Uplift Modeling for Online Advertising

Artem Betlei

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques et Informatique

Arrêté ministériel : 25 mai 2016

Présentée par

Artem BETLEI

Thèse dirigée par **Massih-Reza AMINI**,Professeur, Université Grenoble Alpes
et co-encadrée par **Eustache DIEMERT**, Criteo

préparée au sein du **Laboratoire Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Modélisation d'Uplift pour la Publicité en ligne

Uplift Modeling for Online Advertising

Thèse soutenue publiquement le **24 novembre 2021**,
devant le jury composé de :

Monsieur MASSIH-REZA AMINI
PROFESSEUR DE UNIVERSITES, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur GILLES GASSO
PROFESSEUR DES UNIVERSITES, INST NAT SC APPLIQ ROUEN, Rapporteur
Monsieur SZYMON JAROSZEWICZ
PROFESSEUR, Polska Akademia Nauk, Rapporteur
Madame SIHEM AMER-YAHIA
DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES, Présidente
Monsieur ALAIN RAKOTOMAMONJY
PROFESSEUR DES UNIVERSITES, INST NAT SC APPLIQ ROUEN, Examinateur

Acknowledgements

First of all, I would like to express my deepest appreciation to my supervisors – Massih-Reza Amini and Eustache Diemert. I am grateful for your guidance, patience, support, valuable advice, time-involved, and for experience and variety of research ideas shared with me during these three years. Then, I would like to extend my gratitude to Gilles Gasso and Szymon Jaroszewicz, who kindly agreed to review my PhD thesis. I must also thank Sihem Amer-Yahia and Alain Rakotomamonjy, who agreed to be part of my thesis committee. Special thanks to my teammates of Criteo AI Lab, Causal Learning team (past and present) for their unrelenting encouragement and helpful contributions: Thibaud, Matthieu, Houssam, Panayotis, Amelie, Theophane, Mohamed, Kiewan. I am also grateful to all the APTIKAL members (past and present) for their support and useful advice: Sasha, Nastya, Lies, Vasya, Sami, Karim, Saeed. Thanks to LIG, Universite Grenoble Alpes and Criteo for their logistical support. Thanks to Criteo engineers for providing data and advice in this PhD. I very much appreciate my friends for their interest and stimulating discussions. Finally, my success would not have been possible without the encouragement and nurturing of my girlfriend, brother and parents.

Abstract

Uplift modeling is a machine learning-based technique for treatment effect prediction at the individual level, which has become one of the main trends in application areas where personalization is key, such as personalized medicine, performance marketing, social sciences, etc.

This thesis is intended to expand the scope of uplift modeling for experimental data by developing new theory and solutions for several open challenges in the field, inspired by the online advertising applications perspective.

Firstly we release a publicly available collection of 13.9 million samples collected from several randomized control trials, scaling up available datasets by a 210x factor. We formalize how uplift modeling can be performed with this data, along with relevant evaluation metrics. Then, we propose synthetic response surfaces and treatment assignment providing a general set-up for Conditional Average Treatment Effect (CATE) prediction and report experiments to validate key traits of the dataset.

Secondly, we assume imbalanced treatment conditions and propose two new data representation-based methods inspired by cascade and multi-task learning paradigms. We provide then series of experimental results over several large-scale real-world collections to check the benefits of the proposed approaches.

We then cover the problem of direct optimization of the Area Under the Uplift Curve (AUUC), a popular metric in the field. Using the relations between uplift modeling and bipartite ranking we provide a generalization bound for the AUUC and derive an algorithm optimizing this bound, usable with linear and deep models. We empirically study the tightness of the proposed bound, its efficacy for hyperparameters tuning, and investigate the performance of the method compared to a range of baselines on two real-world uplift modeling benchmarks.

Finally, we consider the problem of learning uplift models from aggregated data. We propose a principled way to learn group-based uplift models from data aggregated according to a given set of groups that define a partition of the user space, using different unsupervised aggregation techniques, such as feature binning by value or by

quantile. We proceed by introducing a bias-variance decomposition of the Precision when Estimating Heterogeneous Effect (PEHE) metric for models learned on a given grouping and show how this decomposition enables us to derive a theoretical optimal number of groups as a function of data size. Experimental results highlight the bias-variance trade-off and confirm theoretical insights concerning the optimal number of groups. In addition, we show that group-based uplift models can have comparable performance to baselines with full access to the data.

Keywords: uplift modeling, large-scale benchmark, multi-task learning, bipartite ranking, generalization bounds, learning from aggregated data, differential privacy, privacy-utility trade-off.

Résumé

La modélisation d'uplift est une technique basée sur l'apprentissage automatique pour la prédiction de l'effet d'un traitement au niveau individuel, qui est devenue l'une des principales tendances dans les domaines d'application où la personnalisation est essentielle, comme la médecine personnalisée, le marketing de performance, les sciences sociales, etc.

Cette thèse a pour but d'étendre la portée de la modélisation d'uplift pour les données expérimentales en développant une nouvelle théorie et des solutions pour plusieurs défis ouverts dans le domaine, inspirés par la perspective des applications de la publicité en ligne.

Tout d'abord, nous mettons à la disposition du public une collection de 13,9 millions d'échantillons collectés à partir de plusieurs essais de contrôle aléatoires, ce qui multiplie par 210 les ensembles de données disponibles. Nous formalisons la façon dont la modélisation d'uplift peut être effectuée avec ces données, ainsi que les mesures d'évaluation pertinentes. Ensuite, nous proposons des surfaces de réponse synthétiques et une affectation de traitement fournissant une configuration générale pour la prédiction de l'effet de traitement moyen conditionnel (CATE) et nous rapportons des expériences pour valider les caractéristiques clés de l'ensemble de données.

Ensuite, nous supposons des conditions de traitement déséquilibrées et proposons deux nouvelles méthodes basées sur la représentation des données, inspirées des paradigmes d'apprentissage en cascade et multi-tâches. Nous fournissons ensuite une série de résultats expérimentaux sur plusieurs collections à grande échelle pour vérifier les avantages des approches proposées.

Nous abordons ensuite le problème de l'optimisation directe de l'Area Under the Uplift Curve (AUUC), une métrique populaire dans le domaine. En utilisant les relations entre la modélisation d'uplift et le classement bipartite, nous fournissons une limite de généralisation pour l'AUUC et dérivons l'algorithme d'optimisation de cette limite, utilisable avec des modèles linéaires et profonds. Nous étudions empiriquement l'étanchéité de la limite proposée, son efficacité pour le réglage des hyperparamètres

et nous examinons les performances de la méthode par rapport à une série de lignes de base sur deux repères de modélisation d'uplift du monde réel.

Enfin, nous considérons le problème de l'apprentissage de modèles d'uplift à partir de données agrégées. Nous proposons une méthode pour apprendre des modèles d'uplift basés sur des groupes à partir de données agrégées selon un ensemble donné de groupes qui définissent une partition de l'espace utilisateur, en utilisant différentes techniques d'agrégation non supervisées, telles que le regroupement de caractéristiques par valeur ou par quantile. Nous introduisons une décomposition biais-variance de la métrique Precision when Estimating Heterogeneous Effect (PEHE) pour les modèles appris sur un groupe donné et montrons comment cette décomposition nous permet de dériver un nombre optimal théorique de groupes en fonction de la taille des données. Les résultats expérimentaux mettent en évidence le compromis biais-variance et confirment les idées théoriques concernant le nombre optimal de groupes. En outre, nous montrons que les modèles d'uplift basés sur les groupes peuvent avoir une performance comparable à celle des modèles de base avec un accès complet aux données.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

1.1.1 Motivation

Selecting subjects that should be exposed to a given treatment is a problem of growing interest in a variety of application domains related to personalization – such as medicine, social sciences, credit scoring, insurance, performance marketing, or online advertising.

Response modeling has long been considered as a standard machine learning-based solution for this task. In response modeling, the model is trained to predict the outcome after the treatment – or response – and then a treatment assignment rule is derived based on model predictions. Despite its popularity in the industry, this method has a significant disadvantage of not taking into account potential outcomes obtained in the absence of treatment. Consider for instance the situation where an experimenter needs to identify which of the customers to send a promotional newsletter to maximize the amount of sales. The four following scenarios are possible (the true scenario is unknown for an experimenter):

1. Client buys a product after receiving a newsletter and does not buy otherwise – treatment is effective.
2. Client buys a product regardless of receiving a newsletter – treatment is pointless.
3. Client does not buy a product regardless of receiving a newsletter – treatment is pointless.
4. Client does not buy a product after receiving a newsletter and buy otherwise – treatment is harmful.

Originating mainly in the online marketing domain, *uplift modeling* takes a step forward by estimating the causal impact of treatment at the individual level, i.e. the difference in the outcome of the subject when she is being treated or not. For achieving this goal, the whole population is divided randomly into two groups: treatment is assigned to units from the treatment group while control group units do not receive treatment. Then, uplift models are learned based on individual features, treatment flag and final outcome (see Section 2.2).

Illustrative example. Figure 1.1 illustrates a typical uplift modeling pipeline, where data are available from prior, randomized experiments. It could be a pilot study using a randomized control trial (RCT) with a placebo for medicine or an A/B test for marketing (step 1). Then, different models predicting the individual uplift can be learned and evaluated (step 2). A popular metric to value the quality of a model is the Area Under the Uplift Curve (AUUC) [90]. This metric measures the cumulative uplift along individuals sorted by model predictions. A good model (with a high AUUC) scores higher those individuals for which the prediction is high (beneficial) compared to ones for which the prediction is low (neutral or even detrimental). Finally, practitioners use predictions to *rank* future instances and assign treatment to individuals with the highest scores (step 3) [37, 43].

When a new cohort of individuals is available, the predictions of the model will be used to target treatment: highest scored individuals would be treated (green individuals in Figure 1.1) whilst lowest scored ones would be excluded from treatment (blue individuals). This strategy is useful as soon as the treatment effect is *heterogeneous* (i.e. depending on observable covariates). Note that the prediction value itself is not as interesting here as the ranking induced by the predictions.

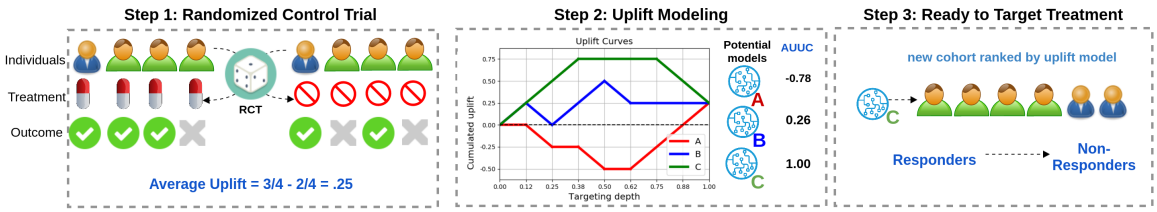


Figure 1.1: Typical uplift modeling pipeline schematized in three steps. Step 1 starts with a randomized control trial using an A/B test. Then Step 2 consists of learning and evaluating several uplift models and selecting the best performing one by Area Under the Uplift Curve (or another metric) on data gathered at Step 1. Finally, at Step 3, the best uplift model is used to target treatment on the next cohort of individuals.

It is worth mentioning here that the problem of uplift modeling is not alone in its role – thus, the task of *Conditional Average Treatment Effect prediction* is closely related to uplift modeling and developed in parallel by researchers from spheres of medicine, econometrics, and social sciences (see Section 1.2 for details).

1.1.2 Challenges in online advertising

Notably for Criteo as an online advertising company, uplift modeling plays an important role in several applied tasks such as incrementality measurement and optimization. Specifically, uplift models can be incorporated as a building block in the real-time bidding process.

