to directly observe the true CATE makes it challenging to establish a direct learning target for the gradient boosting algorithm.

[4] It is well-established that the output of a differentially-private algorithm can be post-processed or manipulated without compromising the privacy guarantees provided by the original algorithm (Kasiviswanathan et al. 2011, Dwork et al. 2014). This property allows further analysis and computation to be performed on the CATE modes estimated using LDP data while preserving its privacy guarantees.

The direct application of gradient boosting to CATE estimation is not straightforward due to the unobservable nature of the true CATE. Unlike traditional machine learning tasks such as classification or regression where the target variable is directly observable, CATE, which represents the individual-level incremental effect of an intervention, is not directly measurable (Rubin 1974). This introduces complexities in setting a direct learning target for the gradient boosting algorithm, which typically relies on observed targets for model construction.

Therefore, we must rely on proxy variables to approximate the true CATE. Several techniques such as Robinson's transformation (Nie and Wager 2021) or the doubly robust score (Kennedy 2020b) have been used to tackle this issue. In our method, we employ Robinson's transformation (Nie and Wager 2021) to approximate the true CATE:

$$\breve{\tau}_i = \frac{Y_i - \widehat{m}(\mathbf{X}_i)}{W_i - \widehat{e}(\mathbf{X}_i)},$$

where $\widehat{m}(\mathbf{X}_i)$ is a model for the conditional mean outcome $\mathbb{E}[Y_i|\mathbf{X}_i]$, and $\widehat{e}(\mathbf{X}_i)$ is a model for the propensity score $\mathbb{P}[W_i|\mathbf{X}_i]$.[5] By calculating the difference between the predicted CATE and $\breve{\tau}_i$, we obtain the proxy residual $\breve{\tau}_i - \widehat{\tau}(\mathbf{X}_i)$ that serves as the learning target for the boosting algorithm.

Challenge 2: The proxy variable derived under LDP may be severely biased due to the injected noise. This could potentially lead the boosting algorithm to learn from wrong signals, rather than correcting predictions based on the true error.

As highlighted in Section 3, the noise introduced through LDP can substantially impact CATE estimators. Similarly, this noise also make $\breve{\tau}_i$ a biased proxy of the true CATE. To illustrate this point, let us express the proxy variable derived under LDP as follows:

$$\breve{\tau}_i = \frac{(Y_i + \eta_i^Y) - \widehat{m}_{\widetilde{Y}}(\mathbf{X}_i + \boldsymbol{\eta}_i^X)}{W_i - \widehat{e}(\mathbf{X}_i + \boldsymbol{\eta}_i^X)},$$

where η_i^Y represents the noise injected into the outcome variable, and $\boldsymbol{\eta}_i^X$ denotes the noises injected into the covariates. It is clear that the variance of the proxy variable $\breve{\tau}_i$ increases as the variance of η_i^Y increases. Furthermore, the noise injected into Y_i can also increase the bias of $\widehat{m}_{\widetilde{Y}}$ when it is a highly non-linear model like random forest (Arlot and Genuer 2014). On the other hand, the noise $\boldsymbol{\eta}_i^X$ in the covariates will cause measurement error bias for the nuisance models ($\widehat{m}$ and $\widehat{e}$). As a result, $\breve{\tau}_i$ becomes a biased and/or high-variance proxy for the true CATE. In this situation, adjusting predictions using the proxy CATE can introduce additional bias, as the proxy CATE may be severely biased due to the influence of the LDP noise.

[5] In situations where the treatment assignment is completely randomized, we can use the treatment proportion as the propensity score instead of building a model.

To overcome this challenge, we introduce a *local learning* strategy. Rather than relying on the proxy residual error across the entire population, we direct our focus to a specific subgroup of individuals who: (i) have a significant "true error", and (ii) can significantly enhance the predictive accuracy of the initial model. Specifically, we first group customers by their predicted CATE values to identify customers with significant true errors. This strategy is effective because CATE models that generate extreme predictions, i.e., large positive or negative values, are more likely to be inaccurate due to the injected noise (Proposition 1). Therefore, subgroups with extreme predictions may be more informative in updating the model. Next, we update the initial model using the subgroup that can yield the best performance improvement. This step ensures that the algorithm always corrects the error in the most efficient manner. We provide details of this strategy in Section 4.3.

Challenge 3: Adaptive learning from subgroups can result in overfitting bias.

The process of selecting the most informative subgroup for model correction (i.e., local learning) can inherently lead to overfitting. Overfitting occurs when the model is excessively complex and fits the training data too closely, which can reduce its ability to predict out-of-sample observations accurately. This risk becomes particularly pronounced during the algorithm's later iterations, where the method might inadvertently identify subgroups that reduces the residual error on the current data, but not applicable to the larger population. This issue is a well-documented drawback when models are adaptively constructed (e.g., Varma and Simon 2006, Cawley and Talbot 2010, Berk et al. 2013, Dwork et al. 2015).

To address this challenge, we augment the local learning strategy with a *global optimization* approach, whereby the degree of adjustment based on the learned calibrator (commonly known as the *step size*) is determined by minimizing the sum of squared residuals across *all* individuals in the calibration set, rather than just those within the subgroup used to construct the calibrator. This approach considers information from the entire calibration set, mitigating the risk of overfitting posed by the local learning technique. In Section 4.3, we provide a detailed discussion on how to determine the optimal step size and why global optimization can indeed mitigate overfitting.

4.3. Solution Details: Debiasing by Model Auditing and Calibration

The discussion so far has highlighted the complexities in applying gradient boosting in our specific context, along with the conceptual solution. Now, we delve into the details of our proposed solution, which we call *Debiasing by Model Auditing and Calibration*. The pseudo-code for the algorithm is outlined in Algorithm 1. In the remaining of this section, we describe the main steps in the algorithm and explain its implementation details.

Algorithm 1 Debiasing by Model Auditing and Calibration

Input: $Q \in \mathbb{N}$, $R \in \mathbb{N}$

Output: $\widehat{\tau}(\mathbf{X}_i)$

Data: Experiment Data $D = \{D_{\text{train}}, D_{\text{cal}}, D_{\text{val}}\}$

Construct the conditional mean model $\widehat{m}$ and the propensity score model $\widehat{e}$ using D_{train}.

Derive the Robinson's transformations $\breve{\tau}_i = \frac{Y_i - \widehat{m}(\mathbf{X}_i)}{W_i - \widehat{e}(\mathbf{X}_i)}$ for $i \in \{D_{\text{cal}}, D_{\text{val}}\}$.

Construct the initial CATE model $\widehat{\tau}^{[0]}(\mathbf{X}_i)$ using D_{train} and generate predictions for D_{cal}.

Divide D_{cal} into Q subgroups $(\mathcal{G}^1_{\text{cal}}, \cdots, \mathcal{G}^Q_{\text{cal}})$ based on the (sorted) predictions from $\widehat{\tau}^{[0]}$.

for $r = 1, \cdots, R$ **do**

Calculate the proxy residual error $\breve{\tau}_i - \widehat{\tau}^{[r-1]}(\mathbf{X}_i)$ for $i \in D_{\text{cal}}$.

for $q = 1, \cdots, Q$ **do**

(Local Learning) Use individuals within the calibration set $\mathcal{G}^q_{\text{cal}}$ to construct a machine learning model, denoted as $\widehat{c}^{[r]}_q(\mathbf{X}_i)$, which is trained to predict $\breve{\tau}_i - \widehat{\tau}^{[r-1]}(\mathbf{X}_i)$ using $\mathbf{X}_i$.

(Global Optimization) Determine the step size by minimizing the R-loss across all the individuals in D_{cal}:

$$\rho^{[r]}_q = \arg\min_{\rho} \sum_{i \in D_{\text{cal}}} \left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho \widehat{c}^{[r]}_q(\mathbf{X}_i) - \breve{\tau}_i\right]^2$$

end

Pick the subgroup that leads to the smallest R-loss in the validation set:

$$q^\star = \arg\min_{q \in \{1, \cdots, Q\}} \sum_{i \in D_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho^{[r]}_q \widehat{c}^{[r]}_q(\mathbf{X}_i) - \breve{\tau}_i\right]^2.$$

Generate new predictions for the validation set:

$$\widehat{\tau}^{new}(\mathbf{X}_i) = \widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho^{[r]}_{q^\star} \widehat{c}^{[r]}_{q^\star}(\mathbf{X}_i), \ i \in D_{\text{val}}.$$

if $\sum_{i \in D_{\text{val}}} \left[\widehat{\tau}^{new}(\mathbf{X}_i) - \breve{\tau}_i\right]^2 - \sum_{i \in D_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i) - \breve{\tau}_i\right]^2 < 0$ **then**

Update the CATE model:

$$\widehat{\tau}^{[r]}(\mathbf{X}_i) = \widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho^{[r]}_{q^\star} \widehat{c}^{[r]}_{q^\star}(\mathbf{X}_i).$$

end

else

Stop the boosting algorithm and return $\widehat{\tau}^{[r-1]}$.

end

end

4.3.1. Iterative Error Correction Using Gradient Boosting. The core of our approach relies on the concept of *gradient boosting*. This method involves constructing a *calibrator* to correct the errors made by the previous ensemble of models (including the initial predictor and the calibrators). Often, a step-size learning procedure is employed to adjust the rate of error correction. Specifically, following the notation introduced earlier, each step aims to create a new ensemble model based on the following update rule:

$$\widehat{\tau}^{[r]}(\mathbf{X}_i) = \widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho^{[r]}\widehat{c}^{[r]}(\mathbf{X}_i), \tag{2}$$

where $\widehat{c}^{[r]}(\mathbf{X}_i)$ denotes the calibrator at each step r, and $\rho^{[r]}$ represents its step size.

Gradient boosting performs *gradient descent* on a loss function to guide the error reduction (i.e., determine $\rho^{[r]}$ and $\widehat{c}^{[r]}$) from the preceding ensemble. Each calibrator is trained to move in the direction that most effectively reduces the residual, which is the negative gradient of a squared-error loss function. Our goal is to minimize the R-loss function (Nie and Wager 2021)

$$\mathcal{L}(\widehat{\tau}^{[r-1]}, \breve{\tau}) = \frac{1}{2}\sum_i \left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i) - \breve{\tau}_i\right]^2,$$

and its the negative gradient is

$$-\frac{\partial \mathcal{L}(\widehat{\tau}^{[r-1]}, \breve{\tau})}{\partial \widehat{\tau}^{[r-1]}(\mathbf{X}_i)} = \breve{\tau}_i - \widehat{\tau}^{[r-1]}(\mathbf{X}_i).$$

Therefore, the calibrator $\widehat{c}^{[r]}$ is designed to predict the negative gradient for unseen individuals, which can be accomplished by building a machine learning model of $\breve{\tau}_i - \widehat{\tau}^{[r-1]}(\mathbf{X}_i)$ on $\mathbf{X}_i$. We refer to the step of estimating the residuals as *model auditing*, as it evaluates the error made by the CATE model. The subsequent updating step is termed *model calibration*, as it refines the CATE model for enhanced accuracy.

There are two common methods to determine the step size in (2). The first approach is to fix the step size, $\rho^{[r]} = \rho$, at each step and perform cross-validation to identify the optimal step size (Beygelzimer et al. 2015, Chen and Guestrin 2016). The second approach utilizes the *steepest descent approach* (Friedman 2001, 2002), where $\rho^{[r]}$ is determined by minimizing the R-loss in each adjustment. Our algorithm adapts the steepest descent approach with novel modifications required to overcome **Challenge 3**. (Further details in Section 4.3.3 and 4.3.4.)

4.3.2. Sample Splitting and Roles of Each Dataset. We partition the experiment data D into three subsets, namely $D_{\text{train}}, D_{\text{cal}}, D_{\text{val}}$. Each subset plays a distinct role for the algorithm: D_{train} is used for constructing the initial CATE model $\widehat{\tau}^{[0]}$ as well as training the conditional mean and propensity models ($\widehat{m}$ and $\widehat{e}$) essential for Robinson's transformation, D_{cal} is used for model

auditing and calibration, and D_{val} is used for subgroup selection and to determine when to stop the algorithm. The rationale for this partitions stems from multiple considerations.

First, it is essential to use distinct datasets for the initial CATE model construction and model calibration. If the same observation i is used for creating the initial CATE model ($\widehat{\tau}^{[0]}$) and perform model calibration, the prediction error of the initial CATE model ($\widehat{\tau}^{[0]}(\mathbf{X}_i) - \tau(\mathbf{X}_i)$) and the proxy error of Robinson's transformation ($\breve{\tau}_i - \tau(\mathbf{X}_i)$) will be positively correlated, as they are both driven by the LDP noise injected into the same individual. In this case, the true prediction error will be canceled out in the residual error since

$$\breve{\tau}_i - \widehat{\tau}^{[0]}(\mathbf{X}_i) = [\breve{\tau}_i - \tau(\mathbf{X}_i)] - [\widehat{\tau}^{[0]}(\mathbf{X}_i) - \tau(\mathbf{X}_i)].$$

As a result, the residual error would not be an informative learning target for the true prediction error. To break this correlation, we use a distinct dataset, namely D_{cal}, independent of D_{train}, for model calibration. In this case, the bias of $\widehat{\tau}^{[0]}$ is largely driven by the noise in D_{train}, while the bias of $\breve{\tau}_i$ is largely driven by the noise of $i \in D_{\text{cal}}$. Note that both the challenge and the practice of using sample splitting to address this problem have been extensively discussed in the literature on statistical inference involving nuisance models (Newey and Robins 2018, Mackey et al. 2018, Foster and Syrgkanis 2019) and CATE estimation (Wager and Athey 2018, Nie and Wager 2021, Semenova and Chernozhukov 2021, Kennedy 2020b).[6]

Second, when constructing Robinson's transformation ($\breve{\tau}_i$) for $i \in D_{\text{cal}}$, it is crucial to ensure that the nuisance models ($\widehat{m}$ and $\widehat{e}$) are estimated using independent data to ensure that the estimation errors of these models will not cause significant bias when learning from $\breve{\tau}_i$ (Semenova and Chernozhukov 2021, Nie and Wager 2021). To achieve this, we similarly utilize D_{train} to construct the nuisance models in the Robinson's transformation.

Third, we use a holdout set D_{val} to guide subgroup selection and to evaluate out-of-sample predictive accuracy during the iterative error correction process. The use of an independent validation set is a pivotal step, as it enables us to assess the generalization performance of the calibrated model and stop the algorithm once we observe no substantial improvements. This strategy helps mitigate the risk of overfitting and enhances the reliability of the algorithm.

While dividing the entire sample into three subsets may initially seem restrictive in terms of sample size, we can further enhance the sample size efficiency by swapping the roles of $D_{\text{train}}, D_{\text{cal}}, D_{\text{val}}$, repeating Algorithm 1, and averaging the resulting predictions. This process is known as *cross-fitting* and is frequently utilized in procedures that require sample splitting (e.g. Newey and Robins 2018, Chernozhukov et al. 2018a, Nie and Wager 2021, Kennedy 2020b).

[6] Essentially, employing sample splitting is essential to fulfill the Neyman orthogonality condition (Chernozhukov et al. 2018a). This condition ensures that any bias in the initial model $\widehat{\tau}^{[0]}$ does not have a first-order impact on the calibrated predictions.

4.3.3. Local Learning. Importantly, when data are protected by LDP, the proxy CATE ($\check{\tau}_j$) generally does not serve as a high-quality proxy due to the noise injection (**Challenge 2**). To overcome this challenge, we design the algorithm in a way that the calibrator is *trained using only* observations that provide *the most significant information for error correction*. We achieve this goal through a two-step process: First, we divide the calibration data into subgroups of individuals who are more (and less) likely to exhibit significant true errors. Then, we select the subgroup that most substantially enhances the predictive accuracy. The latter step is straightforward as it can be calculated via cross-validation. However, its effectiveness relies on the identified subgroups in the initial step truly displaying significant errors, rather than minor ones.

To identify such a subgroup of individuals, we leverage the (theoretical) finding that individuals with extreme CATE predictions (either large positive or large negative values) are more likely to experience significant prediction errors due to the noise injected into the training set. The following proposition formally characterizes this phenomenon for both cases: when covariates or outcomes are protected by LDP.

PROPOSITION 1. ***(Relationship between Predicted CATE and Injected Noise)***

1. (Protected Covariates) *In the scenario described in Theorem 1, the initial CATE prediction for an unseen individual, $\widehat{\tau}^{[0]}(\mathbf{x}_{\text{new}})$, is more extreme (i.e., has either large positive or large negative values) if the new individual and its influential peers in $\mathcal{I}^o$ (i.e., those with non-zero weights $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$) have large injected noise. The magnitude of the prediction is amplified if the weight function at $\mathbf{x}_{\text{new}}$ is highly sensitive to the noise.*
2. (Protected outcome) *In the scenario described in Theorem 3, the initial CATE prediction for an unseen individual, $\widehat{\tau}^{[0]}(\mathbf{x}_{\text{new}})$, is more extreme if the influential peers in $\mathcal{I}^o$ have significant injected noise. The magnitude and direction of the CATE prediction depends on several factors, including (i) the treatment condition of those peers, (ii) the sign of the injected noise, and (iii) the sign of their weights.*

Proof: See Electronic Companion EC.1.5.

Proposition 1 suggests that the CATE predictions from the *initial model* — particularly those with extreme prediction values — can be valuable in identifying subgroups of individuals for whom the prediction error, driven by LDP noise, is significant. Therefore, by categorizing individuals based on their *initial* predicted CATEs, we anticipate that those in groups with notably small or large predicted CATEs are likely to have the most significant prediction errors. Nonetheless, after addressing the prediction errors for these groups, other individuals may emerge as crucial for further error correction. As a result, our local learning approach employs an iterative error adjustment process with the following three steps:

(i) Group individuals in the calibration set (D_{cal}) into Q groups ($\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_Q$) based on their predicted CATEs from the initial CATE model ($\widehat{\tau}^{[0]}$).

(ii) For each step r, construct a calibrator $\widehat{c}_q^{[r]}(\mathbf{X}_i)$ for each subgroup $\mathcal{G}_q$, and determine the step size $\rho_q^{[r]}$ for each subgroup (Section 4.3.4 describes how the step size is determined).

(iii) Select the calibrator from the subgroup $\mathcal{G}_{q^\star}$ that yields the greatest improvement in the R-loss for all individuals in the validation set D_{val}, i.e.,

$$q^\star = \underset{q \in \{1,\cdots,Q\}}{\arg\min} \sum_{i \in D_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho_q^{[r]} \widehat{c}_q^{[r]}(\mathbf{X}_i) - \breve{\tau}_i\right]^2.$$

(iv) Finally, update the CATE model using the following formula:

$$\widehat{\tau}^{[r]}(\mathbf{X}_i) = \widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho_{q^\star}^{[r]} \widehat{c}_{q^\star}^{[r]}(\mathbf{X}_i).$$

It is important to note that we use a separate validation set, distinct from the data used for constructing the calibrator and determining the step size, for subgroup selection. This approach helps protect against potential overfitting of the calibration data and allows us to identify the subgroups that contribute to a reduction in true prediction error.

