Differentially Private Individual Treatment Effect Estimation from Aggregated Data

Artem Betlei, Théophane Gregoir, Thibaud Rahier, Aloïs Bissuel, Eustache Diemert, Massih-Reza Amini

Differentially Private Individual Treatment Effect Estimation from Aggregated Data

Artem Betlei
CAIL & Université Grenoble Alpes

Théophane Gregoir
CAIL & ENS Paris-Saclay

Thibaud Rahier
CAIL

Aloïs Bissuel
CAIL

Eustache Diemert
CAIL

Massih-Reza Amini
Université Grenoble Alpes

September 9, 2021

Abstract

Individual Treatment Effect (ITE) estimation has become one of the main trends in Causal Inference due to its applications in various areas where personalization is key. In order to circumvent the complex problem of causal identification, the randomized control trial (RCT) set-up is used in several domains which refer to ITE estimation as uplift modeling. If practitioners used to have full access to the user-level data in order to learn uplift models, the rise of privacy concerns in different domains such as healthcare or online advertising motivates to explore how such models could be trained to reach significant performances while ensuring relevant privacy guarantees. We present ϵ-ADUM, an ϵ-differentially private method to learn uplift models from data aggregated according to a given partition of the feature space. After adapting the bias-variance decomposition to the Precision in Estimation of Heterogeneous Effects (PEHE) metric, we propose an upper bound of the performance of ϵ-ADUM under a set of illustrative assumptions, which explicits the privacy-utility trade-off for this class of models and provides insights on how the size of the underlying partition can be adapted to match the privacy constraints. Finally, we provide experiments on both synthetic and real data highlighting that ϵ-ADUM outperforms ϵ-differentially private models with access to individual data for strong privacy guarantees ($\epsilon \leq 5$).

1 Introduction

1.1 Motivations

Estimating the causal effect of an action at the individual level − or *Individual Treatment Effect* (ITE) − is a problem of growing interest in the machine learning community, particularly for healthcare [1], online advertising [2, 3] or socio-economic [4] applications.

Many of these applications imply to handle sensitive data for which there are rising privacy concerns. Consequently, many industries are starting to enforce procedures ensuring individual data protection. In the online advertising sector for example, a series of changes to data access were proposed recently by Google Chrome [5] in order to guarantee web users privacy through data aggregation and differential privacy.

In consequence, the scientific community has grown a strong interest in proposing ITE prediction methods which fully leverage the trade-off between privacy and utility.

1.2 Related Work

1.2.1 ITE, CATE & uplift modeling

Individual treatment effect (ITE) [6] may be formally defined using the *potential outcomes framework* [7], in which each individual i has two potential outcomes $Y_i(1)$ (if i receives the treatment) and $Y_i(0)$ (if i

does not receive the treatment). The ITE of individual i is then given by the difference $Y_i(1) - Y_i(0)$. In practice, individuals are often described by a set of features (contained in a variable X), and one rather aims at estimating the *conditional average treatment effect* (CATE) [6], defined, for an individual i with features $X = x_i$, as:

$$\tau(x_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X = x_i]. \tag{1}$$

Standard approaches in ITE/CATE prediction include model-agnostic methods (meta-learners [8, 9, 10], modified outcome methods [11] and their combinations [12]) that implicitly predict the CATE, as well as tree-based approaches which are particularly suited for direct treatment effect estimation[13, 14]. A prolific series of increasingly performing algorithm targeted towards ITE prediction have been proposed, using a variety of techniques to adjust for the covariate shift, such as generative adversarial nets [15], auto-encoders [16], double machine learning [17], representation learning [18, 19] or confounder balancing algorithms [20]. Another recent trend is to study theoretical limits in ITE prediction and especially generalization bounds [21].

A practical-oriented branch of the work on CATE estimation − called *Uplift Modeling* (UM) [22] − focuses on the *Randomized Control Trial* (RCT) setting, in which individuals are **randomly** split into a *treatment* group and a *control* group. This set-up circumvents the causal identification problem (avoiding selection bias) and ensures the CATE (or uplift) is rightfully given by the difference between the following **conditional** expectations:

$$u(x_i) = \mathbb{E}[Y|X = x_i, T = 1] - \mathbb{E}[Y|X = x_i, T = 0]. \tag{2}$$

Privately learning the function defined in (2) is the main focus of our work.