Meanwhile, it is worth noting a few peculiarities that arise during the application of uplift models in online advertising and in Criteo in particular:

- **Large-scale data:** unlike most existing fields that exploit uplift modeling and CATE prediction, online advertising companies often deal with large amounts of data – as a consequence, desirable uplift model should satisfy several conditions: it needs to be lightweight enough to be used on servers and scalable enough to be effectively parallelized. Finally, potential models should be tested on big datasets – as some uplift models performing well on datasets of small size might not have the same efficiency on large-scale data.

- **Imbalanced treatment assignment:** depending on the notion of treatment in the task, situations arise when treated or untreated individuals may cost very differently. As a result, an obtained dataset might be made up of treatment and control groups of different sizes, so some models may lose quality for this reason.

- **Generalization guarantees:** in online advertising, prediction models are automatically retrained on the new portions of data every fixed period – that is why it is so important to maintain their generalizability.

- **Privacy guarantees:** privacy-preserving algorithms play a key role in maintaining trustworthy AI. Recently, series of changes to data access were proposed by Google Chrome [1] to guarantee web users privacy through data aggregation and differential privacy. For instance, instead of having access to single records describing a user, marketers could only be able to get aggregated data queried through an *aggregate reporting API* [61, 1]. More concretely, the marketers would be able to upload to the browser a prediction model from the user data, but the data would stay on the browser and the marketer would only access it

through aggregated projections. Additive noise could simultaneously be added to guarantee differential privacy [42] of the process.

The focus of this thesis is to develop theory and methods of uplift modeling that can help to address mentioned challenges.

1.2 Problem formulation

In this section, we define necessary notations and formulate the two related problems of individual treatment effect prediction and uplift modeling, explaining the difference between them and providing assumptions under which both have a common objective.

1.2.1 Notations

Let $X \in \mathbb{R}^d$ be a random variable of features and $\mathbf{x} \in X$ be its realization, denoting feature vector that characterised an individual, and $Y \in \mathbb{R}$ be the outcome variable (in this thesis, both continuous and binary outcomes were used, however, determining the outcome as continuous is sufficient to define CATE and uplift). Additionally, let the treatment variable $T \in \{0, 1\}$ denote whether an individual receives treatment ($t = 1$) or not ($t = 0$), so we assume the dataset $(\mathbf{x}_i, y_i, t_i) \overset{\text{iid}}{\sim} P_{X,Y,T}$. We define then $S_t = \{\mathbf{x}_i, y_i, t\}_{i=1\ldots n_t}$ as the particular subset of the training set S of size N, i.e. $S = S_1 \sqcup S_0$ and $N = n_1 + n_0$.

Following the *potential outcomes framework* [89], 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, we denote $Y(0)$ and $Y(1)$ the underlying random variables.

1.2.2 Conditional Average Treatment Effect Prediction

The **individual treatment effect (ITE)** of the individual i is given by the difference of its potential outcomes:

$$\tau_i = y_i(1) - y_i(0), \tag{1.1}$$

Note, that among the two potential outcomes, only $y_i = y_i(t_i)$ – the *factual* outcome – is observed in practice, and never both. The unobserved potential outcome $y_i(1 - t_i)$ is often called *counterfactual* of the observed outcome. In the community, this issue is referred to as the *Fundamental Problem of Causal Inference* (FPCI). Thus, FPCI prevents the access to the true value of the ITE.

Since individuals are described by vectors of features $\mathbf{x}$ it is rather possible to estimate the **conditional average treatment effect (CATE)** defined for any $\mathbf{x} \in X$ as:

$$\tau(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) \mid X = \mathbf{x}]. \tag{1.2}$$

The best CATE estimator is at the same time the best estimator for ITE in terms of the mean squared error, as shown in [63].

1.2.3 Uplift Modeling

A practical-oriented branch of the work with CATE prediction, is **uplift modeling**. Uplift is a term in the marketing application which refers to incremental impact of the treatment (promotion, online banner, etc.).

Formally, for the individual $\mathbf{x}$, the *causal* uplift $U(\mathbf{x})$ is defined as the difference in outcome *should* the individual be selected to take the treatment or not:

$$U(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, do(T = 1)] - \mathbb{E}[Y \mid X = \mathbf{x}, do(T = 0)] \tag{1.3}$$

where $do(\cdot)$ is an intervention operator defined in Pearl's causal inference framework [79].

In turn, the *conditional* uplift $u(\mathbf{x})$ is the expected difference in outcome *when* the individual has taken the treatment or not, that is when we observe it after the fact:

$$u(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 1] - \mathbb{E}[Y \mid X = \mathbf{x}, T = 0] \tag{1.4}$$

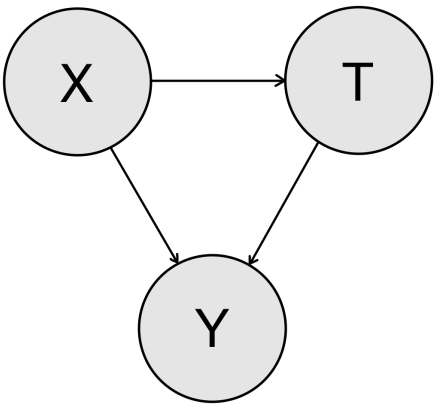
Causal and conditional uplifts are equivalent if treatment is administered to individuals at random:

$$X \perp\!\!\!\perp T \Rightarrow U(\mathbf{x}) = u(\mathbf{x}).$$

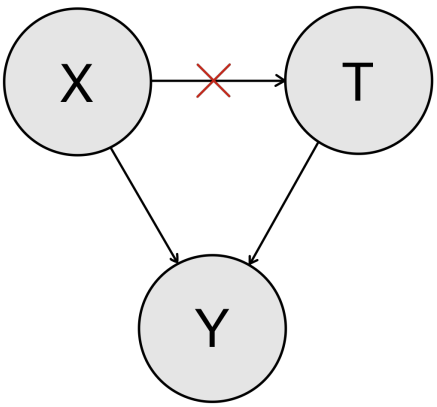
In other words, the data should be obtained from a randomized control trial (RCT), or online controlled experiment (also called A/B test), where individuals are randomly split in two populations: the treatment population which receives the treatment and the control population which does not.

1.2.4 Tasks comparison

Generally, both CATE prediction and uplift modeling have the common goal of determining how changing treatment affects changing outcome. However, a subtle difference between the two can be traced in the data generating process, as can be seen



(a) Causal graph for observational data. Treatment assignment that depends on the covariates leads to selection bias problem.



(b) Causal graph for experimental data. Intervention on treatment could be performed by assigning treatment randomly.

Figure 1.2: Causal graphs for observational and experimental data.

in Figure 1.2: methods developed in CATE prediction community are usually developed for *observational* data (Fig. 1.2a) – therefore these methods are often aimed at solving an additional problem of selection bias between treatment and control populations. In uplift modeling, on the other hand, *experimental* data (Fig. 1.2b) from RCT is assumed, in order to preserve the equivalence between causal and conditional uplift.

As described in [110], there is a set of assumptions, under which objectives of two tasks are equivalent (note that even in this case two tasks are different, as CATE prediction still contains selection bias problem):

- **Overlap**: Any subject has a non-zero probability of receiving or not to receiving the treatment:

$$0 < \mathbb{P}(T = 1 \mid X = \mathbf{x}) < 1 \tag{1.5}$$

- **Stable unit treatment value (SUTV)**: treatment applied to one subject does not affect the outcome of other subjects.

- **Unconfoundedness**:

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X = \mathbf{x} \tag{1.6}$$

We believe that the tasks of CATE prediction and uplift modeling are unreasonably segregated from each other and that they should interact more closely. Notably,

series of model evaluation approaches from uplift modeling can be used in CATE prediction and vice-versa (see Section 2.1) while some of the metrics are related being motivated by different goals (see Section 2.1.3). Besides, particular CATE and uplift models are "universal" in a sense they might be efficiently integrated in both problems (see Section 2.2). However, it is hard to merge the tasks completely as considered conditions of both may still have minor but sufficient differences.

Detailed surveys of the CATE prediction and uplift modeling are provided in [50, 37, 110].

Remark The problem of uplift modeling is the main focus of this thesis – hence throughout the work more attention is given to uplift models with corresponding performance metrics and more contributions are performed to this very problem. In the meantime, several contributions in the thesis address the problem of CATE prediction along with uplift modeling (under the aforementioned assumptions), since the two problems are ultimately closely related.

1.3 Thesis structure

The rest of the thesis consists of five chapters:

- **Chapter 2** represents the background of uplift modeling and CATE prediction, including description of existing models, comparison of different model evaluation approaches and overview of the datasets.
- In **Chapter 3**, a new large-scale benchmark is presented, notably, we provide details on the data collection, sanity checks that validate the use of this data and formalize how uplift modeling and CATE prediction can be performed with this data, along with the relevant evaluation metrics.
- **Chapter 4** contains the data representation-based uplift models for the cases when treatment is imbalanced throughout the data, along with the evaluation of proposed models.
- In **Chapter 5**, the problem of direct optimization of popular uplift modeling metric, namely Area Under the Uplift Curve (AUUC) is considered. In particular, data-dependent generalization error bound for AUUC is derived, using the relation with bipartite ranking, then the corresponding learning objective is presented that optimizes this bound.

- **Chapter 6** describes the problem of learning uplift and CATE prediction models in scenarios when access to both labels and features is available only through aggregated queries upon groups, motivated by a recent increase of privacy constraints in different domains. For this, a principled way to learn group-based uplift models from aggregated data is proposed. Besides, the bias-variance decomposition for the method is identified, highlighting the role of the underlying partition size in the privacy-utility trade-off.

- Finally, in **Chapter 7**, conclusions and future perspectives are provided.

1.4 Corresponding articles

Conferences

- "Uplift Prediction with Dependent Feature Representation in Imbalanced Treatment and Control Conditions" [17] – Artem Betlei, Eustache Diemert, Massih-Reza Amini, published at ICONIP 2018.

- "Uplift Modeling with Generalization Guarantees" [18] – Artem Betlei, Eustache Diemert, Massih-Reza Amini, published at KDD 2021.

Workshops

- "Dependent and Shared Data Representations improve Uplift Prediction in Imbalanced Treatment Conditions" - Artem Betlei, Eustache Diemert, Massih-Reza Amini, presented at Machine Learning for Causal Inference, Counterfactual Prediction, and Autonomous Action (CausalML), ICML 2018 workshop.

- "A large scale benchmark for uplift modeling" [39] - Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, presented at AdKDD, KDD 2018 workshop.

- "Optimization of treatment assignment with generalization guarantees" - Artem Betlei, Eustache Diemert, Massih-Reza Amini, presented at Causal Learning for Decision Making (CLDM), virtual ICLR 2020 workshop.

- "Differentially Private Individual Treatment Effect Estimation from Aggregated Data" [19] – Artem Betlei, Théophane Gregoir, Thibaud Rahier, Aloïs Bissuel, Eustache Diemert, Massih-Reza Amini, presented at Privacy Preserving Machine Learning, virtual ACM CCS 2021 workshop.

Submitted works

- "A Large Scale Benchmark for Individual Treatment Effect Prediction and Uplift Modeling" [40] – Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Theophane Gregoir, Thibaud Rahier, submitted to NeurIPS 2021 Conference, Datasets and Benchmarks Track.

Chapter 2

Background

In this chapter, we introduce detailed background in uplift modeling and CATE prediction. In particular, we provide the survey of model selection and evaluation techniques for the real-life cases, when there is no access to the true CATE or uplift, highlighting the links between some of them. Along with this, we describe the state-of-the-art uplift models and CATE estimators. Finally, we present existing available datasets in both fields.