4.3.4. Global Optimization. While selecting individuals who contribute the most to error reduction through local learning enables more effective model adjustments, the highly adaptive nature of this algorithm can inadvertently lead to significant over-correction and introduce overfitting bias (**Challenge 3**). To alleviate this issue, we adopt a *global optimization* approach for determining the step size, rather than solely depending on $\mathcal{G}_{q^\star}$. Formally, we determine the step size as follows:

$$\rho_q^{[r]} = \underset{\rho}{\arg\min} \sum_{i \in D_{\text{cal}}} \left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i) + \rho \widehat{c}_q^{[r]}(\mathbf{X}_i) - \breve{\tau}_i\right]^2. \tag{3}$$

Through the inclusion of this global optimization step, we ensure that the update minimizes the R-loss across *all* individuals in the calibration set, not just those that contributed most to the error reduction on the validation set. As a result, we mitigate the risk of overcorrection for any specific subgroup, effectively striking a better balance between subgroup correction and overall model performance.[7]

4.3.5. Early Stopping. Finally, it is widely recognized that indefinitely continuing the boosting process can lead to overfitting of the data and inconsistent predictions (e.g. Grove and Schuurmans 1998, Jiang 2004, Zhang and Yu 2005). Hence, it is crucial to appropriately time the

[7] In Section 5, we empirically demonstrate that this adjustment to the algorithm can significantly reduce the overfitting bias.

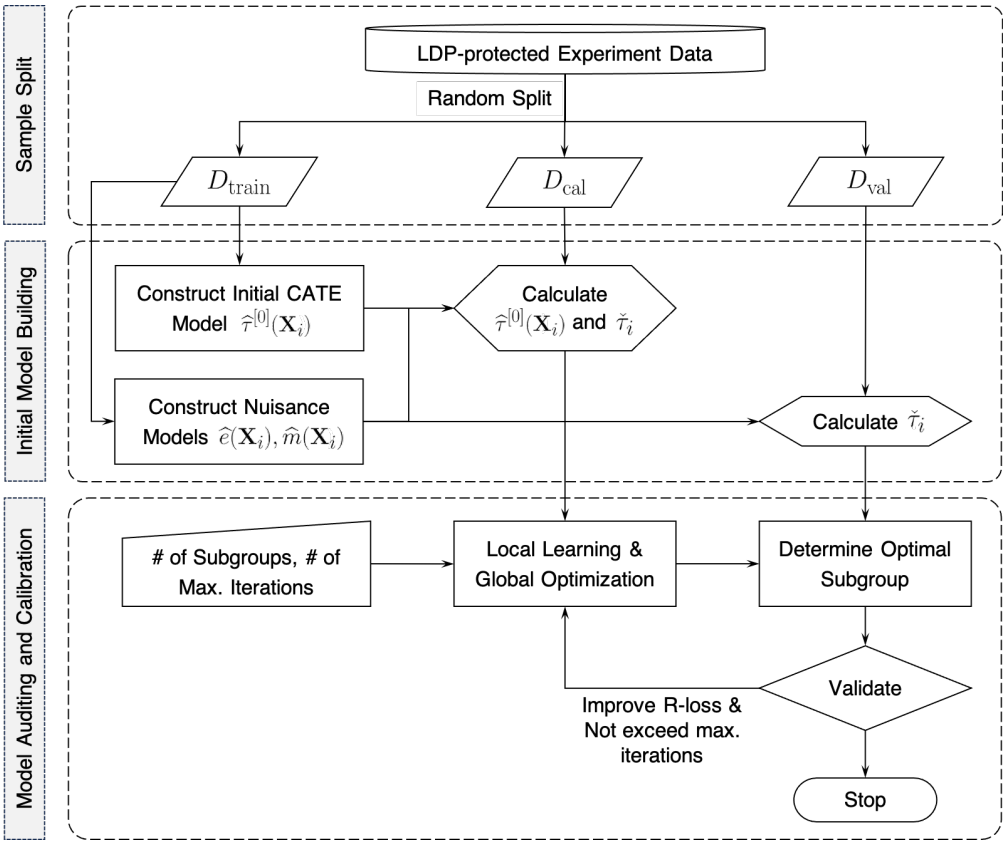
termination of the boosting process — late enough to avoid underfitting (i.e., large enough R), yet early enough to avoid overfitting.

In Algorithm 1, we employ a commonly-used method for early stopping, which is to stop when there is no improvement in predictive accuracy on a validation set. This forms our early stop criterion:

$$\sum_{i\in D_{\text{val}}}\left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i)+\rho_{q^\star}^{[r]}\widehat{c}_{q^\star}^{[r]}(\mathbf{X}_i)-\breve{\tau}_i\right]^2-\sum_{i\in D_{\text{val}}}\left[\widehat{\tau}^{[r-1]}(\mathbf{X}_i)-\breve{\tau}_i\right]^2<0,$$

which means that we truncate the procedure when there is no improvement of R-loss on D_{val}.[8]

Figure 1 Data Partitions and Usage in the Proposed Solution



4.3.6. Summary. Figure 1 offers a concise overview of the proposed algorithm, which encompasses three principal stages. The first stage, sample splitting, partitions the data into three sets: the training set is used to fit the initial CATE model, the calibration set is used to fine-tune the initial model, and the validation set is used for model selection and performance evaluation. The second stage, initialization, constructs the necessary models, including the initial CATE model and

[8] Note that the threshold of improvement can also be set as a tuning parameter in the post-processing algorithm. A larger threshold implies fewer updates being performed, acting as a form of regularization for the boosting process.

the nuisance models for Robinson's transformation. The final stage, model auditing and calibration, involves an iterative process of fine-tuning the model. The algorithm is elatively simple to implement and can be used with a variety of CATE models and data sets.

4.4. Model Selection and Hyperparameter Tuning

Our proposed solution offers high flexibility in the use of different models and hyperparameters, allowing several degrees of freedom for model implementation. We now provide guidelines for model selection and hyperparameter tuning.

- **Select Nuisance Models in Robinson's Transformation:** Two nuisance models are involved in this process, $\widehat{m}$ for the conditional mean outcome and $\widehat{e}$ for the propensity score. As suggested by Nie and Wager (2021), one should choose $\widehat{m}$ and $\widehat{e}$ that yield the highest accuracy as determined by cross-validation.
- **Select the Initial CATE Model ($\widehat{\tau}^{[0]}$):** As we demonstrate in our empirical sections, our proposed solution can enhance targeting performance irrespective of the initial CATE models used for initial predictions (including Causal Forest, T-learner, and R-learner). However, in practice, we recommend choosing the initial CATE model that delivers the best targeting performance (before model calibration) as our analyses consistently show that such a model typically maintains the best performance after post-processing. This initial selection can be easily implemented through the bootstrap validation procedure described in Section 6.3.
- **Select the Class of Calibrators $\left\{\widehat{c}^{[r]}\right\}_r$:** In theory, any type of model (e.g., linear regression, deep neural network) can be used to construct the calibrators. Based on our investigations on both simulated and real-world data, we find that simpler models (like linear regression) usually outperform more complex ones (like regression forest) in terms of both training speed and accuracy. Therefore, we suggest to use simpler models when fitting the calibrators. This recommendation also aligns with the philosophy of boosting, which seeks to enhance predictive performance by ensembling many weak models.
- **Hyperparameter Tuning:** Our algorithm primarily has two hyperparameters: the number of subgroups Q and the number of iterations R.
  - *Number of Subgroups* (Q): The choice for the number of subgroups presents an inherent bias-variance trade-off. If Q is too small, the algorithm may not effectively identify those individuals most informative for model correction, resulting in a larger bias. Conversely, if Q is too large, the algorithm can become unstable, increasing the variance. (This trade-off is further demonstrated using simulation in Appendix EC.2.2.) Overall, we recommend that each subgroup should consist of at least 100 individuals/observations to ensure algorithm stability.

— *Number of Iterations* (R): Through simulation analyses, we have found that setting $R = Q$ is typically sufficient. The reason behind this recommendation is that the step size for subgroup q often approaches zero when it has been previously used for updating the initial model. In other words, once the algorithm has utilized a specific subgroup to update the predictions, the value of updating based on that subgroup again diminishes considerably.

5. Empirical Performance: Simulation

We conduct simulation analyses for three main purposes. First, we validate the theoretical results outlined in Section 3 by investigating the privacy-accuracy trade-off. Second, we showcase the superior performance of our proposed method compared to several benchmarks when the data are protected by LDP. Third, we emphasize the sample size efficiency of our proposed algorithm.

5.1. Simulation Setup

We generate an experimental sample with a binary treatment variable ($W_i \in \{0, 1\}$) where the treatment assignment is completely random, with equal proportions for both treatment and control groups. The outcome variable is generated according to the following process:

$$
\begin{aligned}
Y_i(W_i) &= b(\mathbf{X}_i) + (W_i - 0.5)\tau(\mathbf{X}_i) + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, 5), \\
b(\mathbf{X}_i) &= \sin(\pi X_{i,1} X_{i,2}) - 2(X_{i,1} - X_{i,3} - 0.5)^2 + X_{i,2} X_{i,4} + 2X_{i,5}^2 + X_{i,6}^2, \\
\tau(\mathbf{X}_i) &= 2X_{i,1} - X_{i,2} + 0.5X_{i,3}^2 - X_{i,4} - \log(X_{i,1})(X_{i,4} - 1.5),
\end{aligned}
$$

where each covariate $X_{i,p} \sim Uniform(0, 5)$ is identically and independently distributed.

We examine two scenarios under the above data generating process (DGP). The first scenario assumes that the covariates are protected by LDP. Specifically, we only observe the noisy versions of $X_{i,p}$, denoted as $\widetilde{X}_{i,p} = X_{i,p} + \eta_{i,p}$, where $\eta_{i,p} \sim Laplace(0, \sigma)$ are i.i.d. Laplace noises. The second scenario assumes that the outcome is protected by LDP, implying that we can only observe the noisy outcome $\widetilde{Y}i(W_i) = Y_i(W_i) + \eta_i$, where $\eta_i \sim Laplace(0, \sigma)$ is the independently and identically distributed (i.i.d.) Laplace noise. For each scenario, we vary σ in the simulation to investigate the privacy-accuracy trade-off and manipulate number of individuals in the experiment set to explore the sample size efficiency.

5.2. Methods for Comparison

We evaluate the proposed algorithm against three alternative methods:

1. (Default) This method constructs a CATE model on the entire experiment data without any post-processing.

2. (Global) This method performs global learning and global optimization during the boosting procedure. Specifically, it constructs the calibrator and determines the step size using all of the calibration data.
3. (Local) This method performs local learning and local optimization during the boosting procedure. Specifically, it constructs the calibrator and determines the step size using subgroup data. The model is then updated based on the subgroup results that yield the greatest improvement in validation R-loss.

For methods involving boosting, we divide the experiment data into three equal-sized folds and perform cross-fitting accordingly. The detailed model specification can be found in Electronic Companion EC.2.1.

5.3. Evaluation Procedure

We evaluate predictive accuracy through $B = 100$ bootstrap replications and report the mean of key metrics. Initially, we generate a holdout set D_{holdout} consisting of $N_{\text{holdout}} = 10,000$ individuals which will be used for evaluation across all bootstrap samples. In each replication b, we generate an experimental set D^b (of varying sample size) and use it to construct the CATE estimation using the aforementioned methods. We then generate predictions on the holdout set and calculate the prediction errors for each method. Specifically, when the covariates are protected by LDP, the error of the model is calculated as $\widehat{err}_i^b = \widehat{\tau}^b(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)$. When only the outcome is protected by LDP, the prediction error of a model $\widehat{\tau}^b$ is defined as $\widehat{err}_i^b = \widehat{\tau}^b(\mathbf{X}_i) - \tau(\mathbf{X}_i)$. After executing this process on $B = 100$ bootstrap samples, we calculate the following statistical error metrics:

1. *(Mean Squared Error)* It evaluates the average squared error of the predicted CATEs across individuals in the holdout set, i.e.,

$$\widehat{\mathbb{E}}\left[\left(\widehat{err}_i^b\right)^2\right] = \frac{1}{N_{\text{holdout}}} \sum_{i \in D_{\text{holdout}}} \left[\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{err}_i^b\right)^2\right].$$

2. *(Mean Squared Bias)* It quantifies the average deviation of model predictions from the actual CATEs across individuals in the holdout set, i.e.,

$$\widehat{\mathbb{E}}^2\left[(\widehat{err}_i^b)\right] = \frac{1}{N_{\text{holdout}}} \sum_{i \in D_{\text{holdout}}} \left[\frac{1}{B}\sum_{b=1}^{B}\widehat{err}_i^b\right]^2.$$

3. *(Mean Variance)* It quantifies the average variability in the model's predictions across individuals in the holdout set when the model is trained with different data sets, i.e.,

$$\widehat{\text{Var}}\left[\widehat{\tau}^b(\mathbf{X}_i)\right] = \frac{1}{N_{\text{holdout}}} \sum_{i \in D_{\text{holdout}}} \left\{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\tau}^b(\mathbf{X}_i) - \widehat{\mathbb{E}}\left[\widehat{\tau}^b(\mathbf{X}_i)\right]\right)^2\right\}.$$

In addition to statistical accuracy metrics, we also measure how much our proposed method improves targeting performance when organizations use CATE models estimated on LDP-protected data for their targeting decisions. We assess this improvement using the Area Under the Targeting Operating Characteristic curve (AUTOC) (Yadlowsky et al. 2021), a widely-used metric that gauges a model's ability to correctly rank individuals based on their treatment effects. Specifically, given the predicted CATEs $\widehat{\tau}(\mathbf{X}_i)$ for i in the holdout set, the *Targeting Operating Characteristic* (TOC) is defined as the difference in treatment effects between individuals within the top $\phi \times 100\%$ predicted CATE tier and all individuals. That is,

$$\mathrm{TOC}(\phi; \widehat{\tau}) = \mathbb{E}\left[Y_i(1) - Y_i(0) | F_{\widehat{\tau}}(\widehat{\tau}_i) \geqslant 1 - \phi\right] - \mathbb{E}\left[Y_i(1) - Y_i(0)\right], \tag{4}$$

where $F_{\widehat{\tau}}$ is the cumulative distribution function of the predicted CATEs, and $\widehat{\tau}_i$ is defined as $\widehat{\tau}(\widetilde{\mathbf{X}}_i)$ when the covariates are protected by LDP and as $\widehat{\tau}(\mathbf{X}_i)$ when the outcome is protected by LDP. Then, the AUTOC is defined as

$$\mathrm{AUTOC}(\widehat{\tau}) = \int_0^1 \mathrm{TOC}(\phi; \widehat{\tau}) d\phi. \tag{5}$$

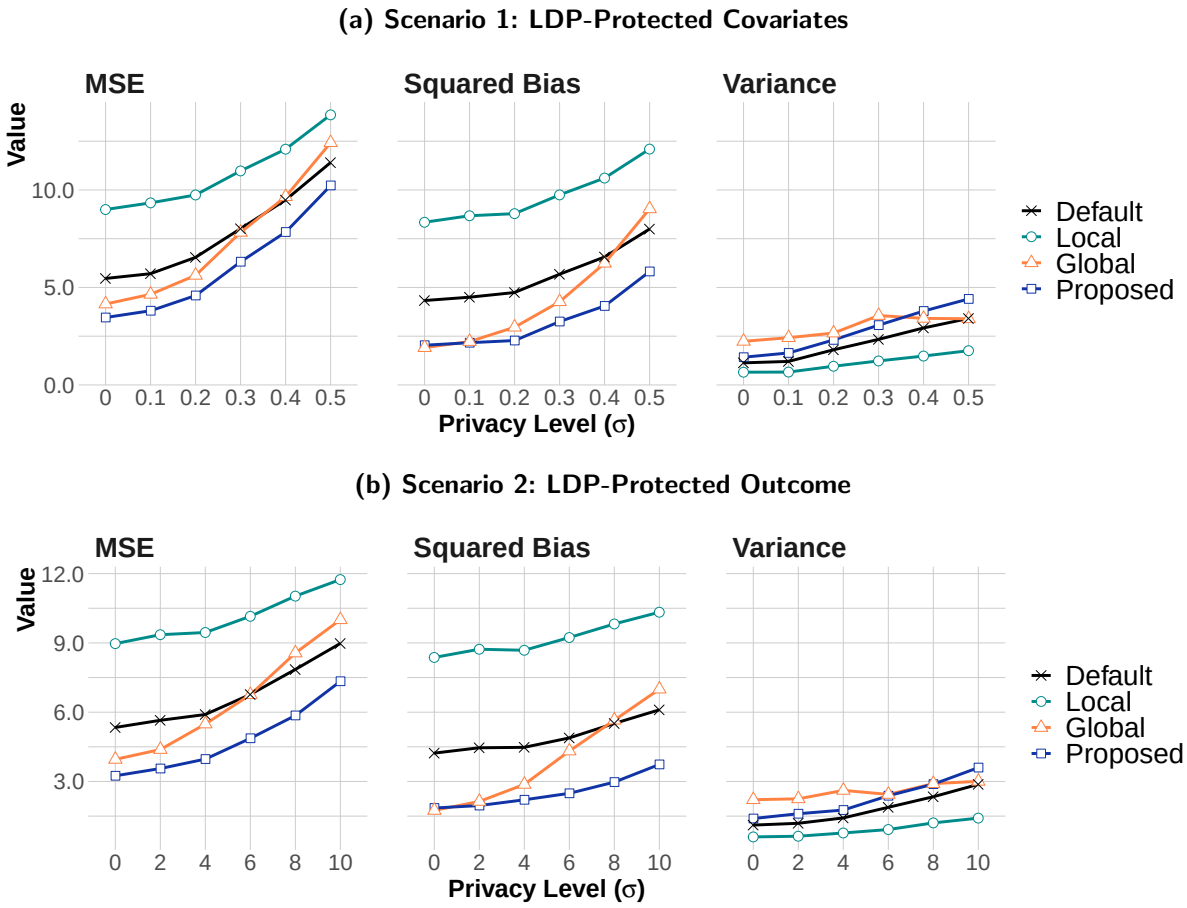
Note that a model $\widehat{\tau}$ outperforms another model $\widehat{\tau}'$ in identifying individuals in the top $\phi \times 100\%$ CATE group if $\mathrm{TOC}(\phi; \widehat{\tau}) > \mathrm{TOC}(\phi; \widehat{\tau}')$. Therefore, a higher AUTOC value suggests that the CATE model is more successful at identifying individuals who demonstrate the strongest sensitivity toward the intervention, leading to more effective targeting policies.