UM methods are often overlapping with ITE prediction techniques or reinventing independently, as the former is subproblem of the latter. For instance, the two-model method is described both in ITE [8] and UM [23] literature, the class variable transformation [24] is a particular case of [11], and the tree-based method from the UM community [25, 26] only fractionally differ from ITE ones. However, there are methods specific to the UM which either extend the two-model method by using certain representations [27] or aim at maximizing the AUUC (see below) directly [28, 29, 30].

1.2.2 Metrics

In order to evaluate CATE estimators, one may use an adapted version of the *Mean-Squared Error* (MSE), namely the *Precision in Estimation of Heterogeneous Effects* (PEHE) [31], which is defined for a model $\hat{\tau}$ as:

$$\epsilon_{PEHE}(\hat{\tau}) = \mathbb{E}\left[\Big(\tau(X) - \hat{\tau}(X)\Big)^2\right], \tag{3}$$

where the expectancy is taken with respect to the distributions of both X and the data from with the model $\hat{\tau}$ is learned.

In practice however, only one of the two potential outcomes is observable for a given individual (the **factual** outcome). This is known as the *Fundamental Problem of Causal Inference* (FPCI), and prevents the computation of the PEHE in real settings. A possible solution is to use a ranking-based metric such as the uplift curve [25], which evaluates a sorting of the individuals according to their predicted uplift score. In that vein, the *Area Under the Uplift Curve* (AUUC) [32] represents the most popular metric in the community.

1.2.3 Learning from aggregated data

Learning individual-level behavior from aggregated-level data has long been known as the ecological inference problem. Plenty of presented aggregated-level methods [33] attempt to tackle the problem of ecological fallacy: when the inferences drawn from aggregate level drastically differs from the ground truth at the individual level.

The most relevant level of aggregation has not yet been completely determined by the research community as the term "aggregated data" has been referring to different frameworks: label similarities with complete access to features [19], aggregated labels with complete access to features [34] or aggregated labels with aggregated features [35]. In this work, both features and labels are considered as sensitive and therefore need to be aggregated, which corresponds to the most restrictive setting.

However, regardless of the selected level of aggregation, most of these methods do not ensure theoretical privacy guarantees without being combined with differential privacy.

1.2.4 Differential Privacy

Differential privacy [36] represents one of the most widely used data protection method in so far as it enables researchers to precisely quantify privacy guarantees while being applicable to general setups. Differential privacy should be considered as a process-oriented method, which allows the private training of models.

In order to learn in a differentially private framework, the most common techniques include result perturbation, objective perturbation [37] or noisy iterative optimization methods which can be performed thanks to a precise budget tracking. In particular, differentially private stochastic methods adding scaled noises for each training batch have already shown great performances when applied to deep learning models [38, 39].

The model we propose enables a one-shot spending of the privacy budget, avoiding both its complex tracking and adaptive spending.

1.3 Our contributions

1. We introduce ϵ-Aggregated Data Uplift Model (ϵ-ADUM), a differentially private method to learn uplift models from data aggregated along a given partition of the feature space.

2. We identify and illustrate a bias-variance decomposition for ϵ-ADUM, highlighting the role of the underlying partition size in the privacy-utility trade-off.

3. Finally we show empirically on both synthetic and real data that, for strong privacy guarantees ($\epsilon \leq 5$), ϵ-ADUM outperforms comparable ϵ-differentially private models with access to individual data.

2 ADUM and its bias-variance trade-off

2.1 Preliminaries

2.1.1 Variables and data

We consider random variables X (features), T (treatment) and Y (outcome) with respective values in $\mathcal{K}$ (a compact convex subset of $\mathbb{R}^d$), $\{0, 1\}$ and $\mathbb{R}$. We additionally suppose there exists treatment/control response functions $f^T, f^C : \mathcal{K} \to \mathbb{R}$ and a real random variable ξ (independent of X) with $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] = \sigma^2$, such that

$$Y = Tf^T(X) \ + \ (1-T)f^C(X) \ + \xi. \tag{4}$$

Under these notations, and for any x, we have that $f^C(x) = \mathbb{E}[Y|T = 0, X = x]$, $f^T(x) = \mathbb{E}[Y|T = 1, X = x]$ and the corresponding uplift is defined as:

$$u(x) = f^T(x) - f^C(x). \tag{5}$$

Finally, we assume we have access to $\mathcal{D} = \{(x_i, t_i, y_i)\}_{1\leq i\leq n}$, a dataset containing n *i.i.d.* realizations of (X, T, Y). Since we are in a randomized controlled trial (RCT) setting, the binary treatment variable T is assumed independent of X. We denote $\mathcal{T}$ and $\mathcal{C}$ the subsets of $\mathcal{D}$ which contain respectively all datapoints from the treatment ($T = 1$) and control ($T = 0$) groups.