2.1 Model evaluation techniques

One of the determinant problems in both CATE prediction and uplift modeling is the inability to observe a given individual in treated and untreated conditions simultaneously due to *Fundamental Problem of Causal Inference* (see Section 1.2.2). Evaluation metrics computing a difference to the true CATE can only work in a simulation setting, where both factual and counterfactual outcomes are available. A popular example of such a metric, used in variety of CATE prediction works [71, 94, 106, 107, 5], is the Precision in Estimation of Heterogeneous Effects (PEHE) [52]:

$$\epsilon_{PEHE}(\hat{\tau}) = \mathbb{E}\left[\left(\tau(X) - \hat{\tau}(X)\right)^2\right] \tag{2.1}$$

However, in real-life scenarios (which have the highest priority in this thesis) only factual outcome is observed, so the true uplift is inaccessible and one cannot utilize PEHE. Applying traditional performance metrics for the classifier or regressor underling CATE predictor or uplift model (e.g. accuracy/MSE of the prediction model) does not guarantee the accuracy of predicting uplift. In addition, since applications of ML models vary greatly from one CATE predictor to another, comparing different predictors in this way is not meaningful.

As a consequence, the process of selecting the right model in this situation also becomes complicated, as reliable model selection requires a model validation procedure.

For both tasks, precision of effect estimation might not be the final metric of interest. Depending on the ultimate goal, practitioners may be also keen to find a good treatment prescription rule or measure treatment prescription quality [112, 94, 12].

Accordingly, finding appropriate ways to evaluate CATE prediction and uplift models without having access to the true CATE or uplift is non-trivial, while being one of the most important problems in both fields.

2.1.1 CATE prediction

A comprehensive overview of model selection approaches for CATE prediction was proposed by [93]. To perform the selection, authors estimate the expected prediction risk for each of the given models and find the model that minimizes this risk. In particular, they include methods either explicitly designed for model selection or adapted from model learning procedures proposed for CATE prediction or policy learning.

Authors then divide approaches into three groups:

- First group of methods that *minimize risk of potential outcomes* include the predictive risk estimation (e.g. by MSE) separately for models $\hat{\mu}_0, \hat{\mu}_1$ and its weighted extension, in which MSE for individual i is divided by propensity score $e(\mathbf{x}_i)$ – for the cases when treatment assignment is conditional on observed covariates.

- In second group, different ways of *maximizing value of treatment policy* are presented, consisting of inverse propensity weighted (see Section 2.1.2.2) and doubly robust value estimators.

- Last, third group represents the approaches of *CATE risk minimization*. Here, three techniques are provided based on matched treated and control individuals, inverse propensity weighting, and Robinson decomposition (see Section 2.2.1) respectively.

Performing series of simulations on both randomized and observational data, authors claim that Robinson decomposition-based CATE risk minimizer, when optimized, most consistently leads to the selection of a high-performing model.

There are several alternative validation methods described in CATE prediction literature.

Influence Functions-based method [5] aims to validate CATE predictors by "predicting" the PEHE using Influence Functions – functional derivatives of a loss function with respect to the data distribution. By analogy to the regular derivatives, if we know PEHE of the model under some known distribution that is close enough to the true distribution, then we can estimate the true PEHE via influence functions using Taylor-like expansion as follows:

$$PEHE(\theta) \approx \underbrace{PEHE(\tilde{\theta})}_{\text{plug-in estimate}} + \underbrace{\dot{PEHE}(\tilde{\theta})\, d(\theta - \tilde{\theta})}_{\text{plug-in bias}} \tag{2.2}$$

where θ denotes a collection of nuisance parameters – conditional potential outcomes μ_0, μ_1 and propensity score e – of the true distribution, $\tilde{\theta}$ is plug-in model and $\dot{PEHE}(\tilde{\theta})$ is influence function of the functional $PEHE(\tilde{\theta})$.

We can briefly describe the procedure in two main steps:

1. **Plug-in estimation:** Fit the plug-in model $\tilde{\theta} = \{\tilde{\mu}_0, \tilde{\mu}_1, \tilde{e}\}$, then compute the plug-in estimate $PEHE(\tilde{\theta})$.

2. **Unplugged validation:** Use the influence functions of $PEHE(\tilde{\theta})$ to predict $PEHE(\theta)$.

For plug-in estimate, authors use Two-Models method (see Section 2.2.1) with gradient boosting algorithm XGBoost [29] in the role of base regressor for both groups and XGBoost classifier for the propensity score.

At the same time, for unplugged validation, authors provide closed-form expression of the first-order influence function [5, Theorem 2] and report that higher-order influence functions of PEHE are intractable.

Counterfactual Cross-Validation [92] is an alternative model selection technique. Proposed idea is that knowing the rank order of the performance of candidate predictors is enough to be able to select the predictor:

$$\underbrace{R_{true}(\hat{\tau}) \leq R_{true}(\hat{\tau}')}_{\text{true performance ranking}} \Longrightarrow \quad \underbrace{\hat{R}(\hat{\tau}) \leq \hat{R}(\hat{\tau}')}_{\text{ranking by evaluation metric}} \tag{2.3}$$

where $R_{true}(\hat{\tau})$ is true PEHE of the model $\hat{\tau}$ and $\hat{R}(\hat{\tau})$ is the potential evaluation metric of $\hat{\tau}$.

Similar to PEHE metric, they provide the following class of evaluation metrics:

$$\hat{R}(\hat{\tau}) := \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\tau}(X_i) - \hat{\tau}(X_i)\right)^2 \tag{2.4}$$

where $\tilde{\tau}$ denotes plug-in estimator, that should satisfy two conditions:

1. $\tilde{\tau}$ should be the unbiased estimator of true CATE
2. $\tilde{\tau}$ should have small expectation of conditional variance

For plug-in estimator authors combine Doubly Robust estimator (see Section 2.2.1) that meets Condition 1, with Counterfactual Regression (see Section 2.2.4), that minimizes expected conditional variance, satisfying Condition 2.

Policy Risk [94] is one more noteworthy metric which measures the risk to target treatment based on the policy implied by model $\hat{\tau}$. For this we can assume some threshold α and assign treatment for $\mathbf{x}$ if $\hat{\tau}(\mathbf{x}) > \alpha$. For the threshold α, policy risk $R_{pol}(\hat{\tau}, \alpha)$ takes the following form:

$$\begin{aligned} R_{pol}(\hat{\tau}, \alpha) = 1 - \Big(&\mathbb{E}[Y \mid \hat{\tau}(\mathbf{x}) > \alpha, T = 1]\mathbb{P}(\hat{\tau}(\mathbf{x}) > \alpha) \\ &+ \mathbb{E}[Y \mid \hat{\tau}(\mathbf{x}) \leq \alpha, T = 0]\mathbb{P}(\hat{\tau}(\mathbf{x}) \leq \alpha)\Big) \end{aligned} \tag{2.5}$$

2.1.2 Uplift modeling

In this subsection, we present an overview of the evaluation metrics provided in uplift modeling literature, which is applicable when the true uplift is unknown.

2.1.2.1 Metrics based on the group uplift

The main concept of this family of metrics is the ability to estimate group-level uplift instead of infeasible, individual-level one. Intuitively, with a good model, individuals with high true uplift should yield a high prediction. This leads to the natural idea to rank the whole population by their uplift score by descending order and compute group-level uplift for a certain proportion of individuals:

$$\hat{u}(X_g) = \mathbb{E}[Y \mid X_g, T = 1] - \mathbb{E}[Y \mid X_g, T = 0] \tag{2.6}$$

where X_g denotes individuals of particular group g.

Several approaches are proposed in the literature on how to divide the population into parts, based on which group-level uplift will be computed. For instance, dividing individuals into 10 bins and computing **uplift per decile** is popular in the marketing community [69, 51, 65]. Having such a division, practitioners may focus more on the behavior of the first k deciles in order to find a good prescription rule, or on the patterns of how the decile uplifts change – naturally, the good uplift model implies a monotonically decreased pattern. **Kendall's Uplift Rank Correlation (KURC)** [15] is one of the metrics trying to catch such a pattern by measuring the correlation between the predicted uplift and the observed one. KURC is defining as follows:

$$\rho = \frac{2}{K(K-1)} \sum_{i<j} \text{sign}\left(\bar{\hat{u}}_i - \bar{\hat{u}}_j\right) \text{sign}\left(\bar{u}_i - \bar{u}_j\right), \tag{2.7}$$

where K is number of bins, $\bar{\hat{u}}_k$ is the average predicted uplift in the bin $k, k \in \{1, ..., K\}$ and $\bar{u}_k$ is the true uplift in the bin k.

While comparison of uplift above a fixed cutoff is a simple and convenient metric, it is not satisfactory enough when the goal is to measure the uplift prediction itself.

As a step forward, the paradigms of Qini [82] and Uplift [90] curves were proposed later with the ideas to extend the "uplift per decile" approaches by adding the ability to compute uplift for any ratio of users. Two curves are quite similar to each other, however, we will describe both of them for completeness.

Let $\hat{u}$ be the uplift model. $\hat{u}(S_1, k)$ and $\hat{u}(S_0, k)$ denote individuals in S_1 and S_0 respectively among the top $100 \cdot k$ percent of whole population, ordered by prediction of $\hat{u}$ with

$$R_1^k = \sum_{i \in \hat{u}(S_1,k)} y_i \text{ and } R_0^k = \sum_{i \in \hat{u}(S_0,k)} y_i$$

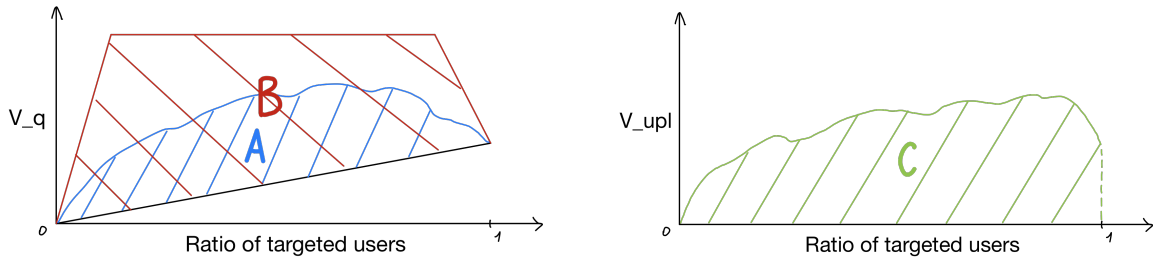
Then, the value related to the cumulative group uplift is expressed as:

$$V_q(\hat{u}, k) = R_1^k - \frac{n_1^k}{n_0^k} R_0^k. \tag{2.8}$$

Note that the resulting value serves as a quality metric of treatment prescription rule (as k corresponds to particular threshold for treatment prescription), being similar by intuition to a group of methods which maximize the value of treatment policy from [93] (see Section 2.1.1).

The Qini curve is drawn by varying k from 0 to 1. Once it is built, one can compute the area under the curve (AUC) as a way to summarize model performance.

There are several building blocks remain to introduce, in order to present the final metric. Let u^{random} be the model, assigning treatment to the users randomly –



(a) Q (b) AUUC

Figure 2.1: Illustrative example of Qini & Uplift curve-based metrics. Q is the ratio of the areas $\frac{A}{B}$ (left), while AUUC is the area C.

this model will act as a baseline. Besides, let u^* be the model that induces optimal ranking of users, namely:

$$u^*(S_1^+) > \left(u^*(S_1^-) \cup u^*(S_0^-)\right) > u^*(S_0^+), \tag{2.9}$$

where S^+ and S^- indicate subsets of positives and negatives respectively. Note that u^* should be considered as an "oracle" model, which might be infeasible to find even theoretically (assume for instance the case when data contains two identical feature vectors with different treatments - no model can separate them).

Qini coefficient Q [82] is proposed as the following ratio:

$$Q(\hat{u}) = \frac{AUC(\hat{u}) - AUC(u^{random})}{AUC(u^*) - AUC(u^{random})} = \frac{\int_0^1 V_q(\hat{u}, x)dx - \int_0^1 V_q(u^{random}, x)dx}{\int_0^1 V_q(u^*, x)dx - \int_0^1 V_q(u^{random}, x)dx} \tag{2.10}$$

In other words, Q quantifies how much better the model is, comparing to the random treatment targeting and at the same time how close the resulting ranking is to the optimal one.