5.4. Results

Figure 2 illustrates the MSE, squared bias, and variance of different methods across a range of privacy levels ($\sigma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ for the covariates and $\sigma \in \{2, 4, 6, 8, 10\}$ for the outcome variable). Specifically, Figure 2a presents the results when the covariates are protected by LDP, and Figure 2b depicts the results when the outcome is protected by LDP. These results are generated using a sample of 3,000 individuals in the experimental set to estimate CATE models, and their performances are evaluated using a holdout set of 10,000 individuals.

Firstly, the results show clear trade-offs between privacy and accuracy: as the privacy level increases, all error metrics correspondingly increase, regardless of the method employed. Note that Causal Forest is generally biased in a finite-sample setting (Wager and Athey 2018). Therefore, the squared bias still increases with the privacy level in the case of LDP-protected outcomes. Secondly, the proposed method consistently outperforms other methods, achieving the lowest MSE across all privacy levels. Specifically, the proposed method yields the lowest bias, and the benefit of bias reduction (compared to the default method) becomes especially pronounced when the noise level is large. Notably, the proposed method has a lower bias than the global method, suggesting that

Figure 2 Predictive Errors of Different Methods Across Varying Privacy Levels



Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of Causal Forest as the initial CATE model. Results derived from different CATE models are available in Electronic Companion EC.2.3.

incorporating local learning strategies can effectively reduce bias. However, if local learning is not balanced with global optimization, the procedure may suffer from substantial overfitting bias, as evidenced by the local method having the highest bias irrespective of the noise level. In terms of variance, the local method yields the lowest variance, while the proposed and global calibration methods exhibit slightly higher variance than the default method.

Next, we present the results for the targeting performance of different methods. Table 1 compares the mean AUTOC values achieved by different methods, along with the percentage of replications (in parentheses) in which the AUTOC value of the focal method exceeds the AUTOC value of the default approach. First, consistent with previous results, there are clear trade-offs between privacy and targeting performance, as the AUTOC values decrease as the privacy level increase. Second,

Table 1 AUTOC Values of Different Methods Across Varying Privacy Levels

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**4.44 (100%)**	4.41 (85%)	4.34 (48%)	4.34	**4.44 (100%)**	4.41 (85%)	4.34 (48%)	4.34
Very Low	**4.39 (88%)**	4.36 (81%)	4.29 (45%)	4.29	**4.44 (91%)**	4.41 (74%)	4.33 (45%)	4.33
Low	**4.28 (91%)**	4.24 (77%)	4.18 (49%)	4.18	**4.37 (88%)**	4.31 (65%)	4.29 (55%)	4.28
Medium	**4.17 (93%)**	4.11 (73%)	4.09 (63%)	4.07	**4.34 (89%)**	4.28 (66%)	4.27 (71%)	4.23
High	**4.01 (85%)**	3.95 (76%)	3.95 (81%)	3.90	**4.25 (87%)**	4.16 (63%)	4.20 (86%)	4.13
Very High	**3.84 (80%)**	3.79 (68%)	3.83 (83%)	3.77	**4.19 (86%)**	4.11 (60%)	4.18 (88%)	4.09

Note: We report the average of AUTOC values from 100 simulation replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses). The results presented here are based on the use of Causal Forest as the initial CATE model. Results derived from different CATE models are available in Electronic Companion EC.2.3.

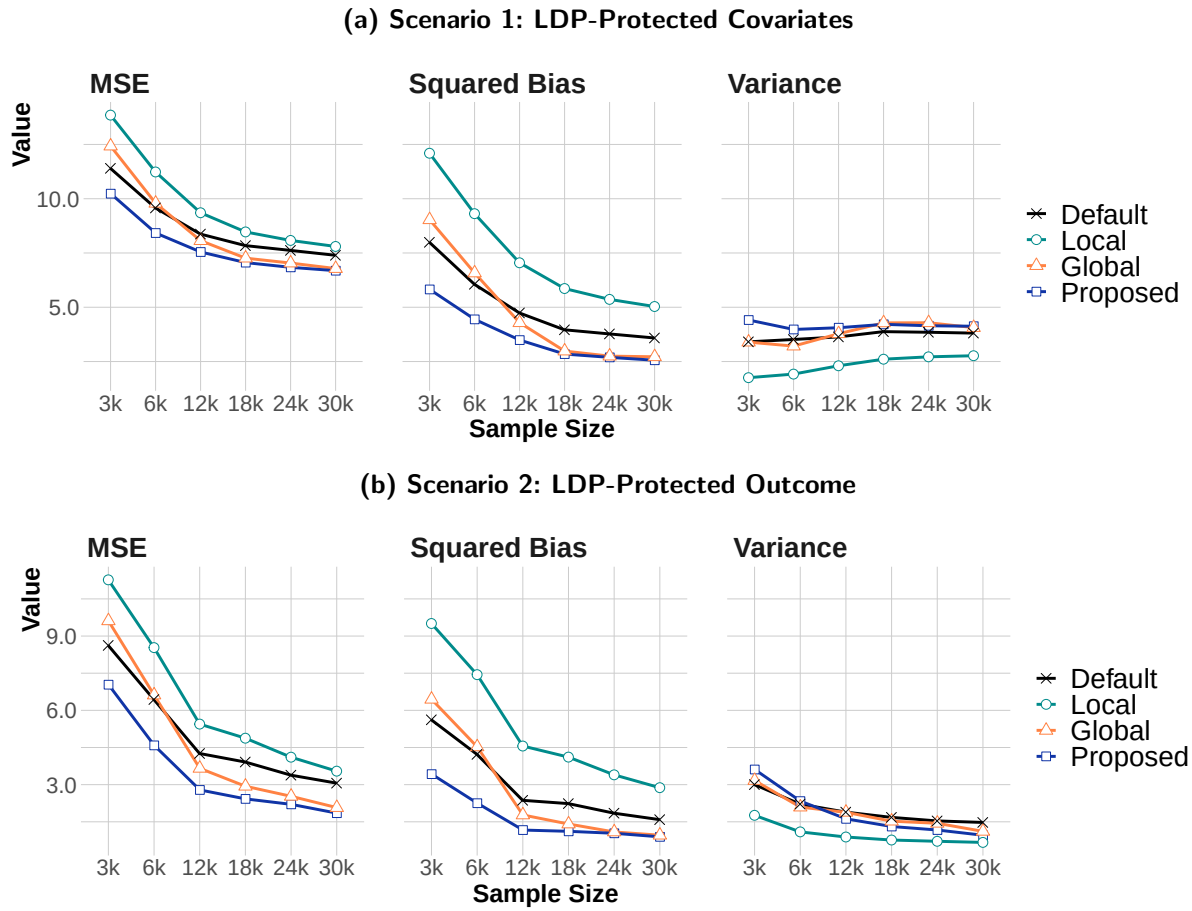
the proposed method consistently outperforms all other methods across all privacy levels. This result highlights that our proposed solution can significantly improve the targeting performance under LDP protection. Third, we find that global calibration notably underperforms when the privacy level is high, suggesting that calibration using all samples may be ineffective in the presence of large injected noise. Conversely, the targeting performance of the local approach improves, and even comparable with the proposed solution, at a very high privacy level. This indicates that the proposed subgroup learning approach is more efficient at identifying informative individuals for model calibration when the noise level is high.

5.5. Sample Size Efficiency

We now investigate the performance of the different methods across varying sizes of experiment data (ranging from 3,000 to 30,000 individuals) while keeping the privacy level fixed ($\sigma = 0.5$ for the covariates and $\sigma = 10$ for the outcome variable). Figure 3 presents the MSE, squared bias, and variance of different methods estimated on different sizes of experiment data.

The results showcase several key findings. Firstly, it is no surprise that both MSE and bias diminish as the sample size grows. Secondly, our proposed method stands out, consistently offering the lowest MSE and bias across varying sample sizes. In particular, the error reduction benefits of the proposed method are especially significant when the sample size is small. For example, our proposed method trained on a 12k sample can achieve a lower MSE and bias than the default method trained on a 30k sample (2.5 times larger). On the other hand, the proposed method and the global-only method demonstrate comparable predictive accuracy when the sample size is large. Finally, the local-only method significantly underperforms compared to the default approach, even

Figure 3 Predictive Errors of Different Methods with Varying Sample Sizes



Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of Causal Forest as the initial CATE model. Results derived from different CATE models are available in Electronic Companion EC.2.3.

when the sample size is large. This underscores the need to strike a balance between the benefits of local learning and the risk of overfitting when improving the accuracy of CATE estimates.

Regarding targeting performance, Table 2 compares the mean AUTOC values achieved by various methods along with the percentage of replications (in parentheses) where the AUTOC value of the proposed method outperforms the AUTOC value of the default approach. Firstly, the proposed method consistently achieves the highest AUTOC value across all sample sizes. This highlights the significant advantage of our proposed solution in terms of sample size efficiency. Secondly, consistent with the results on predictive errors, the global approach achieves similar performance to the proposed solution when the sample size is sufficiently large. In contrast, the local approach still performs worse than the proposed and global approaches, regardless of the sample size.

Table 2 AUTOC Values of Different Methods Across Varying Experiment Sample Sizes

Sample Size	Scenario 1: LDP-Protected Covariates ($\sigma = 0.5$)				Scenario 2: LDP-protected Outcome ($\sigma = 10$)			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
3,000	**3.84 (80%)**	3.79 (68%)	3.83 (83%)	3.77	**4.19 (85%)**	4.11 (60%)	4.18 (88%)	4.09
6,000	**3.92 (98%)**	3.87 (82%)	3.87 (88%)	3.82	**4.37 (95%)**	4.31 (82%)	4.31 (88%)	4.24
12,000	**3.96 (100%)**	3.94 (92%)	3.92 (88%)	3.89	**4.44 (100%)**	4.40 (93%)	4.38 (93%)	4.33
18,000	**4.02 (100%)**	4.02 (98%)	3.98 (80%)	3.96	**4.47 (100%)**	4.45 (98%)	4.42 (96%)	4.36
24,000	**4.03 (100%)**	4.03 (100%)	3.99 (95%)	3.97	**4.49 (100%)**	4.48 (100%)	4.44 (94%)	4.41
30,000	**4.09 (100%)**	4.09 (100%)	4.06 (90%)	4.04	**4.49 (100%)**	4.48 (100%)	4.44 (100%)	4.40

Note: We report the average of 100 simulation replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses). The results presented here are based on the use of Causal Forest as the initial CATE model. Results derived from different CATE models are available in Electronic Companion EC.2.3.

In conclusion, the simulation results demonstrate that our proposed solution can significantly reduce the prediction error and bias of existing CATE models across a wide range of privacy levels, leading to improved targeting performance. Notably, our approach is efficient in sample sizes and balances local learning and global optimization to achieve efficient bias reduction without overfitting. These findings offer valuable insights for businesses seeking to implement effective personalized interventions under the LDP protection.

6. Empirical Performance: Real-world Case Studies

We now validate the proposed solution using two real-world applications that have been used as benchmarks for CATE models (e.g., Rößler and Schoder 2022). Each of the studies corresponds to a randomized marketing campaign run by a firm with the aim to activate purchases among its customers, very common practice in marketing.

6.1. Studies Overview

6.1.1. Study 1: Hillstrom E-mail Campaign. The first study leverages the Hillstrom dataset, which originates from the MineThatData E-Mail Analytics And Data Mining Challenge (Hillstrom 2008). This dataset comprises 64,000 customers who were randomly assigned to one of three groups as part of an e-mail test: those who received an e-mailfor men's merchandise, those who received an e-mail for women's merchandise, and the rest who received no e-mail campaign. In line with previous research utilizing this dataset (Kane et al. 2014, Devriendt et al. 2018, Rößler and Schoder 2022), we focus on the effectiveness of the e-mail promotion for women's merchandise (compared to not receiving any e-mail at all). Consequently, our final sample includes a total of 42,693 customers, evenly split between the treatment group, which contains 21,387 customers, and the control group, which comprises 21,306 customers.

In this study, the primary goal is to establish a targeting policy that optimizes the impact of the intervention on website visits. To estimate the CATEs, we incorporate eight pre-treatment covariates into the analysis, including factors such as the recency of customer purchase, historical expenditure, and whether the customer has previously purchased women's merchandise, among others. The response rates for the treatment and control groups are 15.14% and 10.62% respectively, indicating an average treatment effect of 4.52%. A more detailed summary of this dataset can be found in Electronic Companion EC.3.

6.1.2. Study 2: Starbucks Promotional Campaign Data. The second study utilizes data from a promotional campaign conducted through the Starbucks reward mobile app. The dataset was made available by the Udacity Data Science Program. This experiment data involves a promotional campaign where some customers were randomly offered a promotion (the intevention) to encourage product purchase (the outcome variable). The dataset comprises 126,184 customers and seven anonymous pre-treatment covariates. With 63,112 customers in the treatment group and 63,072 in the control group, the response rates are notably low — 1.68% in the treatment group and 0.73% in the control group, resulting in an average treatment effect of 0.95%. See Electronic Companion EC.4 for more details.

6.2. Implementation of LDP

As these datasets are not protected by LDP, we simulate two scenarios where we implement the most common LDP methods: one where the pre-treatment covariates are proteted by LDP and another where the outcome is protected by LDP. For discrete variables, we apply the randomized response mechanism as suggested by Dwork et al. (2014). With this mechanism, there is a probability $1-f$ of observing the true value of the variable, whereas, with probability f, we solely observe a random draw from all possible values of the variable (each value appears with equal probability). For the continuous variables, we infuse it with $Laplace(0,\sigma)$ noise. We vary the privacy level from a "Very Low" scenario (i.e., small f and σ) to a "Very High" (i.e., large f and σ) scenario. Additional details about the implementation can be found in Electronic Companion EC.3.3 and EC.4.3.

6.3. Performance Evaluation

Similar to Section 5, we compare the proposed approach with the default alternative (estimating CATE directly on the noise-infused data) as well as the global-only and local-only approaches. (See details in Electronic Companion EC.3 and EC.4). To assess the performance of each method, we adopt a bootstrap validation scheme similar to that described by Ascarza (2018).[9] Briefly, we first generate $B = 100$ splits, each comprising an experimental set (70%) and a holdout set (30%). For

[9] Unlike in our simulation analyses, we cannot calculate statistical accuracy metrics as the true CATEs are not observed in real-world data.

each split, we estimate CATEs using each of the four approaches (proposed, default, local-only and global-only), and predict CATEs for all individuals in the holdout set. Subsequently, we calculate the targeting performance on the holdout set for each method using the the AUTOC metric.

It is important to note that in the scenario where the covariates are protected by LDP, we inject noise into covariates for both the experiment and the holdout set. This approach allows us to fully account for the impact of the injected noise on both the model construction process and the subsequent predictions. However, in the scenario when the outcome is protected by LDP, we do not inject noise into the outcome variable for the holdout set to reduce the variance of the AUTOC metric. This approach still provides the true AUTOC in privacy settings since the predicted CATEs for holdout individuals are not calculated using their own outcome values.

6.4. Results

Table 3 presents the AUTOC values (multiplied by 100) of different methods across a range of privacy levels, from "Very Low" to "Very High", under two specific scenarios: differentially-private covariates and differentially-private outcome. We observe three major findings. First, as the level of privacy protection increases, AUTOC values decline proportionally. This highlights the fundamental trade-off between privacy protection and targeting effectiveness. Second, all post-processing methods outperform the default method in terms of AUTOC. This demonstrates the value of post-processing in improving CATE estimation. Third, the proposed method consistently delivers the highest AUTOC of all methods for both datasets, highlighting the value of the proposed local learning with global optimization technique. In conclusion, the proposed method provides a promising approach for improving the targeting effectiveness of CATE models under differential privacy.

7. Conclusions and Future Directions

Local differential privacy has gained significant attention and adoption in recent years, particularly among leading technology companies. Nonetheless, the implications of its implementation on the precision of existing CATE models and firms' targeting capabilities remain largely unexplored. This paper takes the first step in investigating this issue, and proves — both theoretically and empirically — that the added noise from LDP not only increases the variance of the predicted CATEs, but can also introduce bias to existing CATE models, which significantly hinders firms' ability to run effective targeted interventions.

In response to this challenge, we introduce a new paradigm of CATE estimation where firms are encouraged to *post-process* the predicted CATEs before using them for targeting. Leveraging recent advancements in cross-fitting for heterogeneous treatment effect estimation, gradient boosting, and multi-calibration, our proposed *model auditing and calibration* approach enhances the accuracy of CATE predictions while maintaining the existing privacy protections on the experiment data.

Table 3 AUTOC Values (Multiplied by 100) of Different Methods for Hillstrom and Starbucks Data

(a) Study 1 : Hillstrom Data

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**1.62 (66%)**	1.58 (58%)	1.57 (63%)	1.49	**1.62 (66%)**	1.58 (58%)	1.57 (63%)	1.49
Very Low	**1.58 (78%)**	1.52 (71%)	1.49 (62%)	1.36	**1.58 (88%)**	1.51 (82%)	1.49 (79%)	1.29
Low	**1.54 (88%)**	1.45 (73%)	1.4 (71%)	1.28	**1.59 (90%)**	1.50 (78%)	1.42 (74%)	1.27
Medium	**1.44 (87%)**	1.38 (81%)	1.31 (76%)	1.14	**1.43 (88%)**	1.35 (74%)	1.26 (67%)	1.13
High	**1.29 (85%)**	1.23 (78%)	1.12 (65%)	0.99	**1.36 (91%)**	1.23 (83%)	1.16 (65%)	0.99
Very High	**1.22 (82%)**	1.13 (74%)	1.03 (71%)	0.91	**1.25 (85%)**	1.16 (76%)	1.01 (62%)	0.89

(b) Study 2 : Starbucks Data

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**0.59 (91%)**	0.58 (90%)	0.56 (76%)	0.53	**0.59 (91%)**	0.58 (90%)	0.56 (76%)	0.53
Very Low	**0.49 (90%)**	0.48 (87%)	0.44 (75%)	0.41	**0.42 (98%)**	0.38 (83%)	0.33 (69%)	0.29
Low	**0.42 (90%)**	0.4 (83%)	0.34 (58%)	0.33	**0.32 (96%)**	0.27 (87%)	0.23 (73%)	0.18
Medium	**0.37 (97%)**	0.34 (90%)	0.29 (70%)	0.26	**0.26 (96%)**	0.21 (86%)	0.18 (79%)	0.13
High	**0.28 (84%)**	0.26 (76%)	0.22 (47%)	0.22	**0.22 (94%)**	0.19 (88%)	0.16 (77%)	0.12
Very High	**0.21 (86%)**	0.18 (75%)	0.14 (60%)	0.13	**0.18 (88%)**	0.15 (77%)	0.13 (71%)	0.09

Note: We report the average of AUTOC values (multiplied by 100) from 100 bootstrap replications. Additionally, we provide in parentheses the percentage of replications for each post-processing method where its AUTOC value is greater than the AUTOC value of the default approach. The results presented here are based on the use of Causal Forest as the initial CATE model. Results derived from different CATE models are available in Electronic Companion EC.3 and EC.4.