2.1.2 Space partitioning

For a fixed positive integer p we define $\Pi_p(\mathcal{K}) := \left\{\pi \in \{1, \ldots, p\}^{\mathcal{K}} \ : \ \pi \text{ surjective}\right\}$, the set of all possible partitions of $\mathcal{K}$ containing p elements. Let $\pi \in \Pi_p(\mathcal{K})$ be a fixed partition, then there exists $G_\pi^{(1)}, \ldots, G_\pi^{(p)}$ disjoint subsets of $\mathcal{K}$ such that $\bigcup\limits_{1 \leq j \leq p} G_\pi^{(j)} = \mathcal{K}$. For a given $x \in \mathcal{K}$, we denote $G_\pi(x) = \pi^{-1}(\{\pi(x)\})$, the component of π which contains x. For any $G \subset \mathcal{K}$ we denote $|G|^{\mathcal{D}} = \sum\limits_{i \in \mathcal{D}} \mathbb{I}_{x_i \in G}$, *i.e.* the number of points of $\mathcal{D}$ for which the feature vector x_i belongs to G.

2.2 ADUM presentation

We now present *Aggregated Data Uplift Models* (ADUM). For a given partition $\pi \in \Pi_p(\mathcal{K})$, we estimate the uplift of $x \in \mathcal{K}$ by the average treatment effect in the group $G_\pi(x)$. More formally, we define $\hat{u} : \mathcal{K} \to \mathbb{R}$ the function which to all $x \in \mathcal{K}$ assigns:

$$\hat{u}_\pi(x) = \hat{f}_\pi^T(x) - \hat{f}_\pi^C(x), \tag{6}$$

where $\hat{f}_\pi^T$ and $\hat{f}_\pi^C$ refer respectively to aggregated-data based models of the treatment and control response functions, *i.e.*:

$$\hat{f}_\pi^T(x) = \frac{1}{|G_\pi(x)|^T} \sum_{i : x_i \in G_\pi(x)} y_i t_i,$$
$$\hat{f}_\pi^C(x) = \frac{1}{|G_\pi(x)|^C} \sum_{i : x_i \in G_\pi(x)} y_i (1 - t_i).$$

$\hat{f}_\pi^T$ and $\hat{f}_\pi^C$ are piecewise constant functions defined using only aggregated information and would therefore be computable from an aggregate reporting API [40] thanks to `SUM` and `COUNT` queries.

By using aggregated data models for both the treatment and control positive outcome functions, we partially circumvent the FPCI: as long as there are points from both the treatment and control groups in any given component $G_\pi(x)$ of π, the average treatment effect in $G_\pi(x)$ is consistently estimated by $\hat{u}(x)$.

Remark Besides, outside of the RCT setting, additionally assuming $\{\pi(X)\}$ is a valid adjustment set [41] for (T, Y) — e.g. in the case where $X \perp\!\!\!\perp T \,|\, \pi(X)$ which is a strictly weaker assumption than the RCT setting — guarantees ADUM rightfully models the causal effect of T on Y. Nevertheless, finding such a partition represents a non-trivial task which is not the subject of this article.

2.2.1 General PEHE bound for ADUM

Let $\hat{f}_\pi : \mathcal{K} \to \mathbb{R}$ be a model for a given $f : \mathcal{K} \to \mathbb{R}$. For all $x \in \mathcal{K}$ we define:

$$\mathrm{Bias}(\hat{f}_\pi(x)) = f(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_\pi(x)],$$
$$\mathrm{Var}(\hat{f}_\pi(x)) = \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_\pi(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_\pi(x)]\right)^2\right].$$

The (squared) bias term captures how well can f be approached by a piecewise constant function on π: it should typically **decrease when** $|\pi| = p$ **increases**. The variance term captures how close $\hat{f}_\pi$ is to its average in each of the components of π: it should typically **increase when** $|\pi| = p$ **increases**.