The Uplift curve focus on the direct group uplift (avoiding the weighted term from Equation (2.8)), so the point of the curve represents the following:

$$V_{upl}(\hat{u}, k) = R_1^k - R_0^k \tag{2.11}$$

Similarly to the previous case, one can compute the **Area Under the Uplift Curve (AUUC)** [90]:

$$AUUC(\hat{u}) = \int_0^1 V_{upl}(\hat{u}, x)dx \tag{2.12}$$

Besides the described metrics, uplift modeling literature yields multiple modifications of Qini and Uplift curves, with differences residing mainly in i) the way

treatment imbalance is accounted for; and *ii*) whether treated and control groups are ranked separately or jointly. Readers can refer to [38, Table 2] for a comprehensive picture of available alternatives.

2.1.2.2 Inverse Propensity Weighting-based metrics

Metrics of this family are designed mainly for measuring treatment prescription quality. **Expected Response** is introduced in [112] for experimental data cases as an expectation of the outcome of individuals, for which original and prescribed treatment are the same. Formally, random variable Z is as follows:

$$Z = \sum_{t=0}^{K} \frac{1}{e_t} Y \cdot \mathbb{1}[h(X) = t] \cdot \mathbb{1}[T = t] \tag{2.13}$$

where e_t is probability of assigning treatment t and $h(X)$ is treatment prescription policy using uplift model (e.g. $h(X) = \mathbb{1}[\hat{u}(X) > 0]$). Note that K may be greater then 2, allowing to apply metric for multi-treatment problems.

Along with Z, its sample average $\bar{z}$ is unbiased estimator of expected outcome under policy $h(X)$:

$$\mathbb{E}[\bar{z}] = \mathbb{E}[Y \mid T = h(X)] \tag{2.14}$$

Based on $\bar{z}$, authors also proposed the modification of the Uplift curve: subjects are ranked by uplift scores from highest to lowest and $\bar{z}$ is computed cumulatively for different top ratios of the population.

2.1.3 Connections between evaluation techniques of two tasks

As CATE prediction and uplift modeling share the objectives, many of described evaluation approaches are interchangeable between the tasks – thus, some uplift model evaluation methods are reported in CATE model selection surveys [93].

Moreover, we derive the connection between CATE prediction metric Policy Risk and Uplift curve:

Proposition 1 (On the similarity between R_{pol} and V_{upl})**.** *Assume an equivalence between threshold α and top ratio k. The R_{pol} can be expressed similarly to V_{upl} as a weighted difference between R_1^α and R_0^α as follows:*

$$R_{pol}(\hat{u}, \alpha) = 1 - \Big(\frac{P_\alpha}{n_1^\alpha} R_1^\alpha - \frac{1 - P_\alpha}{n_0 - n_0^\alpha} R_0^\alpha + \underbrace{\frac{n_0^+ (1 - P_\alpha)}{n_0 - n_0^\alpha}}_{const(\alpha)} \Big), \tag{2.15}$$

where $P_\alpha = \mathbb{P}(\hat{u}(\mathbf{x}) > \alpha), n_1^\alpha = \sum_{i:\hat{u}(\mathbf{x})>\alpha} y_i \mathbb{1}[t_i = 1]$ and $n_0^{1-\alpha} = \sum_{j:\hat{u}(\mathbf{x})\leq\alpha} y_j \mathbb{1}[t_j = 0]$.

Proof.

$$\mathbb{E}[Y \mid \hat{u}(\mathbf{x}) > \alpha, T = 1] = \frac{\sum_{i:\hat{u}(\mathbf{x})>\alpha} y_i \mathbb{1}[t_i = 1]}{\sum_{i:\hat{u}(\mathbf{x})>\alpha} \mathbb{1}[t_i = 1]} = \frac{R_1^\alpha}{n_1^\alpha},$$

$$\mathbb{E}[Y \mid \hat{u}(\mathbf{x}) \leq \alpha, T = 0] = \frac{\sum_{j:\hat{u}(\mathbf{x})\leq\alpha} y_j \mathbb{1}[t_j = 0]}{\sum_{j:\hat{u}(\mathbf{x})\leq\alpha} \mathbb{1}[t_j = 0]} = \frac{R_0^{1-\alpha}}{n_0^{1-\alpha}}$$

Having $R_0^{1-\alpha} = n_0^+ - R_0^\alpha$ and $n_0^{1-\alpha} = n_0 - n_0^\alpha$ we get

$$\begin{aligned} R_{pol}(\hat{u}, \alpha) &= 1 - \frac{R_1^\alpha}{n_1^\alpha} P_\alpha + \frac{n_0^+ - R_0^\alpha}{n_0 - n_0^\alpha}(1 - P_\alpha) \\ &= 1 - \Big(\frac{P_\alpha}{n_1^\alpha} R_1^\alpha - \frac{1 - P_\alpha}{n_0 - n_0^\alpha} R_0^\alpha + \frac{n_0^+(1 - P_\alpha)}{n_0 - n_0^\alpha}\Big). \end{aligned}$$

□

Based on Proposition 1, we can conclude that, following different intuition and types of usage, in the end, two metrics are estimating similar values, which is a clear indication of the possible use of Uplift curve in CATE prediction and Policy Risk in uplift modeling. Visual comparison between two metrics is provided in Figure 2.2.

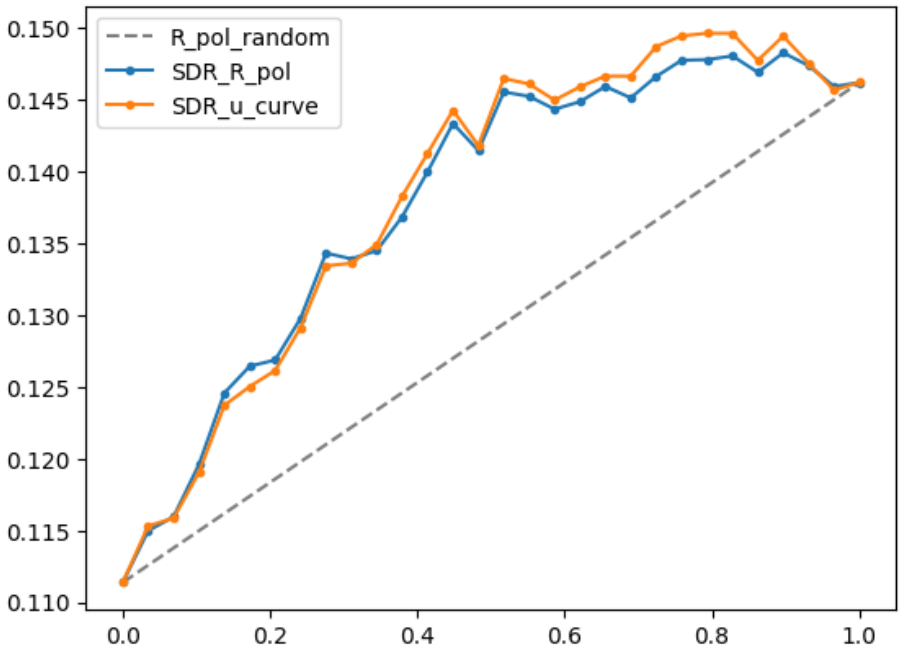


Figure 2.2: Example of curve based on R_{pol} (blue) by varying α and Uplift curve, based on V_{upl} (orange): for policy risk curve, we keep only the part subtracted from 1 of Equation 2.15. Also we calibrate two values to begin from the same point. Both metrics were computed for predictions of the fixed uplift model.

2.2 Existing CATE prediction and uplift models

We provide here the description of the most well-known methods in uplift modeling and CATE prediction, considering their strengths and drawbacks.

We remark that uplift models are often overlapping with CATE prediction techniques or reinvented independently, as the former is a subproblem of the latter (see Section 1.2.4). Besides, a majority of methods are interchangeable between the two tasks.

2.2.1 Model-agnostic methods

Model-agnostic methods (or meta-learners) is a family of methods for which any ML model can be applied as a base learner, there are different ways to compute the uplift though.

Two-Models approach [51] or T-Learner [63] is the most straightforward method to predict CATE (uplift). It uses two separate probabilistic models to predict outcome in treated or untreated conditions:

$$\hat{u}^{TM}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}) \tag{2.16}$$

where $\hat{\mu}_1(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 1]$ and $\hat{\mu}_0(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 0]$.

Any prediction model can be used for the estimation of posteriors and if both models perform well, the uplift model will also perform highly.

At the same time, the main goal of the models is to predict outcomes separately but not exactly the uplift. In cases where the average response is low and/or noisy, there is the risk for the difference of predictions to be very noisy too (see [83] for a detailed critic).

S-Learner [63] is a similar method to TM, estimating the outcome using the treatment indicator as an additional feature. Uplift then could be inferred as:

$$\hat{u}^{S-L}(\mathbf{x}) = \hat{\mu}(\mathbf{x}, 1) - \hat{\mu}(\mathbf{x}, 0) \tag{2.17}$$

where $\hat{\mu}(\mathbf{x}, t) = \mathbb{E}[Y \mid X = \mathbf{x}, T = t]$.

The method is quite simple to use, however, model $\hat{\mu}$ learns only a simple recalibration of the prediction for treated/control, which is usually not enough to find complex interaction between the features and treatment to explain the response. Besides, the importance of the treatment feature may be underestimated by the prediction model.

X-Learner [63] is a method that extends the Two-Models approach especially for scenarios where treatment is imbalanced, i.e. sizes of treatment and control groups significantly differ – in this case, a predictor trained on the smaller group may not be modeled accurately.

The idea of the method is first to predict the counterfactual outcomes for treatment individuals using the model, trained on a control population, and counterfactual outcomes for control individuals using the model, trained on a treatment population. This way, we can get so-called "imputed treatment effects" for each individual:

$$\hat{D}_i^0 = \hat{\mu}_1(\mathbf{x}_i) - Y_i \text{ (i belongs to } \textbf{treatment} \text{ group)} \tag{2.18}$$

$$\hat{D}_j^1 = Y_j - \hat{\mu}_0(\mathbf{x}_j) \text{ (j belongs to } \textbf{control} \text{ group)} \tag{2.19}$$

When imputed effects are computed, the next step is to estimate uplift in two ways by learning two separate regression models based on $\hat{D}^0$ and $\hat{D}^1$:

$$\hat{u}_0 = \mathbb{E}[\hat{D}^0 \mid X, T = 0] \tag{2.20}$$

$$\hat{u}_1 = \mathbb{E}[\hat{D}^1 \mid X, T = 1] \tag{2.21}$$

Finally, the uplift is estimated as the weighted average of $\hat{u}_0$ and $\hat{u}_1$:

$$\hat{u}^{X-L}(\mathbf{x}) = e(\mathbf{x})\hat{u}_0(\mathbf{x}) + (1 - e(\mathbf{x}))\hat{u}_1(\mathbf{x}) \tag{2.22}$$

where $e(\mathbf{x}) = \mathbb{P}(T = 1 \mid X = \mathbf{x})$ denotes probability of assigning treatment to an individual called *propensity score* [88]. $e(\mathbf{x})$ is constant for randomized control trials.

As the approach is tailored particularly for cases when the number of individuals in one group is much larger than in the other, several works [63, 110] confirm the effectiveness in these scenarios.

However, the need to train four prediction models (in observational data cases when one need to estimate propensity score $\hat{e}(\mathbf{x})$, amount of models increases to five) is a disadvantage of the method, as this entails a time-consuming process of hyperparameter tuning and the higher risk of overfitting.

Transformed Outcome methods use different modifications of the original outcome combined with treatment, to obtain a proxy of the true uplift and learn a single model. One frequently used transformation called in this work is the **Modified Outcome Method (MOM)** [11] that is relevant for any range of outcomes, based on the idea of *Inverse Propensity Weighting (IPW)* [54, 88]:

$$Y_i^{MOM} = Y_i \cdot \frac{T_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \tag{2.23}$$

A very important property of this transformation is that Y^{MOM} is an unbiased estimator of the true uplift:

$$\mathbb{E}[Y^{MOM} \mid X = \mathbf{x}] = u(\mathbf{x}) \tag{2.24}$$

and any regression model can be used to estimate uplift by learning Y^{MOM} as a label.