Specifically, we develop a gradient boosting scheme that improves model accuracy without the need for data denoising. Therefore, this approach enhances targeting performance while ensuring the same level of privacy guarantees in the data.

Furthermore, we propose a novel *local learning with global optimization* method, which significantly reduces the bias in the calibration process caused by LDP's noise and overfitting. We evaluate the performance of our proposed method using both simulation analyses and empirical tests on real-world marketing data. The results demonstrate that our solution outperforms existing methods and various alternative benchmarks in terms of predictive accuracy and targeting performance.

Our work has several limitations, which suggest promising directions for future research. Firstly, we have characterized the impact of LDP on CATE estimation, providing key insights into how privacy-preserving methods can affect firms' ability to develop targeted interventions. However, the implications of LDP are far-reaching and extend beyond targeting. Future research could

examine the impact of LDP on other marketing problems, such as demand estimation and product recommendations, and develop new methods that are robust to this noise.

Secondly, we focus on scenarios where firms develop targeting policies based on predicted CATEs. However, the impact of LDP on other targeting methods is also worth exploring. For instance, the policy learning framework (e.g., Kitagawa and Tetenov 2018, Athey and Wager 2021) generates a proxy for the CATE and determines the individuals to be targeted by machine learning models that classify which individuals have positive CATEs. Although our method can also be utilized to enhance the quality of the proxy CATE, it would be interesting to investigate if we can directly enhance policy development, possibly by utilizing more robust machine learning models, to improve targeting ability when data is protected by LDP.

Thirdly, although we have quantified the uncertainty of the proposed algorithm in simulations using bootstrapping, we have yet to propose a theoretically justified method for constructing confidence intervals for the post-processed CATE. This challenge, largely driven by the complexity of our highly adaptive algorithm, could potentially be addressed by leveraging recent advancements in conformal inference, a statistical technique that offers valid prediction intervals under minimal assumptions (e.g., Shafer and Vovk 2008, Lei et al. 2018, Lei and Candès 2021). This integration could provide a measure of the uncertainty associated with these predictions, thereby complementing our current emphasis on predictive accuracy and targeting performance.

Finally, there is potential to further enhance our proposed algorithm. For example, we currently use the same validation set for both subgroup selection and determining the stopping time for the algorithm, which could lead to overfitting the validation set (Varma and Simon 2006, Cawley and Talbot 2010). Several innovative methods have been proposed to tackle this issue by improving algorithmic stability through information theory (Russo and Zou 2016) or differential privacy (Dwork et al. 2015). It would be useful to integrate these methods into our model auditing and calibration approach to further improve the stability and effectiveness of our proposed solution.

References

Acquisti A, Gross R (2009) Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences* 106(27):10975–10980.

Agarwal A, Singh R (2021) Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780* .

Amin K, Kulesza A, Munoz A, Vassilvtiskii S (2019) Bounding user contributions: A bias-variance trade-off in differential privacy. *International Conference on Machine Learning*, 263–271 (PMLR).

Apple (2017) Learning with privacy at scale. Technical report, Apple.

Arlot S, Genuer R (2014) Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939* .

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research* 55(1):80–98.

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *The Annals of Statistics* 47(2):1148–1178.

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

Battistin E, Chesher A (2014) Treatment effect estimation with covariate measurement error. *Journal of Econometrics* 178(2):707–715.

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. *The Annals of Statistics* 802–837.

Beygelzimer A, Hazan E, Kale S, Luo H (2015) Online gradient boosting. *Advances in neural information processing systems* 28.

Bound J, Brown CC, Duncan G, Rodgers WL (1989) Measurement error in cross-sectional and longitudinal labor market surveys: Results from two validation studies.

Burhanpurkar M, Deng Z, Dwork C, Zhang L (2021) Scaffolding sets. *arXiv preprint arXiv:2111.03135* .

Carroll RJ, Roeder K, Wasserman L (1999) Flexible parametric measurement error models. *Biometrics* 55(1):44–54.

Cawley GC, Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11:2079–2107.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen X, Hong H, Tamer E (2005) Measurement error models with auxiliary data. *The Review of Economic Studies* 72(2):343–366.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018a) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1):C1–C68.

Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I (2018b) Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india.

Chesher A (1991) The effect of measurement error. *Biometrika* 78(3):451–462.

Cohen A (2022) Attacks on deidentification's defenses. *31st USENIX Security Symposium (USENIX Security 22)*, 1469–1486.

Cohen J (2013) *Statistical power analysis for the behavioral sciences* (Academic press).

Daljord Ø, Mela CF, Roos JM, Sprigg J, Yao S (2023) The design and targeting of compliance promotions. *Marketing Science* .

Devriendt F, Moldovan D, Verbeke W (2018) A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* 6(1):13–41.

Ding B, Kulkarni J, Yekhanin S (2017) Collecting telemetry data privately. *Advances in Neural Information Processing Systems* 30.

Dwork C (2006) Differential privacy. *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, 1–12 (Springer).

Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A (2015) The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349(6248):636–638.

Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284 (Springer).

Dwork C, Roth A, et al. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4):211–407.

Ellickson PB, Kar W, Reeder III JC (2022) Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science* .

Erlingsson Ú, Pihur V, Korolova A (2014) Rappor: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 1054–1067.

Forbes (2022) Consumer privacy has evolved: How digital-forward companies should adapt. URL `https://www.forbes.com/sites/forbescommunicationscouncil/2022/08/09/consumer-privacy-has-evolved-how-digital-forward-companies-should-adapt/?sh=622db3bd1711`, accessed: 2023-05-15.

Foster DJ, Syrgkanis V (2019) Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036* .

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 1189–1232.

Friedman JH (2002) Stochastic gradient boosting. *Computational statistics & data analysis* 38(4):367–378.

Fry R, McManus S (2002) Smooth bump functions and the geometry of banach spaces: a brief survey. *Expositiones Mathematicae* 20(2):143–183.

Grove AJ, Schuurmans D (1998) Boosting in the limit: Maximizing the margin of learned ensembles. *AAAI/IAAI*, 692–699.

Hausman JA, Newey WK, Ichimura H, Powell JL (1991) Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics* 50(3):273–295.

Hébert-Johnson U, Kim M, Reingold O, Rothblum G (2018) Multicalibration: Calibration for the (computationally-identifiable) masses. *International Conference on Machine Learning*, 1939–1948 (PMLR).

Hillstrom K (2008) The minethatdata e-mail analytics and data mining challenge. Blog entry, Kevin Hillstrom: MineThatData, URL `https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html`.

Hitsch GJ, Misra S, Walter Z (2023) Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957* .

Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81(396):945–960.

Hsiao C (1989) Consistent estimation for some nonlinear errors-in-variables models. *Journal of Econometrics* 41(1):159–185.

Hu Y, Ridder G (2012) Estimation of nonlinear models with mismeasured regressors using marginal information. *Journal of Applied Econometrics* 27(3):347–385.

Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76(1):195–216.

Huang TW, Ascarza E (2023) Doing more with less: Overcoming ineffective long-term targeting using short-term signals. *Available at SSRN 4254202* .

Jiang W (2004) Process consistency for adaboost. *The Annals of Statistics* 32(1):13–29.

Kalantari K, Sankar L, Sarwate AD (2018) Robust privacy-utility tradeoffs under differential privacy and hamming distortion. *IEEE Transactions on Information Forensics and Security* 13(11):2816–2830.

Kane K, Lo VS, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics* 2:218–238.

Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A (2011) What can we learn privately? *SIAM Journal on Computing* 40(3):793–826.

Kennedy EH (2020a) Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* .

Kennedy EH (2020b) Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* .

Kenny CT, Kuriwaki S, McCartan C, Rosenman ET, Simko T, Imai K (2021) The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census. *Science Advances* 7(41):eabk3283.

Kim MP, Ghorbani A, Zou J (2019) Multiaccuracy: Black-box post-processing for fairness in classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254.

Kim MP, Kern C, Goldwasser S, Kreuter F, Reingold O (2022) Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences* 119(4):e2108097119.

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113(523):1094–1111.

Lei L, Candès EJ (2021) Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83(5):911–938.

Lemmens A, Gupta S (2020) Managing churn to maximize profits. *Marketing Science* 39(5):956–973.

Li T (2002) Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics* 110(1):1–26.

Lichtenstein S, Fischhoff B, Phillips LD (1977) Calibration of probabilities: The state of the art. *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, 275–324 (Springer).

Mackey L, Syrgkanis V, Zadik I (2018) Orthogonal machine learning: Power and limitations. *International Conference on Machine Learning*, 3375–3383 (PMLR).

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125 (IEEE).

Newey WK (2001) Flexible simulated moment estimation of nonlinear errors-in-variables models. *Review of Economics and statistics* 83(4):616–627.

Newey WK, Robins JR (2018) Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138* .

Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2):299–319.

Niu F, Nori H, Quistorff B, Caruana R, Ngwe D, Kannan A (2022) Differentially private estimation of heterogeneous causal effects. *Conference on Causal Learning and Reasoning*, 618–633 (PMLR).

Platt J, et al. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3):61–74.

Rößler J, Schoder D (2022) Bridging the gap: A systematic benchmarking of uplift modeling and heterogeneous treatment effects methods. *Journal of Interactive Marketing* 57(4):629–650.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688.

Russo D, Zou J (2016) Controlling bias in adaptive data analysis using information theory. *Artificial Intelligence and Statistics*, 1232–1240 (PMLR).

Sarwate AD, Sankar L (2014) A rate-disortion perspective on local differential privacy. *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 903–908 (IEEE).

Schennach SM (2004a) Estimation of nonlinear models with measurement error. *Econometrica* 72(1):33–75.

Schennach SM (2004b) Nonparametric regression in the presence of measurement error. *Econometric Theory* 20(6):1046–1093.

Schennach SM (2007) Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* 75(1):201–239.

Semenova V, Chernozhukov V (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* 24(2):264–289.

Shafer G, Vovk V (2008) A tutorial on conformal prediction. *Journal of Machine Learning Research* 9(3).

Showkatbakhsh M, Karakus C, Diggavi S (2018) Privacy-utility trade-off of linear regression under random projections and additive noise. *2018 IEEE International Symposium on Information Theory (ISIT)*, 186–190 (IEEE).

Simester D, Timoshenko A, Zoumpoulis SI (2020) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science* 66(6):2495–2522.

Smith AN, Seiler S, Aggarwal I (2021) Optimal price targeting. *Available at SSRN 3975957* .

Sweeney L (1997) Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25(2-3):98–110.

van der Laan L, Ulloa-Pérez E, Carone M, Luedtke A (2023) Causal isotonic calibration for heterogeneous treatment effects. *arXiv preprint arXiv:2302.14011* .

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7(1):1–8.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wang CY, Sullivan Pepe M (2000) Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 62(3):509–524.

Warner SL (1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60(309):63–69.

Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S (2021) Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966* .

Yang J, Eckles D, Dhillon P, Aral S (2023) Targeting for long-term outcomes. *Management Science* .

Yang M, McFowland III E, Burtch G, Adomavicius G (2022) Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science* .

Yoganarasimhan H, Barzegary E, Pani A (2022) Design and evaluation of optimal free trials. *Management Science* .

Yu P, Smith M (2018) Explainer: What is differential privacy and how can it protect your data? URL `https://theconversation.com/explainer-what-is-differential-privacy-and-how-can-it-protect-your-data-90686`, accessed: 2023-05-15.

Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, volume 1, 609–616.

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699.

Zhang T, Yu B (2005) Boosting with early stopping: Convergence and consistency. *The Annals of Statistics* (4):1538 – 1579.

Zhong H, Bu K (2022) Privacy-utility trade-off. *arXiv preprint arXiv:2204.12057* .

Electronic Companion

EC.1. Proofs for Theoretical Results

In the main document, we consider the class of estimators that can be written as the weighted average outcomes (as described in Assumption 2) for simplicity. In the following theoretical proofs, we extend our consideration to other CATE models that incorporate adjustment functions to reduce observed heterogeneity in outcomes, such as the R-learner (Nie and Wager 2021) and DR-learner (Kennedy 2020a). We start by describe the extended class of CATE estimators:

ASSUMPTION EC.1. ***(Extended Class of CATE Estimator)*** *For a given individual with covariates* $\mathbf{x}_{\text{new}}$*, the predicted CATE* $\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o, \mathcal{D}^\ell, \mathcal{D}^m)$ *can be expressed as the difference between two (adjusted) outcome estimators,* $\widehat{\mu}^0$ *and* $\widehat{\mu}^1$*, in the following form:*

$$\widehat{\mu}^w(\mathbf{x}_{\text{new}}) = \sum_{i \in \mathcal{I}^o : W_i = w} \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)],$$

where $\mathcal{I}^o$ *denotes the set of individuals used to impute the two outcome predictions,* $\mathcal{D}^\ell = \{\mathcal{X}^\ell, \mathcal{Y}^\ell, \mathcal{W}^\ell\}$ *represents the data used to construct the weight function* $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$*, and* $\mathcal{D}^m = \{\mathcal{X}^m, \mathcal{Y}^m, \mathcal{W}^m\}$ *is the data used to determine the adjustment function* $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)$*. We also denote* $\mathcal{D}^o = \{\mathcal{X}^o, \mathcal{Y}^o, \mathcal{W}^o\}$ *as the experiment data of individuals in* $\mathcal{I}_o$*, Furthermore, we assume that the estimation process of the CATE model satisfy the following conditions:*

1. *(Honest Estimation) The weight function is independent of* Y_j*,* $\forall j \in \mathcal{D}^o$*. In other words,* $\mathcal{D}^\ell$ *is either independent of* $\mathcal{D}^o$ *or depends only on the covariate values of individuals in* $\mathcal{I}^o$*.*
2. *(Cross Fitting) The adjustment function is either zero or constructed from the dataset* $\mathcal{D}^m$ *that is independent of both* $\mathcal{D}^o$ *and* $\mathcal{D}^\ell$*.*

In comparison to Assumption 2, the only distinction here is the incorporation of an adjustment function to account for the observed heterogeneity in the outcome. To arrive the theoretical results presented in the main document, we can can simply set any terms related to $\widehat{m}^w(\mathbf{X}|\mathcal{D}^m)$ as zero.

EC.1.1. Proof for Theorem 1.

For Theorem 1 and Theorem 2, we impose the following smoothness assumptions:

ASSUMPTION EC.2. ***(Smoothness Conditions)***

1. *(Differentiability) The weight function* $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}, \mathcal{D}^\ell)$*, and the* $\widehat{m}^w(\mathbf{X}_i, \mathcal{D}^\ell)$ *are* K*-th-time continuously differentiable for some* $K > 1$*. We also assume that the true CATE function* $\tau(\mathbf{x}_{\text{new}})$ *is also* K*-th-time continuously differentiable.*[10]

[10] While the weight function is not differentiable for tree-based methods, we can use the differentiable *smooth bump function* (Fry and McManus 2002) to approximate the weight function. Therefore, the Taylor approximation of the bias is still applicable for tree-based methods.

2. *(Bounded Sensitivity) The sensitivity of the weight function towards a specific covariate of each individual is bounded by a dominating function with a finite integral. That is, for any customer i and covariate p, we have*

$$\left|\frac{\widehat{\ell}_i^w(x_{i,p}+\delta_{i,p},\mathbf{x}_{-(i,p)})-\widehat{\ell}_i^w(x_{i,p},\mathbf{x}_{-(i,p)})}{\delta_{i,p}}\right|<g(x_{i,p},\mathbf{x}_{-(i,p)})$$

for any value of $\mathbf{x}_{-(i,p)}$ and bounded $\delta_{i,p}$, where $\int_{-\infty}^{\infty} g(x_{i,p},\mathbf{x}_{-(i,p)})d\mathbf{x}_{-(i,p)}<\infty$. Similarly, we assume that the the adjustment function satisfies the bounded sensitivity assumption. This condition ensures that the derivative and expectation operators are exchangeable.

We now present the extended version of Theorem 1 and provide the proof.

THEOREM EC.1. ***(Extended Version of Theorem 1)*** *Suppose that the CATE model is an unbiased estimator of the true CATE function in the absence of LDP protection. Further, assume that both $\widehat{\tau}$ and τ satisfy the smoothness conditions detailed in Electronic Companion EC.1.1, allowing for the use of a Taylor approximation. In the scenario as described above, the bias in the predicted CATE can be written as*

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}})-\tau(\mathbf{x}_{\text{new}})\right] \approx \underbrace{\Delta_\tau(\mathbf{x}_{\text{new}})}_{\text{Bias driven by }\boldsymbol{\eta}_{\text{new}}} + \underbrace{\sum_{i\in\mathcal{I}^o:\ W_i=1}\Delta_i^1(\mathbf{x}_{\text{new}}) - \sum_{i\in\mathcal{I}^o:\ W_i=0}\Delta_i^0(\mathbf{x}_{\text{new}})}_{\text{Bias driven by noises injected into the experiment set}}.$$

Specifically, the bias resulting from noise injected into the covariates of the new individual is given by $\Delta_\tau(\mathbf{x}_{\text{new}})=\sum_{k=1}^K \frac{1}{k!}\sigma_k trace\left(\partial^k_{\mathbf{x}_{\text{new}}}\tau(\mathbf{x}_{\text{new}})\right)$. The bias introduced by noises in the experimental set can be further broken down into three components: the effects on the weight function, the effects on the adjustment function, and the interaction between these two functions, as expressed by

$$\Delta_i^w(\mathbf{x}_{\text{new}})=\underbrace{\Delta_i^{\ell^w}(\mathbf{x}_{\text{new}})\mathbb{E}\left[Y_i-\widehat{m}^w(\mathbf{X}_i,\mathcal{D}^m)\right]}_{\text{Impact on weight}} - \underbrace{\mathbb{E}\left[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}},\mathcal{D}^\ell)\right]\Delta_i^{m^w}(\mathbf{x}_{\text{new}})}_{\text{Impact on adjustment}} - \Delta_i^{\ell^w}(\mathbf{x}_{\text{new}})\Delta_i^{m^w}(\mathbf{x}_{\text{new}}),$$

where $\Delta_i^{\ell^w}(\mathbf{x}_{\text{new}}) = \sum_{k=1}^K \frac{1}{k!}\sigma_k trace\left(\mathbb{E}\left[\partial^k_{\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right)$ represents the product of the noise magnitude and the average sensitivity of the weight function, and $\Delta_i^{m^w}(\mathbf{x}_{\text{new}}) = \sum_{k=1}^K \frac{1}{k!}\sigma_k trace\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i,\mathcal{X}^m}\widehat{m}^w(\mathbf{X}_i,\mathcal{D}^m)\right]\right)$ is the product of the noise magnitude and the average sensitivity of the adjustment function.