Proposition 1 *Let $\pi \in \Pi_p(\mathcal{K})$ and $\hat{u}_\pi$ the associated ADUM learned wrt data $\mathcal{D} = \mathcal{T} \sqcup \mathcal{C}$, then the PEHE of $\hat{u}_\pi$ satisfies:*

$$\epsilon_{PEHE}(\hat{u}_\pi) \leq 2\mathbb{E}\left[\mathrm{Bias}^2\left(\hat{f}_\pi^{\mathcal{C}}\right) + \mathrm{Bias}^2\left(\hat{f}_\pi^{\mathcal{T}}\right)\right]$$
$$+ 2\mathbb{E}\left[\mathrm{Var}\left(\hat{f}_\pi^{\mathcal{C}}\right) + \mathrm{Var}\left(\hat{f}_\pi^{\mathcal{T}}\right)\right]$$

2.3 ϵ-ADUM : definition and algorithm

In order to get theoretical privacy guarantees, ADUM must be combined with differential privacy. Since ADUM is based on the computation of means, it can be decomposed into a set of `SUM` and `COUNT` queries. Knowing the range of the outcome D_y, the sensitivities of these queries are directly available.

As the partition π creates disjoint subsets of the input domain, the privacy budget ϵ can be entirely spent on each group queries in parallel [42]. Here, we choose to assign an $\frac{\epsilon}{2}$ budget to each `SUM` or `COUNT` query. Therefore, all the queries can be noised thanks to a scaled Laplace noise, turning ADUM into an ϵ-differentially private model: ϵ-ADUM (see Algorithm 1).

Algorithm 1 ϵ-ADUM

```
function TRAIN($(x_i, t_i, y_i)_{i\in[1,n]}, \pi \in \Pi_p(\mathcal{K}), D_y > 0, \epsilon > 0$):
    for $k \in [1, p]$ do
        for $t \in \{0, 1\}$ do
            $E_{k,t} = (y_i \mid \pi(x_i) = k,\ t_i = t)$
            $C_{k,t} = \mathrm{COUNT}(E_{k,t}) + \mathrm{Lap}(\frac{2}{\epsilon})$          ▷ $\frac{\epsilon}{2}$-DP count
            $S_{k,t} = \mathrm{SUM}(E_{k,t}) + \mathrm{Lap}(2\frac{D_y}{\epsilon})$          ▷ $\frac{\epsilon}{2}$-DP sum
            $\widehat{y}_{k,t} = \frac{S_{k,t}}{C_{k,t}}$          ▷ $\epsilon$-DP mean
        end for
        $\widehat{u}_k = \widehat{y}_{k,1} - \widehat{y}_{k,0}$          ▷ $\epsilon$-DP piecewise constant model
    end for
    return $(\widehat{u}_k)_{k\in[1,p]}$
end function

function PREDICT($x_{new} \in \mathcal{K}$):
    return $\widehat{u}_{\pi(x_{new})}$          ▷ Assign value linked to $G_\pi(x_{new})$
end function
```

2.4 The bias-variance trade-off for $\epsilon-$ADUM: insights from an illustrative setting

2.4.1 Simplified setting

For the sake of the result we present in the next subsection, we consider the following illustrative setting: let π be a partition of $\mathcal{K}$, with $|\pi| = p$ components and assume that f^T and f^C are respectively L_T and L_C Lispschitz on $\mathcal{K}$ that we suppose uni-dimensional ($d = 1$) of diameter D_x. Moreover, we denote $\beta_\pi = \max_{G,G'\in\pi}\{\mathrm{diam}(G)/\mathrm{diam}(G')\}$, and make the assumptions that every group $G \in \pi$ is equally populated with respect to $\mathcal{T}$ and $\mathcal{C}$, *i.e.* $\forall G \in \pi$, $|G|^{\mathcal{T}} = |G|^{\mathcal{C}}$