One disadvantage of the method is the ability to discard information by using values of the pairs (X_i, Y_i^{MOM}) instead of the triples (X_i, Y_i, T_i) – in particular cases one may estimate uplift more precisely by exploiting the information in the form of triplets (see Section 3.3 of [11]). Along with this, new outcome Y^{MOM} is suffering from large variance in cases where the variance of propensity score $e(\mathbf{x})$ is small.

Two similar approaches, which are the particular cases of MOM were proposed for binary outcomes, respectively **Four Quadrant method** [65] and **Class Variable Transformation (CVT)** [57]. Following different reasoning, at the end both methods proposed the same transformation:

$$Y_i^{CVT} = \begin{cases} 1 \text{ if } Y_i = 1 \wedge T_i = 1 \\ 1 \text{ if } Y_i = 0 \wedge T_i = 0 \\ 0 \text{ otherwise} \end{cases} \tag{2.25}$$

and uplift can be inferred as:

$$\hat{u}^{CVT}(\mathbf{x}) = 2 \cdot \mathbb{P}(Y^{CVT} = 1 \mid X = \mathbf{x}) - 1 \tag{2.26}$$

As the methods are proposed originally for the uplift modeling problem, the need for experimental data is assumed. For the imbalanced treatment cases, the sample weighting was suggested for CVT to learn the model [57].

R-Learner [75] where R for "residualized" and as an homage to Peter M. Robinson, applies orthogonalization to eliminate selection bias from observational data in two steps.

Firstly, so called nuisance components are estimated, as conditional mean outcome $\hat{\mu}(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}]$ and propensity score $\hat{e}(\mathbf{x})$. Then the uplift is obtained directly by minimizing the objective which is based on the Robinson decomposition [87]:

$$\hat{u}^{R-L}(\cdot) = \underset{u}{\operatorname{argmin}} \left\{ \hat{L}_n \left(u(\cdot) \right) + \Lambda_n \left(u(\cdot) \right) \right\}, \tag{2.27}$$

$$\hat{L}_n \left(u(\cdot) \right) = \frac{1}{n} \sum_{i=1}^{N} \Big(\big(Y - \hat{\mu}^{(-i)}(X_i) \big) - \big(T_i - \hat{e}^{(-i)}(X_i) \big) u(X_i) \Big)^2 \tag{2.28}$$

where $\Lambda_n\left(u(\cdot)\right)$ is interpreted as a regularizer on the complexity of $u(\cdot)$ function and superscript $(-i)$ denotes predictions made without using the i-th training example.

Any loss-minimization method, e.g., penalized regression, deep neural networks, or boosting can be used for each of the steps of R-Learner. Also, authors claim that R-Learner based on penalized kernel regression achieves the same regret bounds as an oracle with a priori knowledge of nuisance components.

At the same time, a drawback of the method is the need of using separate folds of the data for two stages, either by splitting the data or by using cross-fitting [31] to preserve theoretical guarantees. This usually degrades the performance, especially in the small data size cases.

Given nuisance components, R-Learner may also be considered as transformed outcome method with the new outcome of the following form:

$$Y_i^{R-L} = \frac{Y_i - \hat{\mu}(X_i)}{T_i - \hat{e}(X_i)} \tag{2.29}$$

with the sample weights $(T_i - e(X_i))^2$ for the learning. Note that Y^{R-L} is unbiased estimator of the true uplift:

$$\mathbb{E}[Y^{R-L} \mid X = \mathbf{x}] = u(\mathbf{x}). \tag{2.30}$$

DR-Learner [59] where DR for "doubly robust", can be considered as an extension of the Two-Models approach, adding inverse probability weighting term on the residuals of both prediction models.

At the beginning, data is divided into three parts of equal size. On the first stage, conditional mean outcomes $\hat{\mu}_0(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 0]$, $\hat{\mu}_1(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}, T = 1]$ and propensity score $\hat{e}(\mathbf{x})$ are learned based on first and second part of data respectively. Then, the outcome is transformed in the following way:

$$Y_i^{DR-L} = \frac{T_i - \hat{e}(X_i)}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \left(Y_i - \hat{\mu}_{T_i}(X_i)\right) + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \tag{2.31}$$

and is regressed based on third part of data. New outcome appears to be unbiased estimator of the true uplift:

$$\mathbb{E}[Y^{DR-L} \mid X = \mathbf{x}] = u(\mathbf{x}). \tag{2.32}$$

Authors also recommend applying cross-fitting by performing both stages three times varying data parts, so that the final estimator is an average of three regressors.

An estimator is consistent if either the propensity score or conditional mean outcome model is correctly determined. The need of using separate splits of the data for two stages seems to be a disadvantage of the method, as it is for R-Learner.

2.2.2 Tree-based methods

The tree-based methods for CATE prediction and uplift modeling are quite similar to the traditional ML models that are based on decision trees. The principal difference lies in applying the different splitting criteria. The comparison of splitting criteria of different approaches is provided in Table 2.1.

The advantages of tree-based methods are twofold:

- Methods provide direct and "transparent" uplift estimation, that is why they are interpretable
- Is it possible to control the granularity level (maximum number of data points in the leaf) to find the homogeneous subgroups

The key flaw of the methods lies in their instability due to the hierarchical nature of the process of splitting, moreover, instability is compounded by the fact that uplift modeling is a second-order problem [49].

Ensemble-based methods naturally extend tree-based ones by combining several trees into a single model. Following the idea of the regular ensembling technique, the model reduces variance by smoothing out individual tree errors, thereby increasing the stability.

Possible ways to apply bagging for the Two-Models with decision trees as the base learners for uplift decision trees [90] was proposed in [83, 96]. Several adapted versions of random forest for uplift modeling were also introduced: thus, a regular random forest algorithm based on uplift decision trees called Uplift Random Forest was explained in [49], besides [96] present Double Uplift Random Forest using bagged ensemble of the Two-Models with randomized trees as the base learners. In Causal Conditional Inference Forest [48], enhances Uplift Random Forest solving both overfitting and biased variable selection problems – the idea is to separate the variable selection and the splitting procedures along with using of statistically motivated and computationally efficient stopping criterion. Finally, Causal Forest was implemented in [103] based on causal trees [10]. The significant property of estimations of Causal Forest is that they are asymptotically Gaussian and unbiased for the true uplift.

Main drawbacks of ensemble-based methods are that they do not have the interpretability by design and tend to perform best with a large number of trees, which usually hurts the inference latency – in the case of small datasets this is not an important point, but it is not convenient for some applications, such as online advertising, that is the main domain of interest of this thesis.

Table 2.1: Comparison of different splitting criteria of tree-based methods for CATE prediction and uplift modeling. For the methods, $\hat{u}_L$ ($\hat{u}_R$) and n_L (n_R) denote respectively predicted uplifts and number of units for left (right) subtree. n is number of all units.

Method	Splitting Criteria	Notations/Comments
Uplift Incremental Value Modelling [51]	$\lvert\hat{u}_L - \hat{u}_R\rvert$	
Uplift Decision Tree [90]	$\frac{n_L}{n}\hat{u}_L^2 + \frac{n_R}{n}\hat{u}_R^2$	Designed for binary outcomes. Here, Euclidean distance is reported, two more criterias − KL divergence and χ^2 divergence − are also proposed in the paper.
Balance-Based Uplift Tree [83]	$\lvert\hat{u}_L - \hat{u}_R\rvert \cdot \left(1 - \lvert\frac{n_L - n_R}{n_L + n_R}\rvert^\alpha\right)$	$\alpha-$ hyperparameter Designed for binary outcomes, idea is to maximize the uplift difference in two subtrees minimising the gap in size between the nodes.
Significance-Based Uplift Tree [83]	$\frac{(n-4)\cdot(\hat{u}_L - \hat{u}_R)^2}{SSE\cdot\left(\frac{1}{n_L^T} + \frac{1}{n_L^C} + \frac{1}{n_R^T} + \frac{1}{n_R^C}\right)}$	$SSE-$ weighted sum of the population variances, Designed for binary outcomes, idea is to apply linear model to each candidate variable and compute significance of the interaction between treatment and potential splitting variable as measure of split quality.
Causal Tree [10]	$\left(\frac{n_L}{n}\hat{u}_L^2 + \frac{n_R}{n}\hat{u}_R^2\right) - \left(\frac{1}{n} + \frac{1}{n_s}\right)\cdot\left(\frac{S_L^{T\,2}}{p} + \frac{S_L^{C\,2}}{1-p} + \frac{S_R^{T\,2}}{p} + \frac{S_R^{C\,2}}{1-p}\right)$	n_s- size of splitting set, $p-$ average treatment in training set, S_{1L}, S_{0L} (S_{1R}, S_{0R}) − sample variances of treated and control units in left (right) subtree, idea is to divide training set into two parts − the split is defined based on units in splitting set while CATE within the node is computed based on units in the estimation set, CATE is estimated using inverse propensity weighting.

2.2.3 Support Vector Machines-based methods

There are several modifications of Support Vector Machines (SVM) method [33] tailored specifically for uplift modeling task, we provide the descriptions of two of them.

Uplift SVM [108] proposes SVM optimization task that has been reformulated to explicitly model the difference in class behavior between the treatment and control populations.

Method uses two parallel hyperplanes:

$$H_1 : \langle \mathbf{w}, \mathbf{x} \rangle - b_1 = 0, \quad H_2 : \langle \mathbf{w}, \mathbf{x} \rangle - b_2 = 0, \tag{2.33}$$

where $b_1, b_2 \in \mathbb{R}$ are the intercepts and $\mathbf{w}$ is the weight vector.
So the model predictions are specified by the following equation:

$$\hat{u}(\mathbf{x}) = \begin{cases} +1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle > b_1 \text{ and } \langle \mathbf{w}, \mathbf{x} \rangle > b_2, \\ 0 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle \leq b_1 \text{ and } \langle \mathbf{w}, \mathbf{x} \rangle > b_2, \\ -1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle \leq b_1 \text{ and } \langle \mathbf{w}, \mathbf{x} \rangle \leq b_2, \end{cases} \tag{2.34}$$

where function $\hat{u}(\mathbf{x}) : \mathbb{R}^m \to \{-1, 0, 1\}$ is an uplift model that predicts for each individual one of the values $+1$, 0 and -1, corresponding to positive, neutral and negative impact of the action respectively.

The intuition of the main objective is that the points in the neutral area are penalized for crossing one hyperplane, which prevents all points from being classified as neutral. Points that are misclassified are penalized for crossing two hyperplanes, and such points should be avoided.

Disadvantages of the approach include training complexity which is at least quadratic with respect to the standard SVM due to the additional hyperplane and corresponding slack variables.

SVM for differential prediction (SVM-DP) [64] is another variant of SVM, relevant for binary outcomes only, that aims to directly maximize frequently used metric in the field, namely Area Under the Uplift Curve (AUUC) (see Section 2.1.2.1). In the work, "absolute, separate" version of uplift curve [38] was used.

Authors propose to express AUUC as a difference between two Areas Under the Lift curve (AUL) of treatment and control populations (AUL_T and AUL_C respectively) [100]:

$$AUUC = AUL_T - AUL_C, \tag{2.35}$$

Applying the relations between AUL and Area Under the ROC Curve (AUC), we get the following form of AUUC maximization problem:

$$max\ (AUUC) \equiv max\ \left(AUC_T - \lambda AUC_C\right) \tag{2.36}$$

where $\lambda = \frac{n_T \cdot \sum_{i=1}^{N} Y_i(1-T_i) \cdot \sum_{i=1}^{N}(1-Y_i)(1-T_i)}{n_C \cdot \sum_{i=1}^{N} Y_i T_i \cdot \sum_{i=1}^{N}(1-Y_i)T_i}$.