Proof: For all the following proofs, we denote $\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m,\widetilde{\mathcal{D}}^o$ as the noise-injected experiment data. First, by the unbiasedness assumption of $\widehat{\tau}$, we can write the bias as

$$\begin{aligned}\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}})\right]-\tau(\mathbf{x}_{\text{new}}) &= \mathbb{E}\left[\widetilde{\widehat{\mu}}^1(\widetilde{\mathbf{x}}_{\text{new}})-\widetilde{\widehat{\mu}}^0(\widetilde{\mathbf{x}}_{\text{new}})\right]-\mathbb{E}\left[\widehat{\mu}^1(\mathbf{x}_{\text{new}})-\widehat{\mu}^0(\mathbf{x}_{\text{new}})\right]\\ &= \mathbb{E}\left[\widetilde{\widehat{\mu}}^1(\widetilde{\mathbf{x}}_{\text{new}})-\widehat{\mu}^1(\mathbf{x}_{\text{new}})\right]-\mathbb{E}\left[\widetilde{\widehat{\mu}}^0(\widetilde{\mathbf{x}}_{\text{new}})-\widehat{\mu}^0(\mathbf{x}_{\text{new}})\right],\end{aligned}$$

where $\widetilde{\widehat{\mu}}^w(\widetilde{\mathbf{x}}_{\text{new}}) = \sum_{i\in\widetilde{\mathcal{D}}^o:W_i=w} \widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)[Y_i - \widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}^m)]$ is the implied outcome model under LDP protection. Therefore, to quantify the bias, our goal is to derive the bias of the implied outcome models, i.e.,

$$\mathbb{E}\left[\widetilde{\widehat{\mu}}^w(\widetilde{\mathbf{x}}_{\text{new}}) - \widehat{\mu}^w(\mathbf{x}_{\text{new}})\right].$$

We start by considering the bias for each individual i in $\mathcal{D}^o$ with treatment assignment w. We can write the total impact of the injected noises on the predicted CATE can be written as

$$\begin{aligned}
&\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)\left[Y_i - \widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m)\right] - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\left[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right] \\
&\quad= \left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right] Y_i - \widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) + \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m) \\
&\quad= \left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\left[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right] - \widehat{\ell}^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right] \\
&\quad= \underbrace{\left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\left[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]}_{A_i} - \\
&\quad\quad \underbrace{\left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]}_{B_i} - \\
&\quad\quad \underbrace{\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]}_{C_i}.
\end{aligned} \tag{EC.1}$$

Now, our objective is to determine the approximate discrepancies in the weight and adjustment functions, namely, $\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$ and $\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)$. Using Taylor approximation, the approximate deviation in weight functions due to the introduced noises in $\widetilde{\mathbf{x}}_{\text{new}}$ and $\widetilde{\mathcal{X}}^\ell$ can be written as:

$$\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) \approx \sum_{k=1}^{K}\left(\frac{1}{k!}\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right)\boldsymbol{\eta}^{\circ k}_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell},$$

where $\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$ is the k-th order derivative of $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$ with respect to $\mathbf{x}_{\text{new}}$ and $\mathcal{X}^\ell$, $\boldsymbol{\eta}_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}$ denotes the noises being injected to $\mathbf{x}_{\text{new}}$ and $\mathcal{X}^\ell$, and $Z^{\circ k}$ represents a vector or matrix where each element is raised to the power k from the corresponding element in Z.

Similarly, we can write the impact of injected noises on the adjustment function as:

$$\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m) \approx \sum_{k=1}^{K}\left(\frac{1}{k!}\partial^k_{\mathbf{X}_i,\mathcal{X}^m}\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right)\boldsymbol{\eta}^{\circ k}_{\mathbf{X}_i|\mathcal{D}^m}.$$

Using the above Taylor approximations and the fact that $Y_i,\ i \in \mathcal{D}^o$, $\mathcal{D}^\ell$, and $\mathcal{D}^m$ are independent, we can write the expectation of A_i in Equation (EC.1) as

$$\begin{aligned}\mathbb{E}\left[A_i\right] &= \mathbb{E}\quad \underbrace{\left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]}_{\approx \sum_{k=1}^K \left(\frac{1}{k!}\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right)\boldsymbol{\eta}^{\circ k}_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}} \quad \mathbb{E}\left[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right] \\ &\approx \mathbb{E}\left[Y_i - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]\cdot \sum_{k=1}^K \frac{1}{k!}\sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right).\end{aligned}$$

Similarly, the expectation of B_i in Equation (EC.1) can be written as

$$\begin{aligned}\mathbb{E}\left[B_i\right] &= \mathbb{E}\left[\widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\mathbb{E}\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right] \\ &\approx \left(\sum_{k=1}^K \frac{1}{k!}\sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right)\right)\times\left(\sum_{k=1}^K \frac{1}{k!}\sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i,\mathcal{X}^m}\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]\right)\right).\end{aligned}$$

Finally, the expectation of C_i in Equation (EC.1) can be written as

$$\begin{aligned}\mathbb{E}\left[C_i\right] &= \mathbb{E}\left[\ell^w(\widetilde{\mathbf{x}}_{\text{new}}|\mathcal{D}^\ell)\right]\mathbb{E}\left[\widehat{m}^w(\widetilde{\mathbf{X}}_i|\widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right] \\ &\approx \mathbb{E}\left[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}},\mathcal{D}^\ell)\right]\cdot\sum_{k=1}^K \frac{1}{k!}\sigma_k \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i,\mathcal{X}^m}\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]\right).\end{aligned}$$

Next, we try to isolate the impact of imperfect measurement of $\widetilde{\mathbf{x}}_{\text{new}}$ when we have access to the oracle CATE function $\tau(\widetilde{\mathbf{x}}_{\text{new}})$. Note that by the unbiasedness assumption of $\widehat{\tau}$, we have

$$\begin{aligned}\tau(\mathbf{x}_{\text{new}}) &= \mathbb{E}\left[\widehat{\tau}(\mathbf{x}_{\text{new}})\right] \\ &= \sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}\left[\widehat{\ell}_i^1(\mathbf{x}_{\text{new}},\mathcal{D}^\ell)\right]\left(\mathbb{E}\left[Y_i(1) - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)\right]\right) - \\ &\qquad \sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}\left[\widehat{\ell}_i^0(\mathbf{x}_{\text{new}},\mathcal{D}^\ell)\right]\left(\mathbb{E}\left[Y_i(0) - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)\right]\right).\end{aligned}$$

As a result, we have

$$\begin{aligned}\partial^k_{\mathbf{x}_{\text{new}}}\tau(\mathbf{x}_{\text{new}}) = &\sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}}\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\mathbb{E}\left[Y_i(1) - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)\right] - \\ &\sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}}\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\mathbb{E}\left[Y_i(0) - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)\right].\end{aligned}$$

Using the fact that

$$\text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right) = \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{x}_{\text{new}}}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right) + \text{trace}\left(\mathbb{E}\left[\partial^k_{\mathcal{X}^\ell}\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right),$$

we can further write the bias caused by A_i as follows:

$$\begin{aligned}
&\sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}[A_i]-\sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}[A_i]\\
&\quad\approx\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\text{trace}\left(\partial^k_{\mathbf{x}_{\text{new}}}\tau(\mathbf{x}_{\text{new}})\right)+\sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}\left[Y_i-\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]\cdot\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\text{trace}\left(\mathbb{E}\left[\partial^k_{\mathcal{X}^\ell}\widehat{\ell}^w_i(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right)-\\
&\qquad\sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}\left[Y_i-\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]\cdot\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\text{trace}\left(\mathbb{E}\left[\partial^k_{\mathcal{X}^\ell}\widehat{\ell}^w_i(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right).
\end{aligned}$$

Combining all of them together, we can write the bias of the predicted CATE as

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}})\right]-\tau(\mathbf{x}_{\text{new}})\mathbb{E}\left[\widetilde{\widehat{\mu}}^1(\widetilde{\mathbf{x}}_{\text{new}})-\widehat{\mu}^1(\mathbf{x}_{\text{new}})\right]-\mathbb{E}\left[\widetilde{\widehat{\mu}}^0(\widetilde{\mathbf{x}}_{\text{new}})-\widehat{\mu}^0(\mathbf{x}_{\text{new}})\right]\\
&\quad=\sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}[A_i-B_i-C_i]-\sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}[A_i-B_i-C_i]\\
&\quad=\Delta^\tau(\mathbf{x}_{\text{new}})+\\
&\qquad\sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}\left[Y_i-\widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)\right]\cdot\Delta_i^{\ell^1}(\mathbf{x}_{\text{new}})-\sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}\left[Y_i-\widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)\right]\cdot\Delta_i^{\ell^0}(\mathbf{x}_{\text{new}})-\\
&\qquad\sum_{i\in\mathcal{I}^o:\ W_i=1}\mathbb{E}\left[\widehat{\ell}^1_i(\mathbf{x}_{\text{new}},\mathcal{D}^\ell)\right]\cdot\Delta_i^{m^1}(\mathbf{x}_{\text{new}})-\sum_{i\in\mathcal{I}^o:\ W_i=0}\mathbb{E}\left[\widehat{\ell}^0_i(\mathbf{x}_{\text{new}},\mathcal{D}^\ell)\right]\Delta_i^{m^0}(\mathbf{x}_{\text{new}})-\\
&\qquad\sum_{i\in\mathcal{I}^o:\ W_i=1}\Delta_i^{\ell^1}\Delta_i^{m^1}-\sum_{i\in\mathcal{I}^o:\ W_i=0}\Delta_i^{\ell^0}\Delta_i^{m^0},
\end{aligned}$$

where $\Delta^\tau(\mathbf{x}_{\text{new}})=\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\text{trace}\left(\partial^k_{\mathbf{x}_{\text{new}}}\tau(\mathbf{x}_{\text{new}})\right)$, $\Delta_i^{\ell^w}(\mathbf{x}_{\text{new}})=\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\text{trace}\left(\mathbb{E}\left[\partial^k_{\mathcal{X}^\ell}\widehat{\ell}^w_i(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\right]\right)$, and $\Delta_i^{m^w}(\mathbf{x}_{\text{new}})=\sum_{k=1}^{K}\frac{1}{k!}\sigma_k\text{trace}\left(\mathbb{E}\left[\partial^k_{\mathbf{X}_i,\mathcal{X}^m}\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)\right]\right)$.

EC.1.2. Proof for Theorem 2

Proof: By Law of Total Variance, we have the following variance decomposition:

$$\begin{aligned}
&\text{Var}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)]\\
&\quad=\mathbb{E}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\text{Var}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m]\right\}+\\
&\qquad\text{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\mathbb{E}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m]\right\},
\end{aligned}\tag{EC.2}$$

where $\boldsymbol{\eta}$ denotes the noises injected into the covariates.

First, note that the Taylor approximation of $\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)$ is

$$\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)\approx\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)+\sum_{k=1}^{K}\frac{1}{k!}\left[\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^o,\mathcal{X}^\ell,\mathcal{X}^m}\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right]\boldsymbol{\eta}^{\circ k},$$

For simplicity, we use ∂^k in the rest of the proof to denote $\partial^k_{\mathbf{x}_{\text{new}},\mathcal{X}^o,\mathcal{X}^\ell,\mathcal{X}^m}$.

Then, we have

$$\begin{aligned}
&\mathrm{Var}_\eta[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m] \approx \mathrm{Var}_\eta\left[\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)+\sum_{k=1}^K\frac{1}{k!}\left[\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right]\boldsymbol{\eta}^{\circ k}\right]\\
&\quad=\sum_{k=1}^K\frac{1}{k!}\mathrm{Var}\left[\eta_{i,p}^k\right]\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right).
\end{aligned}$$

Therefore, the first term in Equation EC.2 can be written as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\mathrm{Var}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m]\right\}\\
&\quad\approx\sum_{k=1}^K\frac{1}{k!}\mathrm{Var}\left[\eta_{i,p}^k\right]\mathbb{E}\left[\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right)\right].
\end{aligned}$$

Similarly, we can write the second term as

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\mathbb{E}_{\boldsymbol{\eta}}[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m]\right\}\\
&\quad\approx\mathrm{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)+\sum_{k=1}^K\frac{1}{k!}\sigma_k\left[\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right])\right\}\\
&\quad=\mathrm{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left[\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right]+\mathrm{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left[\sum_{k=1}^K\frac{1}{k!}\sigma_k\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right)\right]+\\
&\qquad2\sum_{k=1}^K\frac{1}{k!}\sigma_k\mathrm{Cov}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m),\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right)\right\}\\
&\quad=\mathrm{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left[\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right]+\sum_{k=1}^K\frac{1}{(k!)^2}\sigma_k^2\mathrm{Var}\left[\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right)\right]+\\
&\qquad\sum_{k_1\neq k_2}\frac{1}{k_1!}\frac{1}{k_2!}\sigma_{k_1}\sigma_{k_2}A_{k_1,k_2}+2\sum_{k=1}^K\frac{1}{k!}\sigma_kB_k,
\end{aligned}$$

where $A_{k_1,k_2}=\mathrm{Cov}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left[\mathrm{trace}\left(\partial^{k_1}\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right),\mathrm{trace}\left(\partial^{k_2}\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right)\right]$ and $B_k=\mathrm{Cov}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left\{\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m),\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right)\right\}$.

Combining these two terms together, we have

$$\begin{aligned}
&\mathrm{Var}\left[\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}|\widetilde{\mathbf{x}}_{\text{new}},\widetilde{\mathcal{D}}^o,\widetilde{\mathcal{D}}^\ell,\widetilde{\mathcal{D}}^m)\right]\\
&\quad=\mathrm{Var}_{\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m}\left[\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right]+\\
&\qquad\sum_{k=1}^K\frac{1}{k!}\mathrm{Var}\left[\eta_{i,p}^k\right]\mathbb{E}\left[\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}})^{\circ 2}\right)\right]+\\
&\qquad\sum_{k=1}^K\frac{1}{(k!)^2}\sigma_k^2\mathrm{Var}\left[\mathrm{trace}\left(\partial^k\widehat{\tau}(\mathbf{x}_{\text{new}}|\mathcal{D}^o,\mathcal{D}^\ell,\mathcal{D}^m)\right)\right]+\\
&\qquad\sum_{k_1\neq k_2}\frac{1}{k_1!}\frac{1}{k_2!}\sigma_{k_1}\sigma_{k_2}A_{k_1,k_2}+2\sum_{k=1}^K\frac{1}{k!}\sigma_kB_k.
\end{aligned}$$

EC.1.3. Proof for Theorem 3.1

Proof: Note that by the unbiasedness assumption of $\widehat{\tau}$, $\widehat{\tau}_{\widetilde{Y}}(\mathbf{x})$ is an unbiased estimator of

$$\begin{aligned}
&\mathbb{E}[\widetilde{Y}_i|\mathbf{X}_i=\mathbf{x},W_i=1]-\mathbb{E}[\widetilde{Y}_i|\mathbf{X}_i=\mathbf{x},W_i=0]\\
&\quad=\mathbb{E}[Y_i+\eta_i|\mathbf{X}_i=\mathbf{x},W_i=1]-\mathbb{E}[Y_i+\eta_i|\mathbf{X}_i=\mathbf{x},W_i=0]\\
&\quad=\mathbb{E}[Y_i|\mathbf{X}_i=\mathbf{x},W_i=1]-\mathbb{E}[Y_i|\mathbf{X}_i=\mathbf{x},W_i=0]+(\mathbb{E}[\eta_i|\mathbf{X}_i=\mathbf{x},W_i=1]-\mathbb{E}[\eta_i|\mathbf{X}_i=\mathbf{x},W_i=1])\\
&\quad=\mathbb{E}[Y_i|\mathbf{X}_i=\mathbf{x},W_i=1]-\mathbb{E}[Y_i|\mathbf{X}_i=\mathbf{x},W_i=0]=\tau(\mathbf{x}).
\end{aligned}$$

Therefore, $\widehat{\tau}_{\widetilde{Y}}(\mathbf{x})$ is also an unbiased estimator of $\tau(\mathbf{x})$.

EC.1.4. Proof for Theorem 3.2

To prove Theorem 3.2 in the presence of adjustment functions, we need to impose the following assumption on the class of them to bound their variance.

Assumption EC.3. *The adjustment function, $\widehat{m}^w(\mathbf{x}|\mathcal{D}^m)$, can be written as $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)=\sum_{j\in\mathcal{D}^m}S_j(\mathbf{x})Y_j$, where S_j is (i) independent of any outcome information or (ii) can be represented as a potentially n-nearest-neighbors estimator as described in Wager and Athey (2018).*

Note that (i) includes methods such as kernel regression and nearest neighbor models, while (ii) is specifically designed for tree-based methods.

Proof for Theorem 3.2 Since the experiment data are i.i.d. samples, the variance of the predicted CATEs can be written as

$$\begin{aligned}
&\mathrm{Var}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\text{new}})]\\
&\quad=\sum_{i\in\mathcal{I}^o:W_i=1}\mathrm{Var}\left\{\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)[Y_i+\eta_i-\widehat{m}^1(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]\right\}+\\
&\qquad\sum_{i\in\mathcal{I}^o:W_i=0}\mathrm{Var}\left\{\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)[Y_i+\eta_i-\widehat{m}^0(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]\right\}.
\end{aligned}$$

Note that for each individual i, we can write its variance as

$$\begin{aligned}
&\mathrm{Var}\{\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]\}\\
&\quad=\mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)]^2\right\}\mathrm{Var}[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]+\\
&\qquad\mathrm{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)]\mathbb{E}^2\left[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)\right]
\end{aligned}\tag{EC.3}$$

Now, our goal is to bound the variance by showing (a) the variance of $\widehat{m}^w(\mathbf{X}_i|\mathcal{D}^m)]$ scale with $\mathrm{Var}[\eta_i]$, (b) the expectation of the adjustment function do not scale with $\mathrm{Var}[\eta_i]$, and (iii) terms related to $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell)$ do not scale with $\mathrm{Var}[\eta_i]$.