2.4.2 PEHE bounding for $\epsilon-$ADUM

Corollary 1 *For a given $\Delta \in (0, 1)$, let $p, n \in \mathbb{N}$, $\mathcal{D}$ a dataset of size n, $\pi \in \Pi_p(\mathcal{K})$ and $\epsilon \geq \frac{8p\log(1/\Delta)}{n}$. Let $\hat{u}_\pi$ be the corresponding $\epsilon-$ADUM (defined in Algorithm 1), then the following inequality holds with probability $\geq 1 - \Delta$:*

$$\begin{aligned} \epsilon_{PEHE}(\hat{u}_\pi) &\leq 2(L_C^2 + L_T^2)D_x^2\beta_\pi^2\mathbf{p}^{-2} && \textit{ADUM Bias} \\ &+ 4\left(2\sigma^2 + (L_C^2 + L_T^2)D_x^2\right)\frac{\mathbf{p}}{\mathbf{n}} && \textit{ADUM Variance} \\ &+ (24D_y)^2\frac{\mathbf{p^2}}{\mathbf{n^2}\epsilon^\mathbf{2}} && \epsilon\textit{−DP term.} \end{aligned} \tag{7}$$

When making $\epsilon-$differentially private queries, it is typical to constrain ϵ to be significantly bigger than the inverse of the population of the group upon which the query is made [43], which is consistent with the condition on ϵ stated in the Corollary 1. For instance, if $n = 2 \cdot 10^4$, $p \leq 20$ and $\Delta = 0.01$, the bound holds with probability 99% for any $\epsilon \geq 0.04$.

The number of groups p^{opt} that minimizes the upper bound in (7) has the following asymptotic variations with respect to ϵ and n:

- when ϵ is small compared to $\sqrt{p/n}$, (7) is dominated by its first and last terms and $p^{opt} = \Theta(n\epsilon)$,
- when ϵ is large compared to $\sqrt{p/n}$, (7) is dominated by its two first terms and $p^{opt} = \Theta(n^{1/3})$ does not depend on ϵ.

This shows the flexibility of the class of ADUM models, which robustly adapt to noise addition when the size of the underlying partition is rightfully tuned.

3 Experiments and results

3.1 Synthetic data

3.1.1 Data generation

First, ϵ-ADUM is tested in a synthetic framework in order to observe its performance in terms of PEHE. Each of the n generated individuals are attributed a covariate $X \sim \mathcal{U}(-1, 1)$ $(d = 1)$ and a treatment $T \sim Bernoulli(0.5)$. The treatment effect surface is defined by the difference between response surfaces of treatment and control populations. Each individual couple of potential outcomes is generated following $f^C(X) = 0$, $f^T(X) = \sin X$ and $\xi \sim \mathcal{N}(0, \sigma)$ in order to observe a simple but non-monotonic and noisy treatment effect surface. Moreover, ϵ-ADUM is computed on a regular cut of $\mathcal{K}$ in order to have balanced groups (as $X \sim \mathcal{U}(-1, 1)$) and be consistent with Corollary 1.

3.1.2 Performance comparison

Here, ϵ-ADUM is compared with a *Two-Model* (TM) [23] uplift modeling method, formed by two ϵ-differentially private linear regressions [44] with polynomial features which have access to individual data, denoted ϵ-TM. For each ϵ, we respectively tune the polynomial degree and the number of groups for ϵ-TM and ϵ-ADUM. As highlighted by Figure 1, ϵ-ADUM reaches better performances than individually-trained models for $\epsilon \leq 5$, while $\epsilon = 5$ is often presented as a realistic parameter for the future of the tech industry (including advertising [5]). Indeed, the ADUM framework offers a more robust and easily implementable adaptation to noise addition than individual frameworks thanks to its query architecture. Nevertheless, when considering large ϵ (corresponding to low privacy guarantees), we observe that the great interaction between aggregation and noise addition is being overruled by individual models which benefit from their complete access to granular information. The significant drop in PEHE for ϵ-TM can be explained by the privacy cost of using a supplementary polynomial degree becoming profitable for a privacy budget $\epsilon \geq 2$.