One can obtain the sum instead of difference by reverting labels of control group:

$$max\ (AUUC) \equiv max\ \left(AUC_T + \lambda AUC_C^-\right) \tag{2.37}$$

where AUC_C^- denotes AUC of control group with reverted labels.

Finally, optimization problem (2.37) is solved by utilizing the SVM^{perf} algorithm [58], which is designed to directly optimize AUC.

Drawbacks of the method are the time-consuming learning process and the lack of probabilities of belonging to the class.

2.2.4 Deep learning-based methods

Deep learning-based methods have proven to be efficient in finding non-linear interactions between covariates and outcomes, that is why nowadays they have become very popular in a variety of applied domains. CATE estimators based on deep learning are relatively new but have already established themselves in the field.

Counterfactual Regression (CFR) [94] is a deep learning-based method that extends the Two-Models approach. In the beginning, whole training data is propagating through shared layers, resulting in representations Φ. Representations are divided then into two groups, depending on the treatment flag and thus forming two separate heads that are used to estimate outcomes under treatment and control. Importantly, to adjust for the bias induced by treatment group imbalance, representations Φ are balanced by minimizing the distance between treatment and control populations distributions respectively. Such a distance is also known as Integral Probability Metric (IPM). Authors utilize two IPMs: the Maximum Mean Discrepancy and the Wasserstein distance. Resulting architecture is called Counterfactual Regression Network (CFRNet) and is illustrated in Figure 2.3.

For the data from randomized control trials two populations are already balanced, so IPM becomes 0 and the task reduces to the problem of learning two functions of

potential outcomes. The corresponding, simplified version of the architecture is called Treatment-Agnostic Network (TARNet).

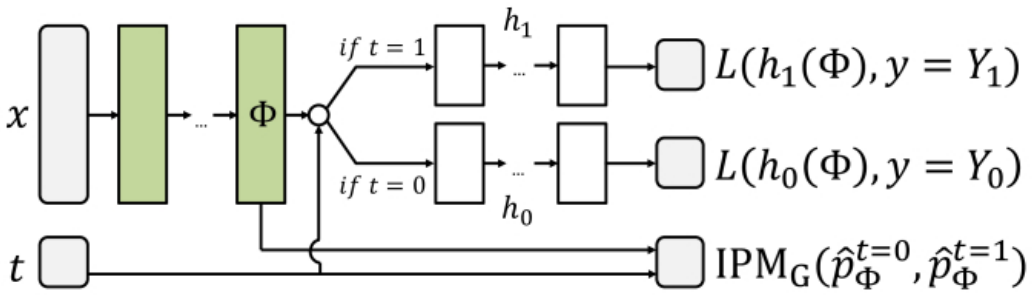


Figure 2.3: Architecture of CFRNet (figure is taken from [94].)

Generative Adversarial Network for inference of Individualized Treatment Effects (GANITE) [107] consists of two GAN blocks each of which consists of a generator and discriminator. In the first, counterfactual block, the generator produces counterfactual outcome Y^{cf} from input triplet $(\mathbf{x}, Y, T)$, at the same time the task of discriminator here is, giving the triplet $(\mathbf{x}, Y, Y^{cf})$ to recognize which outcome is factual. After the training of the counterfactual block, output in form of $(\mathbf{x}, Y, Y^{cf}, T)$ is propagating to the second, ITE block. There, the generator outputs both potential outcomes from input $\mathbf{x}$, discriminator then tries to identify if generated outcomes are the outputs from the counterfactual block. All the blocks are operating during the training stage, while only the ITE block generator is used to predict CATE for unseen data. Model architecture is shown in Figure 2.4.

Other deep learning-based algorithms include:

- DragonNet [95] – network with three separate groups of layers dedicated to the prediction of both potential outcomes and propensity score $e(\mathbf{x})$.
- Causal Effect Variational Autoencoder (CEVAE) [71], which uses a variational auto-encoder to learn a latent confounding set from the observed covariates, and then uses this set to estimate the CATE.
- Similarity preserved Individual Treatment Effect (SITE) [106] – estimation method based on the idea of CFR, that improves the learning of representations Φ by using a position-dependent deep metric and middle-point distance minimization constraints.

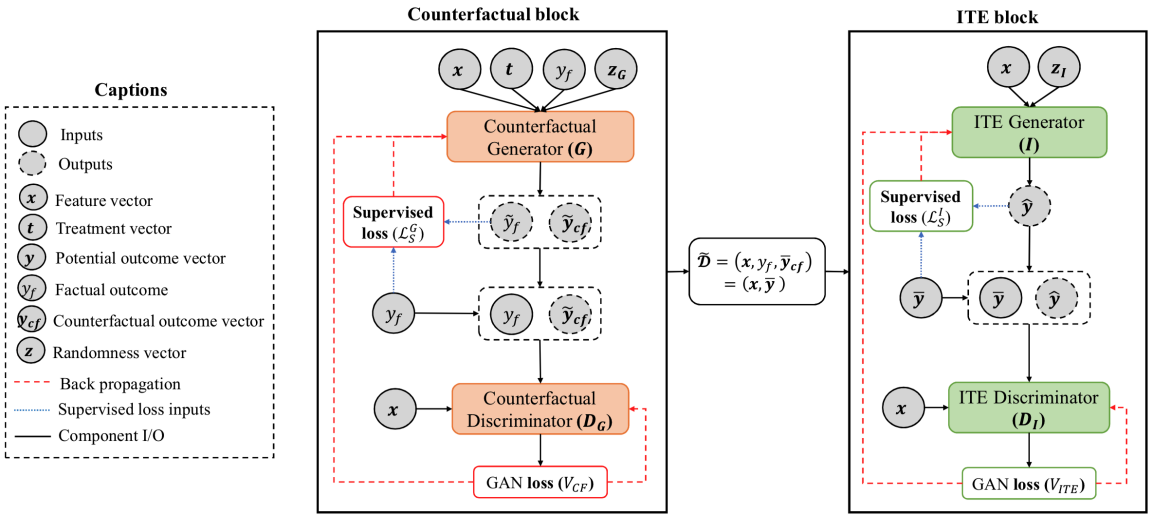


Figure 2.4: Architecture of GANITE (figure is taken from [107].)

2.2.5 Other methods

Other CATE and uplift models include family of bayesian methods [52, 6, 7], approaches applying interactions between treatment and covariates [69, 46], listwise ranking techniques [38], reinforcement learning approaches [68]. In addition, series of works are focused on the methods for multitreatment case [91, 112, 113, 76].

2.3 Overview of datasets

In this section, an overview of public CATE prediction and uplift modeling datasets is provided.

2.3.1 CATE prediction

Infant Health and Development Program (IHDP) [52] is semi-synthetic dataset adapted for CATE prediction. Covariates are obtained from a randomized experiment studying the effects of specialist home visits on future cognitive test results. The treatment group has been made imbalanced artificially, by removing a biased subset of the treatment group. The dataset comprises 747 units (139 treated, 608 control) and 25 covariates estimating the aspects of children along with their mothers.

Jobs dataset [94] is classification dataset, which is made as a combination of Lalonde randomized controlled trial (297 treated records and 425 control records)

and Panel Study of Income Dynamics (2490 control records). Each unit includes 8 covariates, such as age, education, ethnicity, previous earnings, etc. The outcome is the employment status with or without job training (acting as a treatment).

Twins dataset [8] is mined from all births in the USA between 1989 and 1991 years. Practitioners usually focus on the same gender twin pairs which weights less than 2000g. There are 46[1] covariates in the dataset, related to the information about the parents, the pregnancy, and the birth. Treatment T = 1 denotes the heavier one in the twins, and T = 0 – the lighter one. The outcome is determined as the one-year mortality. Note, that for this dataset, the true CATE is known as twins are characterized by the same feature vectors.

2.3.2 Uplift modeling

Hillstrom e-mail marketing dataset [53] comprises results of an e-mail campaign for an Internet-based retailer. The dataset contains information about 64,000 customers who last purchased within at most twelve months. The customers were involved in an e-mail test and were randomly assigned to receive an e-mail campaign featuring men's merchandise, women's merchandise, or not receive an e-mail.

X5 RetailHero dataset [47] contains raw retail customer purchases, raw information about products and general info about customers. The dataset was provided by X5 Retail Group at the RetailHero 2019 hackaton.

[1] We provide here data characteristics based on the work [71], meanwhile several works [107, 106] perform pre-processing of the dataset differently resulting in characteristics that differ from described version

Chapter 3

A Large-Scale Benchmark for Uplift Modeling and CATE Prediction

This chapter is based on submitted paper [40]: "A Large Scale Benchmark for Individual Treatment Effect Prediction and Uplift Modeling" – Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Theophane Gregoir, Thibaud Rahier, submitted to NeurIPS 2021 Conference, Datasets and Benchmarks Track.

3.1 Motivation

Generally speaking the benefit of having more datasets at hand is that it enables to draw more robust conclusions when experimenting new methods as algorithms are run in a variety of settings with different characteristics. In particular when very few benchmarks are available in a very active field such as causal inference there is always a risk of "conceptual overfitting" as the research community tries to beat the state of the art on few datasets representing only a very narrow range of real applications. Moreover, the lack of public, realistic datasets prompts sometimes researchers to publish results on private data, thus making unreproducible claims. Having large scale datasets is also a blessing for researchers in that they can run experiments on new methods with a greater chance to capture significant performance differences as the variance of metrics dwindles with size.

Our dataset brings a size increase for CATE prediction of 4 orders of magnitude and 2 orders for uplift modeling compared to established benchmarks. In terms of covariate dimensionality it brings a much harder setting with some features having thousands of possible values, which is more representative of modern problems in Web

applications. In the same line, the dataset is proposing challenges in both target and treatment imbalance with only a small portion of the users assigned to the control population and an overall low positive outcome rate. The anonymization strategy employed on this dataset with features represented as hashed tokens is perhaps also a characteristic of future applications where privacy enhancing technologies are pervasive. Finally, the dataset and its documentation are being maintained and improved over time as for instance a new version has been released one year after the initial one to correct a problem potentially impairing fair evaluation of models (see section 3.3 for details).

For CATE prediction prediction specifically, the common usage is to design semi-synthetic experiments using real features and simulated outcomes defined by simple response surfaces (constant or exponential) [52]; we additionally propose here realistic surfaces enriching the diversity of existing propositions.

3.2 Contributions

1. We present publicly available large-scale dataset, CRITEO-UPLIFTv2. In the spirit of [45] we detail the key elements from the datasheet of the dataset.

2. We introduce new synthetic response surfaces inspired by real observations that permit to use our large scale dataset for CATE prediction.

3. We report experiments to validate key characteristics of the dataset.

3.3 CRITEO-UPLIFTv2 dataset

Dataset is publicly available on the Criteo website.

Motivation and supporting organization Criteo is a private company which has been committed to the advancement of reproducible Advertising Science for a long time with a track record of releasing 7 large scale datasets in the last 7 years, some of which became industry and academic standards. In general these datasets[1] are interesting in that they showcase problems at the frontier of current ML theory and practice, for instance high-dimensional, tera-scale prediction problems [34], counterfactual learning [67] and credit assignment learning [41]. In order to provide a realistic benchmark for uplift modeling, the Criteo AI Lab built a dataset through a

[1] see `https://ailab.criteo.com/ressources/` for the complete list of published datasets

collection of online controlled experiments (A/B tests) on the Web in order to better study the individual effect of advertising on ad clicks and sales. More precisely, the dataset is constructed by assembling data resulting from several *incrementality tests*, a particular Randomized Control Trial procedure where a part of an online population is prevented from being targeted by advertising whilst the other part is subject to it.