To show (a), note that the variance of the adjustment function can be written as:

$$\mathrm{Var}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] = \sum_{j\in\mathcal{D}^m} \mathrm{Var}[S_j(\mathbf{X}_i)(Y_j+\eta_j)] =$$

$$\sum_{j\in\mathcal{D}^m} \mathrm{Var}[S_j(\mathbf{X}_i)Y_j] + \sum_{j\in\mathcal{D}^m} \mathrm{Var}[S_j(\mathbf{X}_i)]\mathrm{Var}[\eta_j]$$

$$\sum_{j\in\mathcal{D}^m} \mathrm{Var}[S_j(\mathbf{X}_i)Y_j] + \left\{ \mathrm{Var}[S_j(\mathbf{X}_i)] \underbrace{\mathbb{E}[\eta_j^2]}_{=\mathrm{Var}[\eta_j]} + \mathbb{E}^2[S_j(\mathbf{X}_i)]\mathrm{Var}[\eta_j] \right\} = \Theta(\mathrm{Var}[\eta_j]).$$

For Assumption EC.3 (i), since $S_j(\mathbf{X}_i)$ is independent of the outcome (and therefore η_j), all the terms with respect to $S_j(\mathbf{X}_i)$ in constant in $\mathrm{Var}[\eta_j]$. Therefore, $\mathrm{Var}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] = \Theta(\mathrm{Var}[\eta_j])$. For Assumption EC.3 (ii), by the fact from Wager and Athey (2018) that $\frac{1}{2(n-1)|\mathcal{D}^m|} \lesssim \mathrm{Var}[S_j(\mathbf{X}_i) \lesssim \frac{1}{(n-1)|\mathcal{D}^m}$ and $\mathbb{E}[S_j(\mathbf{X}_i)] \leqslant \frac{1}{n}$, we can see that $\mathrm{Var}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] = \Theta(\mathrm{Var}[\eta_j])$.

To show (b), note that the expectation of the adjustment function satisfying Assumption EC.3 (i) is

$$\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] = \sum_{j\in\mathcal{D}^m} \mathbb{E}[S_j(\mathbf{X}_i)(Y_j+\eta_j)] = \sum_{j\in\mathcal{D}^m} \mathbb{E}[S_j(\mathbf{X}_i)]\mathbb{E}[Y_j+\eta_j] = \sum_{j\in\mathcal{D}^m} \mathbb{E}[S_j(\mathbf{X}_i)](\mathbb{E}[Y_j]+\mathbb{E}[\eta_j]).$$

Therefore, $\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] = \Theta(\mathbb{E}[\eta_j])$. For the adjustment function satisfying Assumption EC.3 (ii), since $0 \leqslant S_j(\mathbf{X}_i) \leqslant 1/|\mathcal{D}^m|$, we have

$$\underbrace{-\sum_{j\in\mathcal{D}^m} \frac{1}{|\mathcal{D}^m|}|\mathbb{E}[Y_j+\eta_j]|}_{=-\sum_{j\in\mathcal{D}^m}\frac{1}{|\mathcal{D}^m|}|\mathbb{E}[Y_j]|} \leqslant \mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] = \sum_{j\in\mathcal{D}^m} \mathbb{E}[S_j(\mathbf{X}_i)(Y_j+\eta_j)] \leqslant \underbrace{\sum_{j\in\mathcal{D}^m} \frac{1}{|\mathcal{D}^m|}|\mathbb{E}[Y_j+\eta_j]|}_{=\sum_{j\in\mathcal{D}^m}\frac{1}{|\mathcal{D}^m|}|\mathbb{E}[Y_j]|}$$

Therefore, it does not scale with $\mathbb{E}[\eta_j]$ or $\mathrm{Var}[\eta_j]$.

Now, we consider the case when the weight function does not use the outcome information. In this case, we have $\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\widetilde{\mathcal{D}}^\ell) = \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)$. We can write the variance as:

$$\mathrm{Var}\{\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]\}$$

$$= \underbrace{\mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]^2\right\}}_{\equiv C_1^w} \Big(\mathrm{Var}[Y_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] + \mathrm{Var}[\eta_i]\Big) +$$

$$\underbrace{\mathrm{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)]}_{\equiv C_2^w} \mathbb{E}[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)].$$

Note that

$$\mathrm{Var}[Y_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]$$

$$= \mathrm{Var}[Y_i] + \mathrm{Var}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)] - 2\sqrt{\mathrm{Var}[Y_i]}\sqrt{\mathrm{Var}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]}\mathrm{Cor}[Y_i, \widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]$$

$$= \Theta(\mathrm{Var}[\eta_i]) - \Theta(\sqrt{\mathrm{Var}[\eta_i]}) = \Theta(\mathrm{Var}[\eta_i]).$$

Similarly, we have

$$\mathbb{E}^2[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]=\mathbb{E}^2[Y_i]+\mathbb{E}^2[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]+\mathbb{E}^2[\eta_i]+$$
$$-2\mathbb{E}[Y_i]\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]-2\underbrace{\mathbb{E}[\eta_i]\mathbb{E}[\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]}_{=\Theta(\mathbb{E}^2[\eta_i])}+2\mathbb{E}[Y_i]\mathbb{E}[\eta_i].$$

Note that this term does not scale with $\mathrm{Var}[\eta_i]$ as $\mathbb{E}[\eta_i]=0$ the Laplace or Gaussian mechanisms. For response randomization, this term is $\Theta(\mathrm{Var}[\eta_i])$ as $\mathbb{E}[\eta_i]=\Theta(\mathrm{Var}[\eta_i])$ and $\mathbb{E}^2[\eta_i]=\Theta(\mathrm{Var}[\eta_i])$. As a result, we have $\mathrm{Var}\{\widehat{\ell}_i^w(\mathbf{x}_{\mathrm{new}}|\mathcal{D}^\ell)[Y_i+\eta_i-\widehat{m}^w(\mathbf{X}_i|\widetilde{\mathcal{D}}^m)]\}=\Theta(\mathrm{Var}[\eta_i])$, which implies that $\mathrm{Var}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\mathrm{new}})]=\Theta(\mathrm{Var}[\eta_i])$.

For the case when the weight depends on the outcome information, we only consider tree-based models, such as Causal Forest as it is the most common, if not the only, case in which the weight depends on the outcome information. Note that for this class of models, the induced outcome models are also potentially n-nearest neighbors estimators for some n (as shown by Wager and Athey (2018)). Therefore, we have

$$\frac{1}{2(n-1)|\mathcal{D}^\ell|}\lesssim\mathrm{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\mathrm{new}}|\mathcal{D}^\ell)]\lesssim\frac{1}{(n-1)|\mathcal{D}^\ell|}\quad\text{and}\quad 0\leqslant\mathbb{E}[\widehat{\ell}_i^w(\mathbf{x}_{\mathrm{new}}|\mathcal{D}^\ell)]\leqslant\frac{1}{|\mathcal{D}^\ell|}.$$

As a result, $\mathbb{E}\left\{[\widehat{\ell}_i^w(\mathbf{x}_{\mathrm{new}}|\widetilde{\mathcal{D}}^\ell)]^2\right\}$ and $\mathrm{Var}[\widehat{\ell}_i^w(\mathbf{x}_{\mathrm{new}}|\widetilde{\mathcal{D}}^\ell)]$ in Equation (EC.3) will not scale with the variance of the noise. On the other hand, the remaining terms in Equation (EC.3) are $\Theta(\mathrm{Var}[\eta_i])$ functions based on the above analysis. Consequently, we can conclude that $\mathrm{Var}[\widehat{\tau}_{\widetilde{Y}}(\mathbf{x}_{\mathrm{new}})]=\Theta(\mathrm{Var}[\eta_i])$.

EC.1.5. Proof for Proposition 1

Proof for Proposition 1.1: By Taylor approximation, we can write

$$\Delta_i^{\ell,w} \equiv \widehat{\ell}_i^w(\widetilde{\mathbf{x}}_{\text{new}}, \widetilde{\mathcal{D}}^\ell) - \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) \approx \sum_{k=1}^{K} \left(\frac{1}{k!} \partial^k_{\mathbf{x}_{\text{new}}, \mathcal{X}^\ell} \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) \right) \boldsymbol{\eta}^{\circ k}_{\mathbf{x}_{\text{new}}, \mathcal{X}^\ell},$$

$$\Delta_i^{m,w} \equiv \widehat{m}^w(\widetilde{\mathbf{X}}_i, \widetilde{\mathcal{D}}_m) - \widehat{m}^w(\mathbf{X}_i, \mathcal{D}^m) \approx \sum_{k=1}^{K} \left(\frac{1}{k!} \partial^k_{\mathbf{X}_i, \mathcal{X}^m} \widehat{m}^w(\mathbf{X}_i, \mathcal{D}^m) \right) \boldsymbol{\eta}^{\circ k}_{\mathbf{X}_i, \mathcal{X}^m}.$$

Then, the predicted CATE can be written as:

$$\begin{aligned}\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}}) = \widehat{\tau}(\mathbf{x}_{\text{new}}) + \\ \sum_{i \in \mathcal{I}^o : W_i = 1} \Delta_i^{\ell,1}[Y_i - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)] - \sum_{i \in \mathcal{I}^o : W_i = 1} \widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\Delta_i^{m,1} - \sum_{i \in \mathcal{I}^o : W_i = 1} \Delta_i^{\ell,1}\Delta_i^{m,1} - \\ \sum_{i \in \mathcal{I}^o : W_i = 0} \Delta_i^{\ell,0}[Y_i - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)] + \sum_{i \in \mathcal{I}^o : W_i = 0} \widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)\Delta_i^{m,0} + \sum_{i \in \mathcal{I}^o : W_i = 1} \Delta_i^{\ell,0}\Delta_i^{m,0}.\end{aligned}$$

Therefore, the larger absolute values of $\boldsymbol{\eta}$, the more extreme the predicted CATE $\widehat{\tau}(\widetilde{\mathbf{x}}_{\text{new}})$ becomes. Additionally, the greater the magnitudes of $\sum_{k=1}^{K} \left(\frac{1}{k!} \partial^k \widehat{\ell}_i^w(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) \right)$ and $\sum_{k=1}^{K} \left(\frac{1}{k!} \partial^k \widehat{m}^w(\mathbf{X}_i, \mathcal{D}^m) \right)$, the more significant the influence of the injected noise on the predicted CATE will be.

Proof for Proposition 1.2 For an unseen individual with covariates $\mathbf{x}_{\text{new}}$, we can express the predicted CATE as follows:

$$\begin{aligned}\widehat{\tau}(\mathbf{x}_{\text{new}}) &= \sum_{i \in \mathcal{I}^o : W_i = 1} \widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)] - \sum_{i \in \mathcal{I}^o : W_i = 0} \widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)] \\ &= \sum_{i \in \mathcal{I}^o : W_i = 1} \widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i + \eta_i - \widehat{m}^1(\mathbf{X}_i|\mathcal{D}^m)] - \\ &\quad \sum_{i \in \mathcal{I}^o : W_i = 0} \widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell)[Y_i + \eta_i - \widehat{m}^0(\mathbf{X}_i|\mathcal{D}^m)].\end{aligned}$$

Therefore, for $i \in \mathcal{D}^o$:

1. For those who are treated (i.e., $W_i = 1$), the injected noise η_i is positively correlated with $\widehat{\tau}(\mathbf{x}_{\text{new}})$ if $\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) > 0$ and negatively correlated if $\widehat{\ell}_i^1(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) < 0$.
2. For those who are untreated (i.e., $W_i = 0$), the injected noise η_i is positively correlated with $\widehat{\tau}(\mathbf{x}_{\text{new}})$ if $\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) < 0$ and negatively correlated if $\widehat{\ell}_i^0(\mathbf{x}_{\text{new}}|\mathcal{D}^\ell) > 0$.

Therefore, when η_i is extreme, the predicted CATE will also be extremely high or low.

EC.2. Further Details about the Simulation Analyses

In this appendix, we provide implementation details about the simulation analyses described in Section 5 of the main document and present robustness checks.

EC.2.1. Details on Model Specifications

Constructing the CATE proxy

To construct Robinson's transformation, we use a propensity score ($\widehat{e}$) of 0.5 as the treatment assignment is completely randomized. We evaluate three candidate models for the conditional mean ($\widehat{m}$) : linear regression, linear regression with interactions, and regression forest. We use 1,000 individuals (500 treated and 500 non-treated) as the training set and 2,000 individuals (1,000 treated and 1,000 non-treated) as the holdout evaluation set. This is reflective of the sample size when performing model calibration for an experiment set with 3,000 individuals. Table EC.1 presents the mean squared error of each model in predicting Y_i. Given that linear regression with interactions results in the smallest prediction error, we select it as our preferred model for Robinson's transformation. During the calibration process, we also utilize a linear regression with interactions to construct calibrators. Note that the model used for the conditional mean in the Robinson's transformation as well as the calibration process do not follow the same specification as the data generating process. Yet, the algorithm is still able to reduce the MSE of the CATE predictions.

Table EC.1 MSE of Conditional Mean Outcome Models: Simulation Data

Scenario 1: LDP-Protected Covariates

Privacy	Linear Regression	Linear Regression with Interactions	Random Forest
No	146.1	80.2	110.1
Very Low	161.5	82.6	117.7
Low	171.4	99.1	142.3
Medium	174.3	120.8	148.1
High	204.2	144.4	177.3
Very High	215.0	168.7	195.8

Scenario 2: LDP-Protected Outcome

Privacy	Linear Regression	Linear Regression with Interactions	Random Forest
No	146.1	80.2	110.1
Very Low	170.9	93.1	135.8
Low	179.3	110.7	144.6
Medium	223.4	151.4	198.4
High	272.7	204.4	241.1
Very High	338.6	285.3	326.6

Constructing the initial CATE model

Regarding the initial CATE model, we consider three different types of CATE estimation methods. These include:

1. *Causal Forest* (Wager and Athey 2018): which is the model used in all results presented in the main document. We implement causal forest using the `grf` package with default parameters.
2. *T-learner* (Künzel et al. 2019): We also evaluate the results when using t-learner; these are presented in Electronic Companion EC.2.3.1. To do so, we construct two separate regression models to estimate $\mathbb{E}[Y_i|W_i, \widetilde{\mathbf{X}}_i]$, one for the treatment group and another for the control group, and take the difference between the predicted outcomes as our predicted CATEs. (We use random forests with default parameters as provided in the `grf` package when estimating each (separate) regression model.)
3. *R-learner* (Nie and Wager 2021): We also evaluate the performance of our solution when using R-learner as initial CATE; results are presented in Electronic Companion EC.2.3.1. We implement the R-learner following Nie and Wager (2021), with two-fold cross-fitting, and build all the models using random forest with default parameters in the `grf` package.

EC.2.2. Determine Number of Subgroups and Number of Iterations

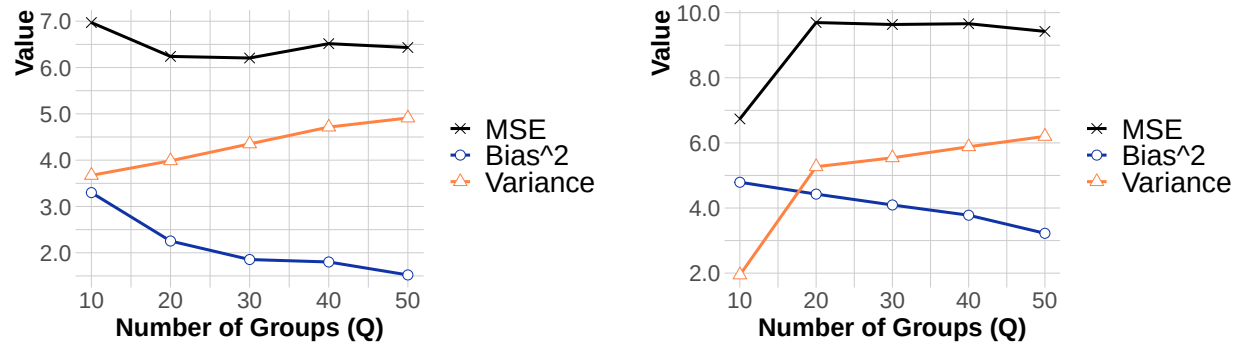
In this section, we investigate the sensitivity of our solution to the pre-specified number of subgroups (Q) and the number of iterations (R). We start by illustrating the bias-variance trade-offs for the number of subgroups. Figure EC.1 displays the predictive error metrics when applying the proposed solution with varying values of Q. As we increase the number of subgroups, the bias decreases for both scenarios, while the variance correspondingly increases. This outcome showcases a classic bias-variance trade-off commonly observed in machine learning models: increasing model complexity may enhance precision but concurrently introduce additional variability. Notably, when the outcome is protected by LDP, changes in Q do not significantly affect overall accuracy. On the other hand, when the covariates are protected by LDP, using the smallest value of Q achieves the lowest MSE.

Figure EC.2 illustrates the step size $\rho_{q^\star}^{[r]}$ in each iteration (capped at 50 iterations) when applying the proposed solution with varying numbers of subgroups (i.e., varying values of Q). Generally, the step sizes approach zero before the completion of the first Q iterations, regardless of the chosen value for Q. This observation suggests that performing $R = Q$ iterations is typically sufficient for the proposed solution, thereby providing a practical guideline for setting this hyperparameter.

EC.2.3. Robustness Checks for Different Initial CATE models

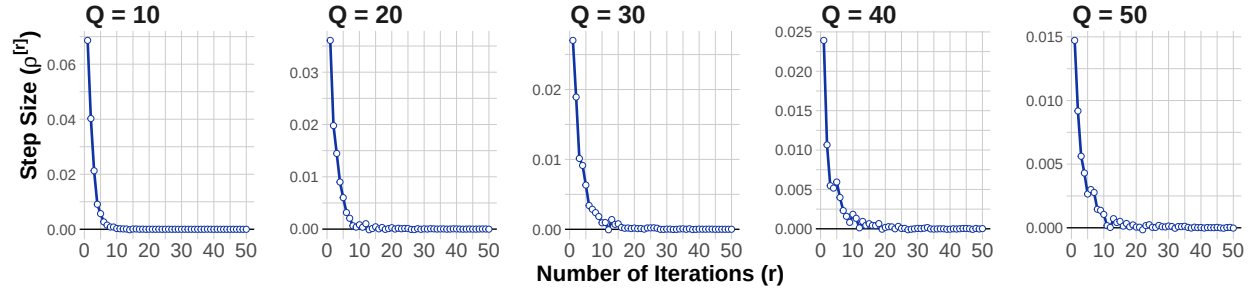
As described in Section EC.2.1, we evaluate the performance of three different approaches for the initial CATE model. The main manuscript includes the results using Causal Forest and this appendix presents the results for the other two methods. Overall, Causal Forests result in the highest predictive accuracy and best targeting performance. In addition, regardless of the initial CATE models, our proposed solution yields to the smallest predictive error and best targeting performance with varying privacy levels and experiment sample sizes.

Figure EC.1 Predictive Errors of Proposed Method with Varying Numbers of Subgroups



Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of Causal Forest as the initial CATE model.