3.1.3 Bias-variance trade-off illustration

The bias-variance trade-off introduced in Corollary 1 is illustrated experimentally in Figure 2. Indeed, for every value of ϵ, as the number of groups increases, the $PEHE$ starts by decreasing because of the bias reduction (first term of (7)) before increasing due to a penalizing variance (second term of (7)) and the rising impact of the privacy-induced noise addition (third term in (7)) $-$ the two latter being due to an insufficient population in the groups. Furthermore, this experiment also highlights the dependency between ϵ and the optimal number of groups for ϵ-ADUM. First, when ϵ increases, the optimal number of groups increases and ϵ-ADUM's best performance improves. Then, as illustrated by the two merged performance curves for $\epsilon = 50$ and $\epsilon = 100$, ϵ-ADUM enters a capped regime for which the ϵ-differentially private perturbation becomes negligible compared to errors inherent to ADUM (see 2 asymptotic regimes in Section 2.4).

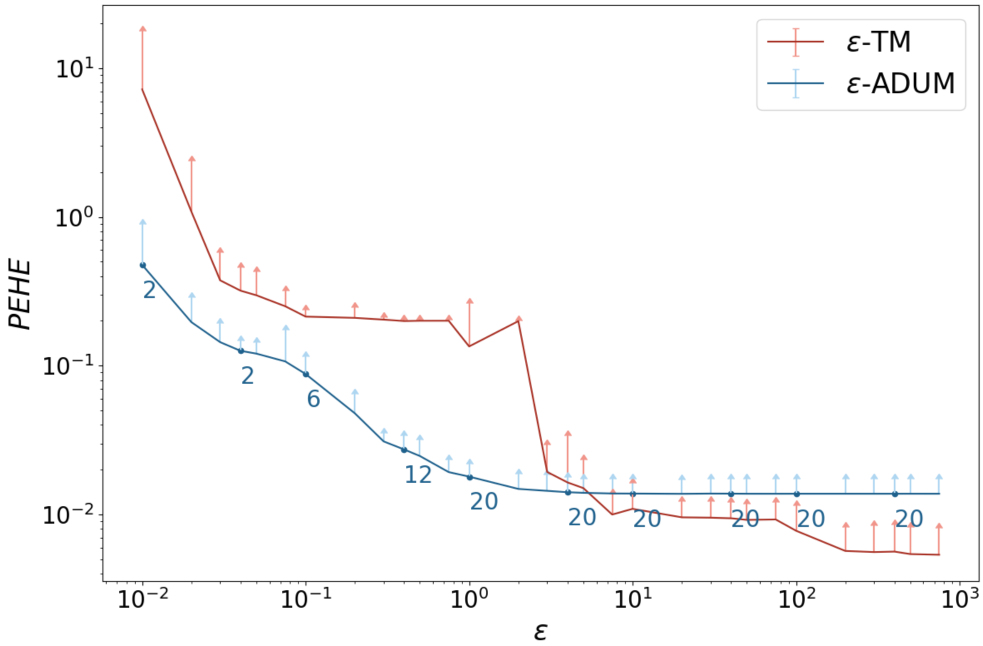

Figure 1: Comparison of test $PEHE$ (lower is better) for ϵ-TM and ϵ-ADUM over 20 random train/test splits selecting 20000 points. Arrows represent standard deviations and the tuned number of groups for ϵ-ADUM is annotated in blue. For this experiment, $\sigma = 1$.