System description The system can be formally described by introducing the following variables: for a given user, X contains their features, T is the binary *treatment* variable, such that $T = 1$ for users in the treatment population and $T = 0$ for user in the control population, and E, V and C are binary variables respectively indicating if the user has had at least one *exposition* to advertisement, *visit* on the website or *conversion* during the A/B testing period (see Figure 3.1b for example timelines of such users). In Figure 3.1a, we present the underlying causal graph [78] associated to this system. It contains both conditional independence and causal relations for example, we see from the causal graph that the treatment (T) is independent on the user features (X), guaranteeing rightful causal effect identification. Additionally, the variables respect the following constraints − purely due to their definition in the online advertising context:

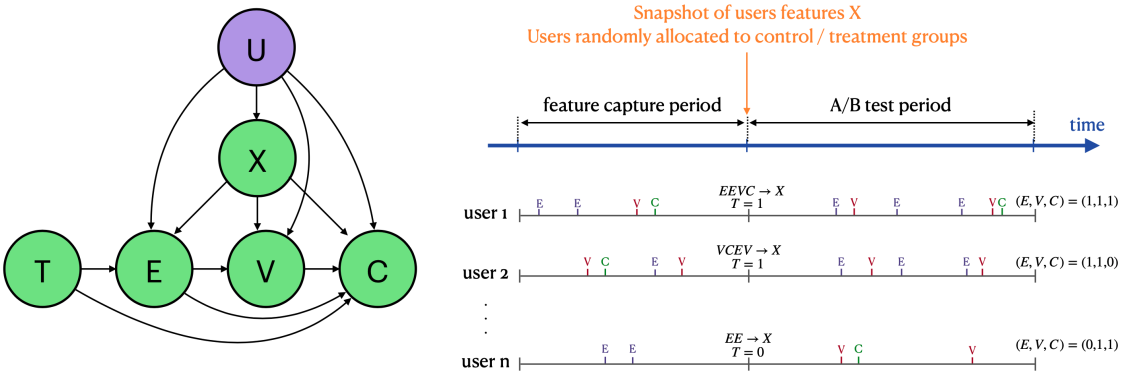
$$T = 0 \Rightarrow E = 0 \qquad \text{no exposition to ads in the control population}$$
$$V = 0 \Rightarrow C = 0 \qquad \text{no conversion can happen without a visit}$$

The online advertising context suggests some additional assumptions that enable more efficient CATE prediction − for example that the effect of T on C or V is only impacted by E [84] − which we will not detail further in this work.

Data collection As illustrated by Figure 3.1b, users − as identified by a browser cookie − leave online traces through advertiser events such as website visits or product views [62]. For a given advertiser, at a pre-defined point in time a random draw assigns each user either to the treated or control population. The period before this assignment is used to capture user features (mostly related to prior user activity). The set of features was chosen so that it is predictive of subsequent user events and we can easily verify from a technical standpoint that they were all captured in the legit period. Once treatment is assigned users are then either subject to personalized advertising (if treated) or not (if in control) until the end of the data collection period. During the first 2 weeks after treatment assignment ad visits and online conversions on the

advertiser website are logged. Then, features observed at treatment assignment times are joined with treatment prescription status, effective ad exposure and observed visits and conversion labels. Finally, the data for several such advertiser tests is merged to obtain the raw dataset.

Modern web platforms typically run numerous randomized experiments concurrently, yet users using these services are usually not fully aware of it. In our case we respected the Criteo Privacy Policy allowing users to opt out of the experiment at any point. The only drawback we can think of for users involved in the experiment was to avoid seeing ads, which is probably benign for most of us.



(a) Causal graph of the online advertising system

(b) Data collection process illustration

Figure 3.1: Data collection process and associated causal graph

Anonymization To protect Criteo industrial assets and user privacy neither test advertiser provenance nor features names and meaning are disclosed. Moreover, feature values were hashed to a random vector space to make them practically impossible to recover while keeping their predictive power. Non-uniform negative sampling on labels has been performed so that the original incrementality level cannot be deduced while preserving a realistic, challenging benchmark.

Considerations to avoid temporal confounding A particular characteristic of the current advertising systems is that they target users dynamically based on observed interactions over time [16]. This means that even in a randomized control trial (A/B test) interactions with the system influence subsequent ad exposure via adjustments of the bids based on user reactions. Notably, interactions after the first one are influenced both by the treatment and by previous interactions. This calls for

either considering only the first interaction of a user during an A/B test or to log the user variables at the start of the test and observe the reward during the test. We have chosen the latter solution as it enforces logging of features at the same time for all users, minimizing chances to observe temporal shifts in their distribution for reasons like sales periods or production platform evolution.

Considerations to concatenate data from different tests The incrementality tests of different advertisers had different treatment ratios, meaning that the features as well as the uplift were correlated with the fact of being part of a given test. In other words the (unreleased) test id was a hidden confounder of the (features, labels) x treatment distribution. To allow for proper use for uplift and CATE prediction we needed that all instances in the final dataset were drawn i.i.d. from the same $P_{X,Y,T}$ distribution. If not, a prediction model could have had a good score by learning to predict which test an instance was coming from and utilizing the fact that some tests were heavily imbalanced in terms of treatment to guess if a treated or untreated positive was more likely. That would have defeated the purpose of the dataset to serve as a realistic benchmark for uplift or CATE modeling. To remedy that situation we sub-sampled all incrementality tests at the same, global treatment ratio. That way the dataset scale is preserved and the task is kept close to reality. This rebalancing is the key difference between the previous version (v1) and v2 of the dataset[2] [39] and has been validated (see Section 3.5).

Dataset description and analysis The final dataset (v2), henceforth referred to as CRITEO-UPLIFTv2, consists of 14M rows, each one representing a user with 12 features, a treatment indicator, an effective ad exposure indicator and 2 binary labels (visits and conversions). The global treatment ratio is 85%, meaning only a small portion of users where observed in the control population for which no advertising is performed by Criteo. It is typical that advertisers keep only a small control population as it costs them in potential revenue. Positive labels mean the user visited/bought on the advertiser website during the test period (2 weeks). Positive ad exposure means the user effectively saw an ad for the advertiser during the label collection period. Among the 12 variables, 4 are continuous and 8 are categorical with a large number of modalities. In order to evaluate the impact of each feature on the visit outcome V, a random forest model (formed by 100 estimators) is trained on each of the treatment and control population to predict V. Then, for each feature, Mean

[2] v1 has been decommissioned and early users warned of that flaw

Decrease in Impurity (MDI) [23] is computed for both models and the corresponding average MDI is reported in Table 3.1. According to this experiment, f0, f2, f8, f9 appear to drive V significantly, while f1, f5, f11 have less influence.

	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
Number of modalities	cont.	60	cont.	552	260	132	1,645	cont.	3,743	1,594	cont.	136
Feature Importance (MDI)	0.110	0.003	0.297	0.036	0.020	0.007	0.064	0.018	0.218	0.170	0.046	0.010

Table 3.1: Summary of CRITEO-UPLIFTv2 feature characteristics. Feature importance is evaluated thanks to MDI which was computed using an average over two Random Forest models (formed by 100 estimators) respectively trained on treatment and control population.

3.4 Generation of synthetic surfaces for CATE prediction

In the spirit of [52], we propose a class of synthetic response surfaces as well as method to design confounded treatment assignment, providing a semi-synthetic version of our dataset, named CRITEO-ITE, that can be used as a benchmark for CATE models evaluation.

3.4.1 Response surfaces

We add two classes of synthetic response surfaces for the CRITEO-ITE.

First, we reproduce the popular semi-synthetic setups from [52]. In the Case 'A', constant treatment effect is generated by two linear response surfaces, namely $\mu_0(\mathbf{x}) = \mathbf{x}\beta$ and $\mu_1(\mathbf{x}) = \mathbf{x}\beta + 4$, where β is a coefficient vector with each component sampled from the same multinomial distribution. Case 'B' uses an exponential control response surface $\mu_0(\mathbf{x}) = \exp((\mathbf{x} + W)\beta)$ and a linear treatment response surface $\mu_1(\mathbf{x}) = \mathbf{x}\beta - \omega$, W here is a fixed offset matrix and ω is a real number, adjusted so that the average treatment effect on the treated (ATT) is consistent with real-world measures.

Second, we propose a novel *multi-peaked* (non monotonous) class of response surfaces in the spirit of radial basis function interpolation − inspired from observations made on projections of the real uplift surface (see Figure 3.2) − which define both a novel and challenging CATE modeling problem.

Formally, we suppose that X is equipped with a norm $||.||$ and define, for $t \in \{0, 1\}$ and $\mathbf{x} \in X$:

$$\mu_t(\mathbf{x}) = \sum_{c \in \mathcal{C}} w_{t,c} \exp\left(-\frac{||\mathbf{x} - c||^2}{2\sigma_c^2}\right), \tag{3.1}$$

where $\mathcal{C}$ is a set of *anchor points*, $\{w_{0,c}, w_{1,c}\}_{c \in \mathcal{C}}$ are the weights and $\{\sigma_c\}_{c \in \mathcal{C}}$ correspond to the width of influence associated to each of those points.

For any $\mathbf{x} \in X$, the associated CATE is therefore given by

$$\tau(\mathbf{x}) = \sum_{c \in \mathcal{C}} \exp\left(-\frac{||\mathbf{x} - c||^2}{2\sigma_c^2}\right)(w_{1,c} - w_{0,c}).$$

If the distance between the different anchor points are large compared to the values of the σ_cs, the CATE of each $c \in \mathcal{C}$ is $\tau(c) \approx w_{1,c} - w_{0,c}$ and for any $\mathbf{x} \in X$, $\tau(\mathbf{x})$ is a weighted sum of the $\tau(c)$'s with weights exponentially decreasing with the ratios $\frac{||\mathbf{x}-c||}{\sigma_c}$.

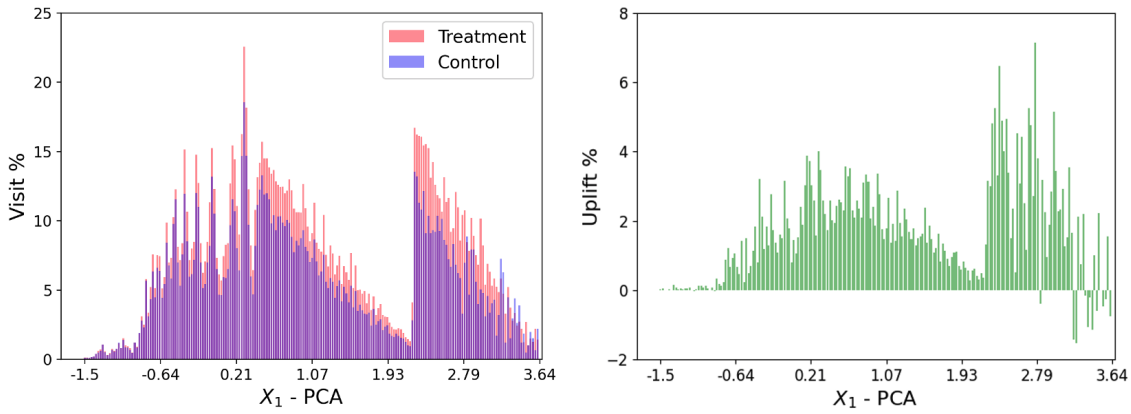


Figure 3.2: On the right, average uplift is reported as a function of the first component of Principal Component Analisys (PCA) computed on continuous features from CRITEO-UPLIFTv2 thanks to regular binning. On the left, average visit proportion is shown, computed with the same binning for control and treatment populations. Note the common multi-peak structure of both outcomes and uplift.

3.4.2 Treatment assignment mechanism

To simulate an observational setting, we design an heterogeneous treatment assignment function $p : \mathbf{x} \mapsto \mathbb{P}(T = 1 \mid X = \mathbf{x})$ so that T is confounded with the outcome Y (note that the case where p is constant corresponds to the RCT setting).