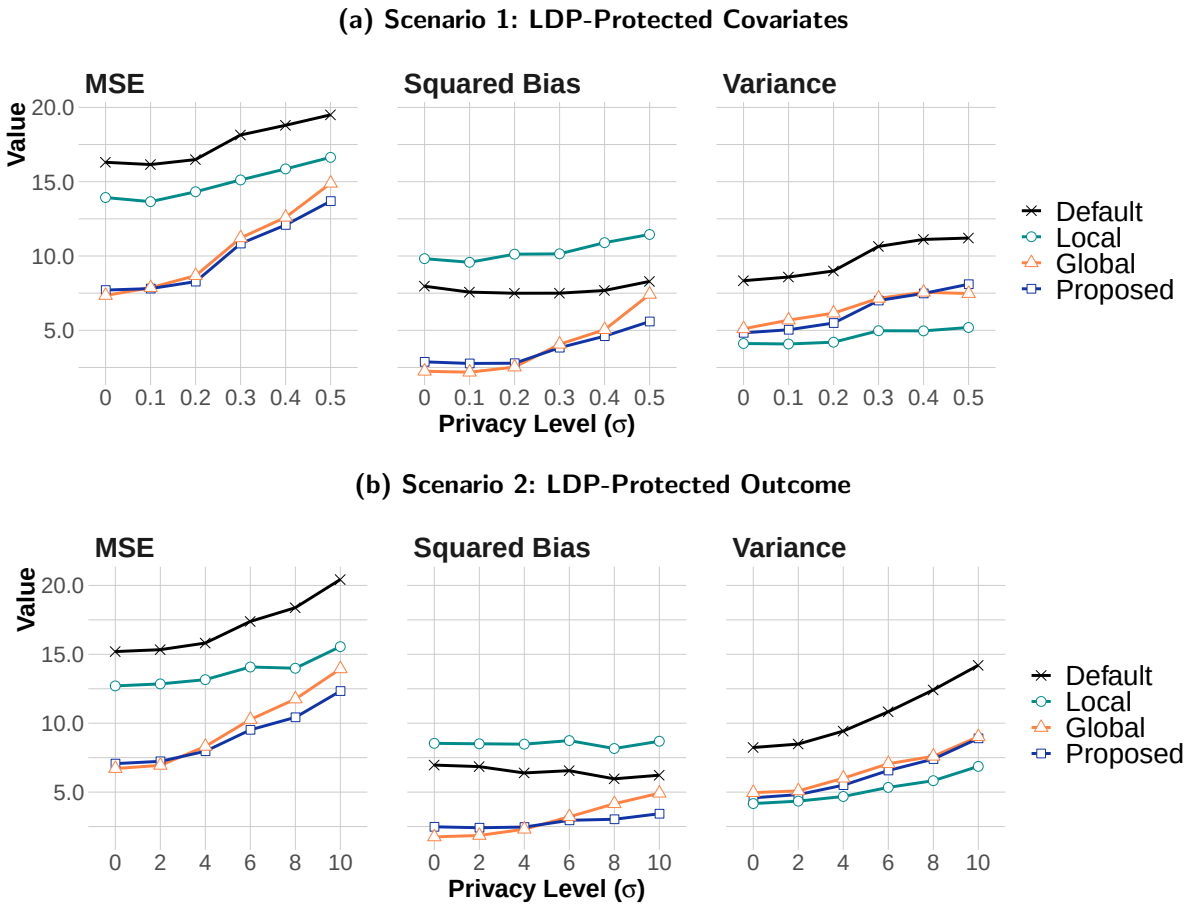
Figure EC.2 Step Size in Each Iteration



Note: Each point is calculated by averaging the mean step size over 100 bootstrap replications. We present results from the case when the outcome is protected by LDP with $\sigma = 10$, but the pattern remains the same across different scenarios. We use Causal Forest as the initial CATE model.

EC.2.3.1. T-learner. First, we present the predictive errors of various methods when using T-learner as the initial CATE model. (The corresponding results for causal forest are presented in Figure 2 and Table 1 of the main manuscript.) As seen in Figure EC.3, our proposed method exhibits similar performance to the global approach when privacy levels are relatively low, and outperforms the global approach at higher privacy levels. This observation is also reflected in Table EC.2, which reports the AUTOC values of different approaches at varying privacy levels. In addition, all the post-processing methods yield smaller predictive errors and higher AUTOC values than the default approach across all privacy levels. This highlights the value of iteratively refining the T-learner model using simpler models.

Figure EC.3 Predictive Errors of Different Methods Across Varying Privacy Levels: T-learner



Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

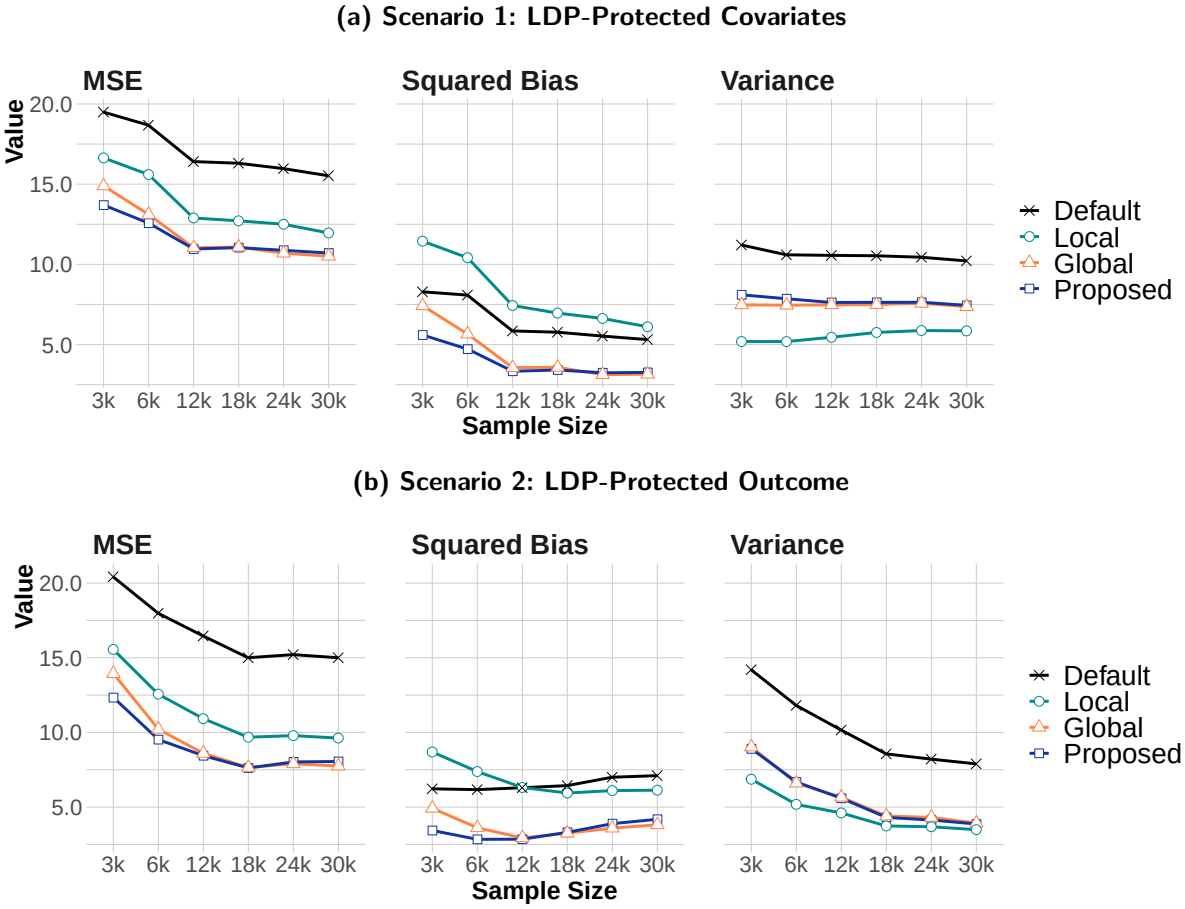
Table EC.2 AUTOC Values of Different Methods Across Varying Privacy Levels: T-learner

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**4.09 (100%)**	4.14 (100%)	3.49 (90%)	3.34	**4.09 (100%)**	4.14 (100%)	3.49 (90%)	3.34
Very Low	**4.04 (100%)**	4.06 (100%)	3.46 (88%)	3.28	**4.08 (100%)**	4.10 (100%)	3.44 (88%)	3.30
Low	**4.01 (100%)**	4.01 (100%)	3.39 (85%)	3.25	**3.97 (100%)**	3.96 (100%)	3.39 (85%)	3.25
Medium	**3.56 (100%)**	3.52 (99%)	3.08 (90%)	2.91	**3.90 (100%)**	3.84 (99%)	3.36 (93%)	3.18
High	**3.53 (100%)**	3.47 (98%)	3.03 (86%)	2.89	**3.78 (100%)**	3.64 (98%)	3.32 (96%)	3.09
Very High	**3.27 (100%)**	3.12 (96%)	2.83 (88%)	2.67	**3.63 (100%)**	3.43 (96%)	3.18 (92%)	2.96

Note: We report the average of 100 simulation replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses).

Figure EC.4 displays the key error metrics of various approaches, constructed using different experiment data sizes. (The corresponding results for causal forest are presented in Figure 3 and Table 2 of the main manuscript.) When the sample size is small, our proposed solution significantly outperforms the global method. However, the performance of both methods converges as the sample size increases. This underscores the value of local learning as a method to identify the most informative individuals, especially in scenarios with small sample sizes and high noise. This observation is further corroborated by the AUTOC values in Table EC.3.

Figure EC.4 Predictive Errors of Different Methods with Varying Sample Sizes: T-learner



Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table EC.3 AUTOC Values of Different Methods Across Varying Experiment Sample Sizes: T-learner

Sample	Scenario 1: LDP-Protected Covariates ($\sigma = 0.5$)				Scenario 2: LDP-protected Outcome ($\sigma = 10$)			
Size	Proposed	Global	Local	Default	Proposed	Global	Local	Default
3,000	**3.27 (100%)**	3.12 (96%)	2.83 (88%)	2.67	**3.63 (100%)**	3.43 (96%)	3.18 (92%)	2.96
6,000	**3.68 (100%)**	3.62 (100%)	3.28 (100%)	3.08	**3.95 (100%)**	3.86 (100%)	3.58 (100%)	3.3
12,000	**3.64 (100%)**	3.62 (100%)	3.36 (100%)	3.17	**4.06 (100%)**	4.01 (100%)	3.69 (100%)	3.44
18,000	**3.68 (100%)**	3.67 (100%)	3.44 (100%)	3.26	**4.10 (100%)**	4.07 (100%)	3.81 (100%)	3.56
24,000	**3.72 (100%)**	3.72 (100%)	3.51 (100%)	3.33	**4.12 (100%)**	4.11 (100%)	3.86 (100%)	3.62
30,000	**3.76 (100%)**	3.77 (100%)	3.52 (100%)	3.34	**4.12 (100%)**	4.11 (100%)	3.86 (100%)	3.64

Note: We report the average of 100 simulation replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses).

EC.2.3.2. R-learner. Figure EC.5 presents the MSE, squared bias, and variance of different approaches at varying privacy levels, while Table EC.4 reports the corresponding AUTOC values. (The corresponding results for causal forest are presented in Figure 2 and Table 1 of the main manuscript.) The results suggest two main points: (i) our proposed solution yields the smallest errors and best targeting performance across all privacy levels, and (ii) the effectiveness of local learning improves as the privacy level increases.

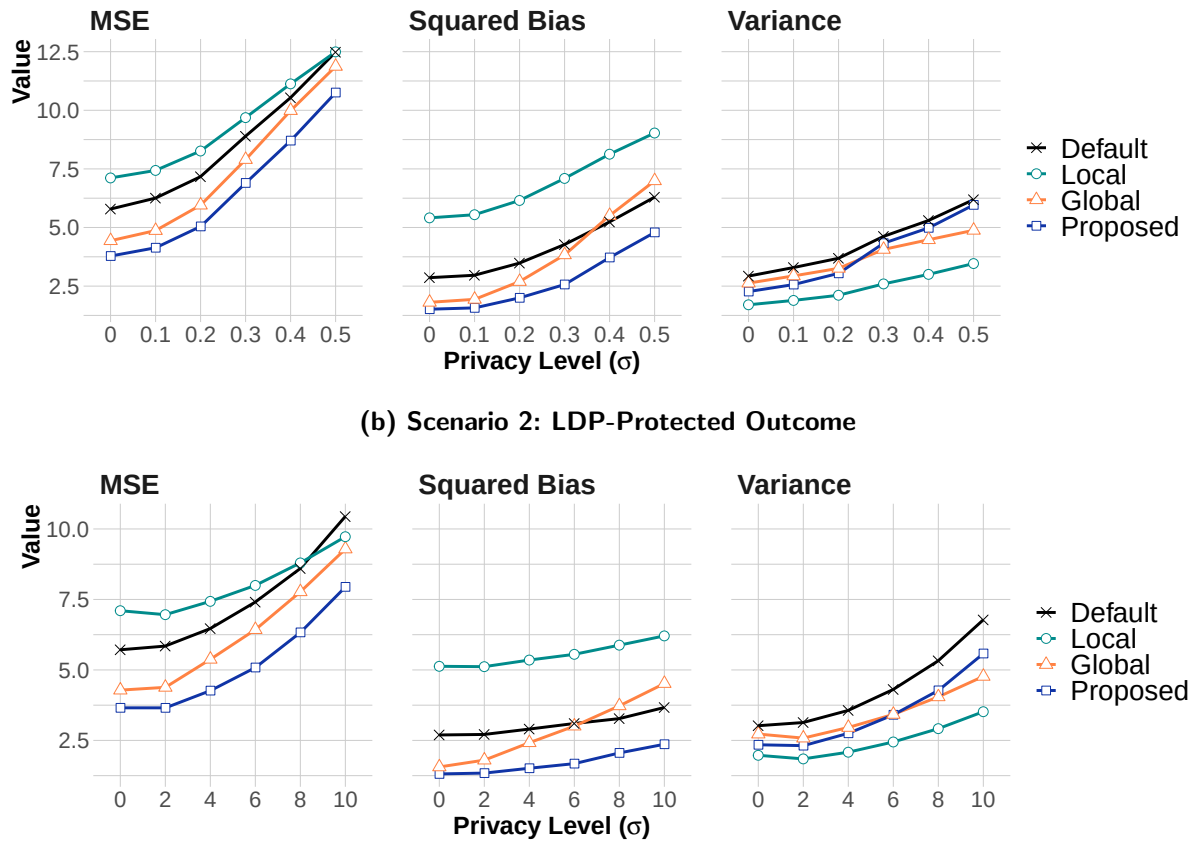
Table EC.4 AUTOC Values of Different Methods Across Varying Privacy Levels: R-learner

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**4.41 (100%)**	4.37 (95%)	4.27 (82%)	4.22	**4.41 (100%)**	4.37 (95%)	4.27 (82%)	4.22
Very Low	**4.28 (100%)**	4.25 (97%)	4.15 (81%)	4.10	**4.37 (100%)**	4.34 (96%)	4.22 (70%)	4.20
Low	**4.26 (100%)**	4.22 (98%)	4.13 (88%)	4.08	**4.33 (100%)**	4.28 (93%)	4.2 (92%)	4.13
Medium	**4.15 (100%)**	4.10 (94%)	4.05 (91%)	3.96	**4.28 (100%)**	4.19 (95%)	4.15 (93%)	4.05
High	**4.02 (98%)**	3.95 (89%)	3.93 (97%)	3.83	**4.15 (98%)**	4.06 (94%)	4.04 (95%)	3.90
Very High	**3.73 (96%)**	3.67 (94%)	3.68 (98%)	3.53	**4.04 (99%)**	3.96 (92%)	3.97 (99%)	3.77

Note: We report the average of 100 simulation replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses).

Below we present the results when varying the sample size. (The corresponding results for causal forest are presented in Figure 3 and Table 2 of the main manuscript.)

Figure EC.5 Predictive Errors of Different Methods Across Varying Privacy Levels: R-learner



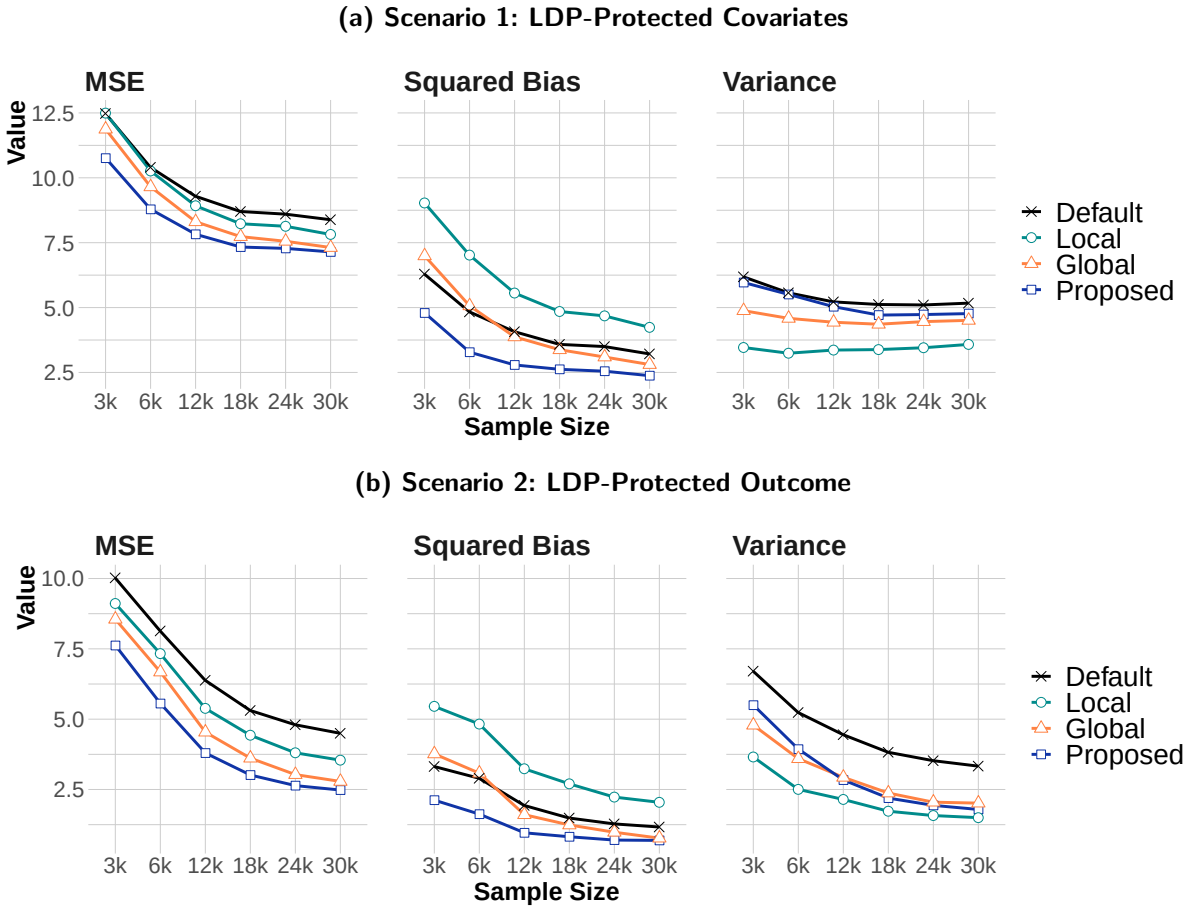
Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table EC.5 AUTOC Values of Different Methods Across Varying Experiment Sample Sizes: R-learner

Sample	Scenario 1: LDP-Protected Covariates ($\sigma = 0.5$)				Scenario 2: LDP-protected Outcome ($\sigma = 10$)			
Size	Proposed	Global	Local	Default	Proposed	Global	Local	Default
3,000	**3.73 (96%)**	3.67 (94%)	3.68 (98%)	3.53	**4.01 (100%)**	3.96 (90%)	3.96 (100%)	3.75
6,000	**3.88 (100%)**	3.81 (95%)	3.81 (100%)	3.70	**4.24 (100%)**	4.17 (98%)	4.15 (100%)	3.98
12,000	**3.92 (100%)**	3.89 (100%)	3.86 (98%)	3.77	**4.40 (100%)**	4.35 (100%)	4.32 (100%)	4.16
18,000	**3.94 (100%)**	3.92 (100%)	3.90 (100%)	3.81	**4.42 (100%)**	4.38 (100%)	4.35 (100%)	4.21
24,000	**4.10 (100%)**	4.08 (100%)	4.06 (100%)	3.98	**4.45 (100%)**	4.42 (100%)	4.38 (100%)	4.25
30,000	**4.06 (100%)**	4.04 (100%)	4.02 (100%)	3.94	**4.47 (100%)**	4.44 (100%)	4.41 (100%)	4.29

Note: We report the average of 100 simulation replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses).