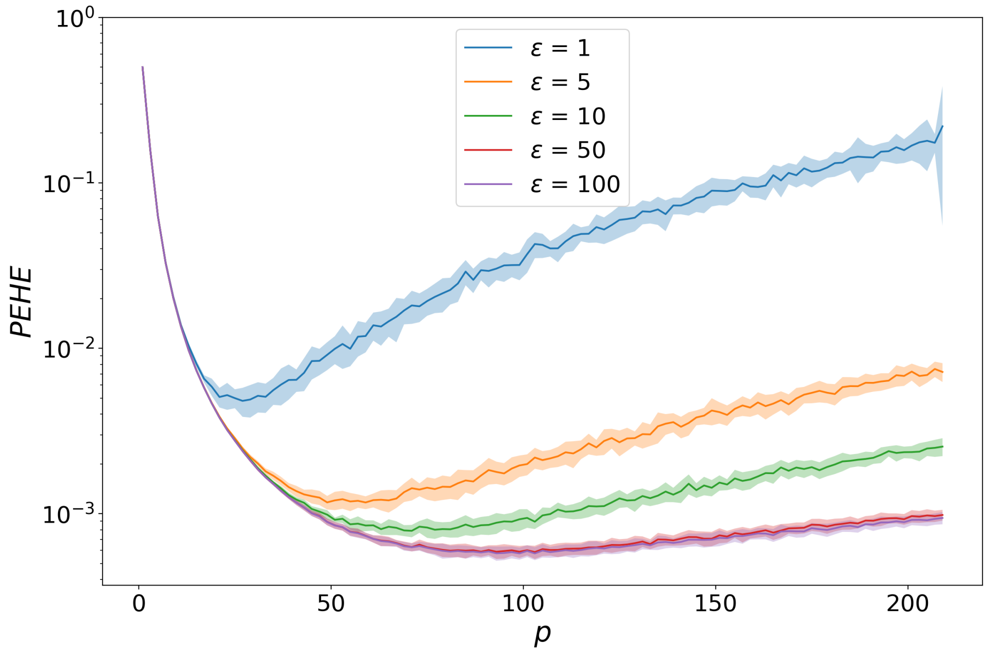

Figure 2: Test $PEHE$ (lower is better) over 20 random train/test splits selecting 20000 points, illustrating the ϵ-ADUM bias-variance trade-off with respect to the number of groups p for 5 selected ϵ. For this experiment, $\sigma = 0.1$.

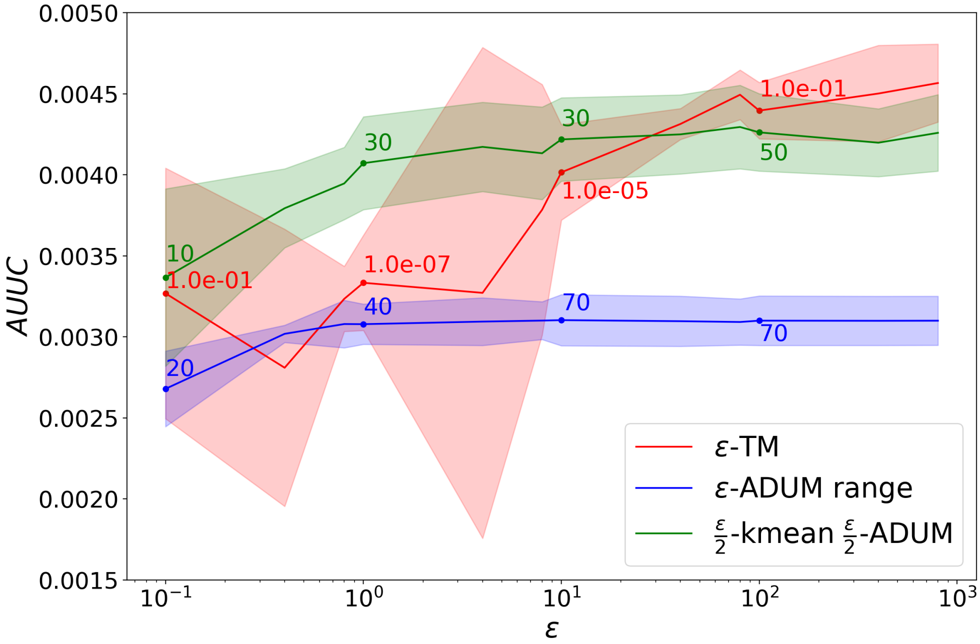

Figure 3: Comparison of test *AUUC* (higher is better) between individually-trained ϵ-TM and two variations of ϵ-ADUM over 4 random train/test splits randomly selecting 1M points from CRITEO-UPLIFTv2. The tuned number of for ϵ-ADUM is annotated in blue and green while the tuned regularization parameter C is in red for ϵ-TM.

3.2 Real data

CRITEO-UPLIFTv2 dataset [2] is an open large scale dataset constructed from incrementality A/B tests. Results are reported for the "visit" binary outcome, hence ϵ-differentially private logistic regressions [37] are used as prediction models in an ϵ-TM method.

As presented in Section 2.2, ϵ-ADUM is partition-dependant. For a real dataset, trivial partitions of $\mathcal{K}$ such as one-dimensional regular cut are not sufficient anymore, and we propose to find a relevant partition while preserving privacy guarantees by decomposing our privacy budget ϵ in an $\frac{\epsilon}{2}$-kmeans partitioning [45] $-$ outputting a partition π $-$ and a consecutive $\frac{\epsilon}{2}$-ADUM along π. It is worth mentioning that in practice, the partition and its corresponding mean queries could be provided by an external actor in order to avoid any access to granular data.