Figure EC.6 Predictive Errors of Different Methods with Varying Sample Sizes: R-learner



Note: We simulate 100 replications to compute the bootstrap mean error metrics for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

EC.3. Further Details about Hillstrom Case Study

In this section, we provide the summary statistics, implementation details, and robustness checks for the Hillstrom case study in Section 6.3.

EC.3.1. Summary Statistics.

Table EC.6 presents the summary statistics for the Hillstrom data. The definitions of the pre-treatment covariates are:

1. Recency: The number of months since the customer's last purchase.
2. History: The actual dollar value the customer has spent in the past year.
3. Mens: A binary variable indicating whether the customer purchased men's merchandise in the past year (1 = Yes).
4. Womens: A binary variable indicating whether the customer purchased women's merchandise in the past year (1 = Yes).
5. Zip_Code: A classification of the customer's zip code as Urban, Suburban, or Rural
6. Newbie: A binary variable indicating whether the customer was acquired in the past twelve months (1 = Yes).
7. Channel: The channel(s) the customer purchased from in the past year.

Table EC.6 Summary Statistics for Hillstrom Data

Discrete Variables

Variable	N	Unique Values	Distributions
visit (Outcome)	42,693	2	0: 37,193, 1: 5,500
email (Treatment)	42,693	2	0: 21,306, 1: 21,387
mens	42,693	2	0: 19,166, 1: 23,527
womens	42,693	2	0: 19,260, 1: 23,433
newbie	42,693	2	0: 21,235, 1: 21,458
channel	42,693	3	Phone: 18,781, Website: 18,727, Multichannel: 5,185
zip_code	42,693	3	Suburban: 19,275, Urban 17,098, Rural: 6,320

Continuous Variables

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
history	42,693	241.71	254.04	29.99	65.16	158.46	326.05	3345.93
newbie	42,693	0.5	0.5	0	0	1	1	1
recency	42,693	5.76	3.5	1	2	5	9	12

EC.3.2. Covariate Balance Check

To assess covariate balance, we compare the distributions of each covariate for both treated and non-treated customers. In particular, we use the *standardized mean difference* measure, which is the mean difference between the treated and non-treated groups divided by the pooled standard deviation. Generally, it is considered small if the value is less than 0.20 (Cohen 2013). Table EC.7

reports the summary for the covariate balance check. Note that the standardized mean differences are close to zero for all covariates, suggesting that the experiment is properly randomized.

Table EC.7 Covariate Balance Check for Hillstrom Data

Variable	Mean Diff.	Pooled St. Dev.	Standardized Mean Diff.
recency	0.021	3.505	0.006
history	1.803	255.307	0.007
mens	−0.003	0.497	−0.007
womens	0.003	0.498	0.006
newbie	0.000	0.500	0.001
channel_Multichannel	−0.002	0.327	−0.005
channel_Phone	0.000	0.496	0.000
channel_Web	0.001	0.496	0.003
zip_code_Rural	0.003	0.356	0.009
zip_code_Suburban	−0.003	0.498	−0.006
zip_code_Urban	0.000	0.490	0.000

EC.3.3. Implementation of Local Differential Privacy

We examine two scenarios: the LDP-protected outcome and the LDP-protected covariates. In the first scenario, we apply the randomized response mechanism described in Section 6.2, setting p to values in the set $0.05, 0.10, 0.15, 0.20, 0.25$. In the second scenario, we deploy the Laplace mechanism for the `recency` variable, setting $\sigma_{\texttt{recency}}$ to values in the set $2, 4, 6, 8, 10$, and for the `history` variable, setting $\sigma_{\texttt{recency}}$ to values in the set $20, 40, 60, 80, 100$. For all discrete variables, we implement the randomized response mechanism, with p in the set $0.05, 0.10, 0.15, 0.20, 0.25$.

EC.3.4. Model Selection and Specification

We now provide details of the specifications and computational implementations for both the Robinson's transformation and various calibration methods. Regarding the Robinson's transformation, we set a constant propensity score ($\widehat{e} = 0.5$), considering that the experiment is completely randomized. As for the conditional mean model ($\widehat{m}$), we evaluate three candidate models: linear regression, logistic regression, and regression forest. The data is split into two sets: 20% allocated as the training set and the remaining 80% serving as the holdout set.[11] The mean-squared error (MSE) metric is reported in Table EC.8, calculated as $\frac{1}{N_{\text{holdout}}} \sum_{i \in \text{holdout set}} [Y_i - \widehat{\mathbb{P}}(Y_i = 1)]^2$. Linear regression is chosen as our final model since it yields the smallest MSE.

For initial CATE estimation, we consider the following models:

(a) *Causal Forest*: We use the causal forest function implemented in the `grf` package with 200 trees and other default parameters. Note that we choose 200 trees instead the default 2,000 trees to accelerate the model training process.

[11] We selected 20% customers for model training as the calibration procedure employs roughly 20% of the data to construct the conditional mean model.

Table EC.8 MSE of Conditional Mean Outcome Models: Hillstrom Data

Scenario 1: LDP-Protected Covariates

Privacy	Linear Regression	Logistic Regression	Random Forest
No	0.1082	0.1084	0.1088
Very Low	0.1082	0.1084	0.1088
Low	0.1106	0.1107	0.1111
Medium	0.1080	0.1082	0.1083
High	0.1146	0.1146	0.1152
Very High	0.1101	0.1102	0.1104

Scenario 2: LDP-Protected Outcome

Privacy	Linear Regression	Logistic Regression	Random Forest
No	0.1078	0.1080	0.1090
Very Low	0.1102	0.1102	0.1112
Low	0.1086	0.1088	0.1105
Medium	0.1157	0.1157	0.1175
High	0.1145	0.1145	0.1164
Very High	0.1160	0.1161	0.1181

(b) *T-learner*: We construct a single predictive model to estimate the outcome as a function of the treatment and covariates using random forest with 200 trees and default parameters in `grf` package). Then, we calculate the predicted CATE as the difference in predicted outcomes between treated and non-treated conditions, while keeping all other covariates constant.

(c) *R-learner*: We implement a ten-fold cross-fitted R-learner, constructing Robinson's transformation using a constant propensity score (0.5) and employing `Xgboost` as the conditional outcome model. We subsequently utilize `Xgboost` to estimate the CATE, based on Robinson's score. For all the `Xgboost` models, configure the model by setting the maximum depth of each tree to 1, selecting a learning rate of 0.5, and setting the number of iterations to 10. These parameters have been chosen because they yield the highest cross-validated accuracy for outcome predictions.

EC.3.5. Results for T-learner and R-learner

Table EC.9 presents the AUTOC values (multiplied by 100) of different methods, utilizing T-learner and R-learner as the initial CATE models. In line with previous findings, our proposed solution consistently outperforms the other methods across a broad spectrum of privacy levels. It's worth noting that R-learner exhibits greater robustness than other CATE models at high privacy levels. Consequently, at the *Very High* privacy level, the proposed method offers only a marginal improvement in the AUTOC value when compared to the default approach.

Table EC.9 AUTOC Values (Multiplied by 100) of Different Methods for Hillstron Data

(a) Results from T-learners

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**1.62 (62%)**	1.61 (61%)	1.60 (62%)	1.51	**1.62 (62%)**	1.61 (61%)	1.60 (62%)	1.51
Very Low	**1.55 (82%)**	1.52 (86%)	1.43 (78%)	1.23	**1.65 (87%)**	1.63 (84%)	1.58 (85%)	1.37
Low	**1.42 (87%)**	1.38 (74%)	1.31 (72%)	1.20	**1.46 (88%)**	1.38 (78%)	1.35 (82%)	1.14
Medium	**1.35 (88%)**	1.28 (87%)	1.20 (73%)	1.03	**1.42 (84%)**	1.34 (76%)	1.27 (68%)	1.11
High	**1.32 (83%)**	1.28 (72%)	1.16 (68%)	1.01	**1.34 (89%)**	1.25 (77%)	1.22 (68%)	1.05
Very High	**1.19 (78%)**	1.11 (72%)	1.02 (69%)	0.86	**1.29 (94%)**	1.22 (84%)	1.15 (84%)	0.91

(b) Results from R-learner

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**1.71 (79%)**	1.69 (75%)	1.68 (73%)	1.53	**1.71 (79%)**	1.69 (75%)	1.68 (73%)	1.53
Very Low	**1.51 (64%)**	1.44 (52%)	1.43 (52%)	1.40	**1.63 (58%)**	1.57 (50%)	1.59 (57%)	1.57
Low	**1.45 (56%)**	1.40 (50%)	1.38 (52%)	1.39	**1.60 (61%)**	1.57 (57%)	1.56 (57%)	1.52
Medium	**1.45 (60%)**	1.45 (52%)	1.39 (52%)	1.40	**1.53 (56%)**	1.49 (49%)	1.47 (53%)	1.51
High	**1.34 (56%)**	1.31 (50%)	1.30 (48%)	1.33	**1.52 (55%)**	1.47 (54%)	1.42 (46%)	1.47
Very High	**1.30 (51%)**	1.26 (44%)	1.20 (48%)	1.30	**1.41 (52%)**	1.35 (38%)	1.31 (46%)	1.37

Note: We report the average of AUTOC values (multiplied by 100) from 100 bootstrap replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses).

EC.4. Further Details about Starbucks Case Study

In this section, we provide the summary statistics, implementation details, and robustness checks for the Starbucks case study in Section 6.3.

EC.4.1. Summary Statistics

Table EC.10 provides summary statistics for the Starbucks data. Note that there are five categorical and two continuous pre-treatment covariates.

Table EC.10 Summary Statistics for Starbucks Data

Discrete Variables

Variable	N	Unique Values	Distribution
Purchase (Outcome)	126,184	2	0: 124,664, 1: 1,520
Promotion (Treatment)	126,184	2	0: 63,072, 1: 63,112
V1	126,184	4	0: 15,846, 1: 47,410, 2: 47,134, 3: 15,794
V4	126,184	2	1: 40,379, 2: 85,805
V5	126,184	4	1: 23,179, 2: 46,597, 3: 48,643, 4: 7,765
V6	126,184	4	1: 31,435, 2: 31,420, 3: 3,1651, 4: 31,678
V7	126,184	2	1: 37,545, 2: 88,639

Continuous Variables

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
V2	126,184	29.98	5.00	7.10	26.596	29.98	33.354	55.108
V3	126,184	0.00	1.00	−1.69	−0.91	-0.04	0.83	1.69

EC.4.2. Covariate Balance Check

Table EC.11 reports the summary statistics for the covariate balance check. The result suggests that the experiment is properly randomized as the standardized mean differences of all the covariates are close to zero.

EC.4.3. Implementation of Local Differential Privacy

In the scenario when the outcome variable is protected by LDP, we apply the randomized response mechanism, setting p to values in the set $0.05, 0.10, 0.15, 0.20, 0.25$. In the scenario when covaraites are protected by LDP, we deploy the Laplace mechanism for the `V2` variable, setting $\sigma_{\texttt{V2}}$ to values in the set $5, 10, 15, 20, 25$, and for the `V3` variable, setting $\sigma_{\texttt{V3}}$ to values in the set $1, 2, 3, 4, 5$. For all discrete variables, we implement the randomized response mechanism, with p in the set $0.08, 0.16, 0.24, 0.32, 0.40$.

EC.4.4. Model Selection and Specification

Similar to the Hillstrom case study, we set a constant propensity score (0.5) and select linear regression as the conditional mean model in Robinson's transformation. The mean-squared error (MSE) for this model is reported in Table EC.12.

For initial CATE estimation, we consider the following models:

Table EC.11 Covariate Balance Check for Starbucks Data

Variable	Mean Diff.	Pooled St. Dev.	Standardized Mean Diff.
V1 = 0	−0.002	0.331	−0.005
V1 = 1	−0.003	0.484	−0.006
V1 = 2	0.002	0.484	0.004
V1 = 3	0.003	0.331	0.009
V2	−0.011	5.001	−0.002
V3	0.012	1.000	0.012
V4 = 1	−0.001	0.466	−0.002
V4 = 2	0.001	0.466	0.002
V5 = 1	0.003	0.387	0.007
V5 = 2	−0.001	0.483	−0.001
V5 = 3	−0.001	0.487	−0.001
V5 = 4	−0.002	0.240	−0.006
V6 = 1	0.000	0.433	0.000
V6 = 2	0.000	0.432	0.000
V6 = 3	0.001	0.433	0.001
V6 = 4	−0.001	0.434	−0.002
V7 = 1	−0.001	0.457	−0.002
V7 = 2	0.001	0.457	0.002

Table EC.12 MSE of Conditional Mean Outcome Models: Starbucks Data

Scenario 1: LDP-Protected Covariates

Privacy	Linear Regression	Logistic Regression	Random Forest
No	0.0120	0.0120	0.0120
Very Low	0.0120	0.0120	0.0120
Low	0.0114	0.0114	0.0114
Medium	0.0123	0.0123	0.0123
High	0.0118	0.0118	0.0118
Very High	0.0117	0.0117	0.0117

Scenario 2: LDP-Protected Outcome

Privacy	Linear Regression	Logistic Regression	Random Forest
No	0.0120	0.0120	0.0120
Very Low	0.0120	0.0120	0.0120
Low	0.0146	0.0146	0.0147
Medium	0.0172	0.0172	0.0173
High	0.0214	0.0214	0.0217
Very High	0.0264	0.0264	0.0268

1. *Causal Forest*: We use the causal forest function implemented in the `grf` package with 200 trees and other default parameters. Similarly, we choose 200 trees instead the default 2,000 trees to accelerate the model training process.
2. *T-learner*: We construct a single predictive model to estimate the outcome as a function of the treatment and covariates using random forest with 200 trees and default parameters in

`grf` package). Then, we calculate the predicted CATE as the difference in predicted outcomes between treated and non-treated conditions, while keeping all other covariates constant.

3. *R-learner*: We implement ten-fold cross-fitted R-learner, constructing Robinson's transformation using a constant propensity score (0.5) and employing `Xgboost` as the conditional outcome model. We subsequently utilize `Xgboost` to estimate the CATE, based on Robinson's score. For all the `Xgboost` models, configure the model by setting the maximum depth of each tree to 1, selecting a learning rate of 0.5, and setting the number of iterations to 10. These parameters have been chosen as they yield the highest cross-validated accuracy for outcome predictions.

EC.4.5. Results for T-learner and R-learner

Table EC.9 presents the AUTOC values (multiplied by 100) of different methods, utilizing T-learner and R-learner as the initial CATE models. In line with previous findings, our proposed solution consistently outperforms the other methods across a broad spectrum of privacy levels. The only exception is that at the *Very High* level of outcome privacy, the proposed method performs the same as the default approach, while the global-only and local-only performs worse than the default method.

Table EC.13 AUTOC Values (Multiplied by 100) of Different Methods for Starbucks Data

(a) Results from T-learners

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**0.66 (62%)**	0.65 (62%)	0.64 (50%)	0.64	**0.66 (62%)**	0.65 (62%)	0.64 (50%)	0.64
Very Low	**0.57 (66%)**	0.57 (58%)	0.56 (56%)	0.55	**0.53 (82%)**	0.50 (70%)	0.47 (58%)	0.46
Low	**0.52 (66%)**	0.51 (62%)	0.49 (47%)	0.49	**0.38 (88%)**	0.34 (76%)	0.30 (64%)	0.29
Medium	**0.46 (68%)**	0.45 (62%)	0.43 (56%)	0.43	**0.31 (86%)**	0.26 (64%)	0.22 (56%)	0.21
High	**0.39 (70%)**	0.37 (58%)	0.36 (48%)	0.37	**0.24 (92%)**	0.20 (76%)	0.17 (66%)	0.14
Very High	**0.32 (61%)**	0.31 (58%)	0.29 (48%)	0.30	**0.21 (92%)**	0.18 (76%)	0.16 (74%)	0.11

(b) Results from R-learner

Privacy	Scenario 1: LDP-Protected Covariates				Scenario 2: LDP-Protected Outcome			
	Proposed	Global	Local	Default	Proposed	Global	Local	Default
No	**0.65 (94%)**	0.65 (94%)	0.63 (88%)	0.60	**0.65 (94%)**	0.65 (94%)	0.63 (88%)	0.60
Very Low	**0.57 (84%)**	0.56 (85%)	0.55 (73%)	0.53	**0.58 (62%)**	0.57 (54%)	0.56 (40%)	0.57
Low	**0.50 (77%)**	0.50 (72%)	0.49 (65%)	0.47	**0.52 (56%)**	0.49 (44%)	0.47 (36%)	0.50
Medium	**0.44 (68%)**	0.43 (61%)	0.43 (55%)	0.42	**0.44 (60%)**	0.41 (48%)	0.41 (46%)	0.43
High	**0.37 (71%)**	0.37 (65%)	0.36 (58%)	0.35	**0.37 (56%)**	0.36 (54%)	0.35 (54%)	0.35
Very High	**0.32 (64%)**	0.31 (57%)	0.30 (50%)	0.30	**0.32 (50%)**	0.31 (45%)	0.30 (40%)	0.32

Note: We report the average of AUTOC values (multiplied by 100) from 100 bootstrap replications, along with the percentage of replications in which the AUTOC value of the focal method is greater than the AUTOC value of the default approach (given in parentheses).