As observed on synthetic data, ϵ-ADUM appears to outperform models with access to individual data for strict privacy guarantees ($\epsilon \leq 5$). Once again, when privacy guarantees loosen up, the ϵ-differentially private TM overtakes ϵ-ADUM thanks to its access to granular data (see Figure 3). Moreover, the significant impact of the partition is illustrated by the difference of performances between one-dimensional regular cut (on the first feature) and $\frac{\epsilon}{2}$-kmeans partitioning even though the consecutive $\frac{\epsilon}{2}$-ADUM is performed with a halved privacy budget.

4 Conclusion

In this article, we introduce ϵ-ADUM, a new uplift ϵ-differentially private method to learn uplift models from aggregated data. Then, a theoretical study of this model is conducted giving insights on its empirical error through the expression of a bias-variance trade-off. Finally, on both synthetic and real data, ϵ-ADUM is tested and appears to outperform classical differentially private methods for strong privacy guarantees ($\epsilon \leq 5$) although the latter can access a granular level of data.

References

[1] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24), 2011.

[2] Eustache Diemert, Artem Betlei, Renaudin Christophe, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London, United Kingdom*, 2018.

[3] Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Theophane Gregoir, and Thibaud Rahier. A large scale benchmark for individual treatment effect prediction and uplift modeling. 2021.

[4] Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1), 2012.

[5] Privacy sandbox. `https://www.chromium.org/Home/chromium-privacy/privacy-sandbox`.

[6] Weijia Zhang, Jiuyong Li, and Lin Liu. A unified survey on treatment effect heterogeneity modeling and uplift modeling. *arXiv preprint arXiv:2007.12769*, 2020.

[7] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[8] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[9] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[10] Thibaud Rahier, Amélie Héliou, Matthieu Martin, Christophe Renaudin, and Eustache Diemert. Individual treatment prescription effect estimation in a low compliance setting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1399–1409, 2021.

[11] Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

[12] Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[13] Susan Athey and Guido Imbens. Recursive Partitioning for Heterogeneous Causal Effects. 4 2015.

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

[15] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[16] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

[17] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018.

[18] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[19] Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 265–274, 2017.

[21] Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138. PMLR, 2018.

[22] Nicholas J Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal*, 1(3):14–21, 2007.

[23] Behram Hansotia and Brad Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35, 2002.

[24] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. *ICML Workshop on Clinical Data Analysis*, 2012.

[25] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling. In *2010 IEEE International Conference on Data Mining*, pages 441–450. IEEE, 2010.

[26] Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, pages 1–33, 2011.

[27] Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift prediction with dependent feature representation in imbalanced treatment and control conditions. In *International Conference on Neural Information Processing*, pages 47–57. Springer, 2018.

[28] Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support Vector Machines for Differential Prediction. *Machine learning and knowledge discovery in databases : European Conference, ECML PKDD ... : proceedings. ECML PKDD (Conference)*, 8725:50–65, 2014.

[29] Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[30] Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift modeling with generalization guarantees. In *Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, 2021.

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

[32] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, volume 46, 2012.

[33] Gary King, Martin A Tanner, and Ori Rosen. *Ecological inference: New methodological strategies.* Cambridge University Press, 2004.

[34] Avradeep Bhowmik, Minmin Chen, Zhengming Xing, and Suju Rajan. Estimagg: A learning framework for groupwise aggregated data. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 477–485. SIAM, 2019.

[35] Avradeep Bhowmik, Joydeep Ghosh, and Oluwasanmi Koyejo. Sparse parameter recovery from aggregated data. In *International Conference on Machine Learning*, pages 1090–1099. PMLR, 2016.

[36] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[37] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12(null):1069–1109, July 2011.

[38] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[39] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data, 2017.

[40] Michael Kleber. Turtledove. `https://github.com/WICG/turtledove/`, 2019.

[41] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, NY, USA, 2000.

[42] Frank McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. *Commun. ACM*, 53(9):89–97, September 2010.

[43] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. *2014 IEEE 27th Computer Security Foundations Symposium*, Jul 2014.

[44] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy, 2012.

[45] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means clustering, 2015.