We propose a simple way to introduce treatment assignment bias by making p depend on the component of x which has the most predictive power with the outcome

Y. Specifically, for a given small $\delta > 0$ we define

$$p(\mathbf{x}) = (1 - 2\delta) \cdot \text{sigmoid}(\alpha^T \mathbf{x}) + \delta, \tag{3.2}$$

where $\alpha = (0, \ldots, 0, 1, 0, \ldots, 0)$ is a sparse $d-$dimensional vector for which the only nonzero component is the one which corresponds to the highest importance component of $\mathbf{x} \in X$ for the prediction of $\mathbb{E}[Y \mid X = \mathbf{x}]$.

This choice of treatment assignment mechanism guarantees that the *strong ignorability assumption* [88] is met since $p(\mathbf{x}) \in [\delta, 1 - \delta]$ for all $\mathbf{x} \in X$, and that all confounders between T and Y are contained in X, ensuring that

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X.$$

3.5 Experiments

In this part, we will be presenting experiments conducted on CRITEO-UPLIFTv2 and CRITEO-ITE datasets. First, sanity checks are performed in order to validate a correct collection and creation of CRITEO-UPLIFTv2. Then, we provide experiments on CRITEO-UPLIFTv2 highlighting the impact of dataset size on separability for uplift modeling task. Finally, we provide a benchmark of CATE prediction methods on classical response surfaces and ours.

3.5.1 Dataset Validation

We perform several sanity checks to verify properties of our dataset. First check is that the treatment is indeed independent of the features: $T \perp\!\!\!\perp X$. A convenient way to verify this assumption is to perform a Classifier Two-Sample Test (C2ST) [70]: a classifier trained to predict treatment should not do better than chance level. The distribution of H_0 in this case is obtained by computing the test loss of classifiers (we use log loss) trained to predict random splits in the data. Table 3.2 gives the result of the test. The empirical loss of the learned treatment classifier is very close to the dummy one from H_0, which is reflected by a high p-value for the one-sided test.

Second check is to make sure that logged features are informative and relevant for predicting outcomes (visit and conversion). This is not necessarily trivial as we sampled features that were technically easy to log and anonymized them. Table 3.3 presents the performance (as measured by log loss) of classifiers learned on the outcomes for treatment, control and the whole dataset. The non-trivial improvement over a dummy baseline indicates that features are indeed informative for the task.

Median Random Loss	Treatment Loss	p-value
0.42269	0.42307	0.13667

Table 3.2: Result of C2ST on treatment predictability with 300 resamples using log loss. The p-value does not allow to reject H_0 and confirms that $T \perp\!\!\!\perp X$.

visit, %	34.74
conversion, %	32.22

Table 3.3: Improvement over the log loss of a dummy classifier for different labels. Baseline is a classifier predicting the average label, and improvement is relative to baseline.

3.5.2 Uplift Modeling

Features. In order to reach a reasonable running time while conserving the great feature complexity of CRITEO-UPLIFTv2, the features used here are formed by the 4 initial continuous features and 100 projections on random vectors of the categorical features which are then one-hot encoded.

Target. To train uplift models, both visits and conversions are available as the labels. However, as presented here, we suggest practitioners to model uplift primarily on visits in so far as conversion uplift signal appears to be too weak due to the high imbalance in the label.

Metric. We pick as a performance measure the "separate, relative" AUUC – evaluations of [38] concluded robustness of this version to treatment imbalance and its ability to capture the intended usage of uplift models to target future treatments. Confidence intervals are computed using AUUC test set bound [18].

Protocol. The focus of this experiment is not on providing the best possible baseline but rather to highlight the fact that CRITEO-UPLIFTv2 is a meaningful alternative to existing uplift modeling datasets, scaling up in challenge while permitting to obtain statistical significance in the results. For this reason, we use 80%/20% train/test splits and AUUC performances were compared on test subsamples of proportional sizes to existing datasets, namely 1000 (IHDP), 5000 (Jobs), 50000 (Hill-

strom), 1M and whole test data. Besides, the training set is used to tune the baseline models via grid search combined with stratified 5-fold cross-validation (to save both treatment and outcome imbalance).

Models. Four uplift models were used as a baselines: Two-Models (TM) [51], Class Variable Transformation (CVT) [57], Modified Outcome Method (MOM) [11] and Shared Data Representation (SDR) [17]. Particular prediction models (from scikit-learn [80]) and hyperparameter grids are the following:

- TM, CVT – Logistic Regression, l_2 regularization term C : $[1e^0, 1e^2, 1e^4, 1e^6, 1e^8]$
- MOM – Ridge, l_2 regularization term α: $[1e^{-8}, 1e^{-6}, 1e^{-4}, 1e^{-2}, 1e^0]$
- SDR – Logistic Regression, l_2 regularization term C : $[1, 10, 100, 1000]$, feature importance term λ: $[0.1, 1]$

Results. Figure 3.3 represents results of the experiment. For the test sizes up to 1M, all presented methods are indistinguishable by their AUUC score as their confidence intervals overlap almost entirely. However, starting from 1M points onwards one can separate approaches and perform model selection. Hence it justifies the need for a large dataset for such a challenging task.

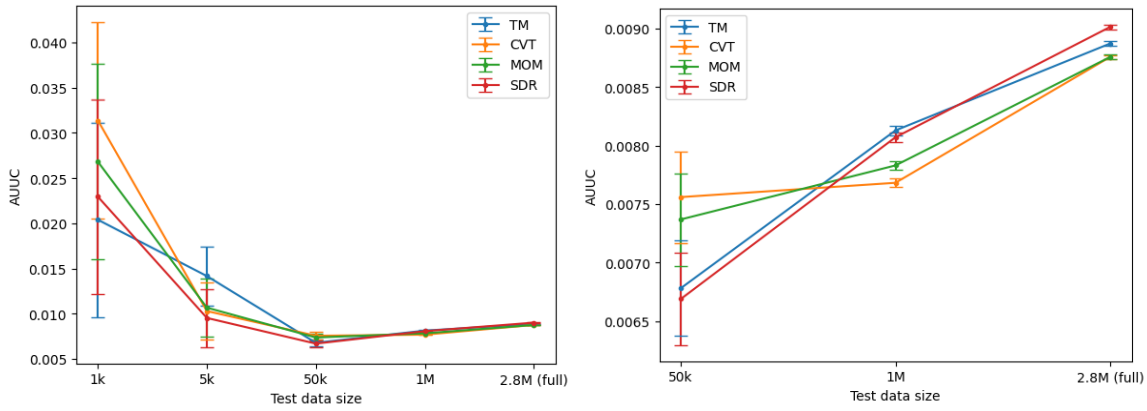


Figure 3.3: Models separability on CRITEO-UPLIFTv2.

3.5.3 CATE prediction

Features. Features used for response surface generation and prediction are the 4 initial continuous and 5 random projections which are then one-hot encoded leading to a total dimensionality of 32.

Target. In order to test CRITEO-ITE, we performed a benchmark of CATE prediction baselines evaluated on 3 generation protocols presented in Section 3.4:

- cases A and B from [52],
- our multi-peaked with 5 randomly selected anchor points, $\sigma_c = 1$ for all $c \in \mathcal{C}$ and $\{w_{0,c}, w_{1,c}\}_{c \in \mathcal{C}}$ drawn from $\mathcal{U}(0, 1)$ then fitted to ensure $ATE \approx 4$ as for the other surfaces.

For the 3 types of surfaces, we defined the same treatment bias consisting in a sigmoid on the highest importance feature with $\delta = 0.01$ (see Section 3.4.2).

Metric. For each baseline, mean $\sqrt{\epsilon_{PEHE}}$ is reported with its standard deviation over 10 experiments.

Protocol. Following the concept of [94], for each of the 3 generation protocols, 10 different realizations are generated. Then, for each realization, using a subsample of 100,000 points, baseline models are tuned thanks to a 5-fold cross-validation and then tested (with 50/50 train/test split).

Models. For this experiment, the baseline models include deep models (TARNet, CFRNet [94]) and meta-learners (T-Learner, X-Learner, R-Learner, DR-Learner) with Random Forest [24] as prediction model [3]. TARNet and CFRNet were implemented in TensorFlow [2]. They were trained for 20 epochs during cross validation and finally for 100 epochs on the entire training set. Batch size was set to 128. For this two deep models, the following hyper parameters were tuned thanks to a randomized search :

- number of layers : $[2, 3]$
- number of units per layer : $[32, 64]$
- regularization term : $[1e^{-4}, 1e^{-6}]$
- IPM regularization term (CFRNet only) : $[1e^{-2}, 1e^{-4}]$

Concerning meta-learners, models were partially implemented in CausalML library [28] (released with Apache License, Version 2.0). Meta-learners are using Random Forest Regressors from scikit-learn [80] which were tuned thanks to a randomized search with the following hyper parameters :

- number of estimators : $[10, 20]$
- maximum depth in range : $[1, 2, 4, 8]$

Results. As illustrated by Table 3.4, performances differ from one type of surface to the other underlining the importance of developing a variety of responses in which our multi-peaked version can anchor. For example, although X-Learner outperforms R-Learner on surfaces from [52], our multi-peaked generation highlights the contrary.

[3] Code for this benchmark experiment is available on the project repository.

	Case A [52]	Case B [52]	Multi-peaked (ours)
T-Learner	0.317 ± 0.001	1.890 ± 0.001	0.333 ± 0.056
X-Learner	0.117 ± 0.001	1.883 ± 0.002	0.413 ± 0.137
R-Learner	0.272 ± 0.034	11.668 ± 2.897	0.356 ± 0.045
DR-Learner	$\mathbf{0.047 \pm 0.005}$	2.510 ± 0.342	0.379 ± 0.080
TARNet	0.104 ± 0.001	0.682 ± 0.067	0.195 ± 0.044
CFRNet (MMD)	0.057 ± 0.001	$\mathbf{0.239 \pm 0.029}$	$\mathbf{0.152 \pm 0.032}$

Table 3.4: CATE prediction experiments on CRITEO-ITE. Mean $\sqrt{\epsilon_{PEHE}}$ performances are reported alongside their standard deviation. Best performance is in bold.

3.6 Summary

We have highlighted the need for large scale benchmarks for causal inference tasks and released an open dataset, several orders of magnitude larger and more challenging than previously available. We have discussed the collection and sanity checks for its use in uplift modeling and CATE prediction. In particular we have shown that it enables research in uplift prediction with imbalanced treatment and response levels (e.g. visits) providing model separability due to its large size. We have also proposed semi-synthetic version of our dataset that can be used as a benchmark for CATE models evaluation.

Chapter 4

Data Representation Methods for Imbalanced Treatment Conditions

This chapter is based on published paper [17]: "Uplift Prediction with Dependent Feature Representation in Imbalanced Treatment and Control Conditions" – Artem Betlei, Eustache Diemert, Massih-Reza Amini, published at ICONIP 2018.

4.1 Motivation

Imbalanced classification is popular machine learning problem [22] that can be solved by using variety of different techniques such as over- or under-sampling, sample or class weighting, etc.

However, in uplift modeling or CATE prediction, except the outcome, treatment variable might be imbalanced as well. This situation often arises with the growth of administrative and online data sources due to the privacy concerns [63], or in fields of medicine or online advertising as a result of different cost of treating and non-treating an individual. Consequently, some models may lose quality for this reason.

4.2 Contributions

Our main contributions are twofold:

1. We introduce two novel model-agnostic approaches that tackle the case of imbalanced treatment and control datasets and discuss their merits.

2. We evaluate the proposed approaches on a real-life collection and produce palpable evidence of their practical usefulness.

4.3 Dependent Data Representation

Dependent Data Representation (DDR) approach is an extension of Two-Models approach. DDR is based on a Classifier Chains method [86] originally developed for multi-label classification problems. The idea is that if there are L different labels, one can build L different classifiers, each of which solves the problem of binary classification and at the training process each next classifier uses predictions of the previous ones as extra features (diagram is provided in Figure 4.1).

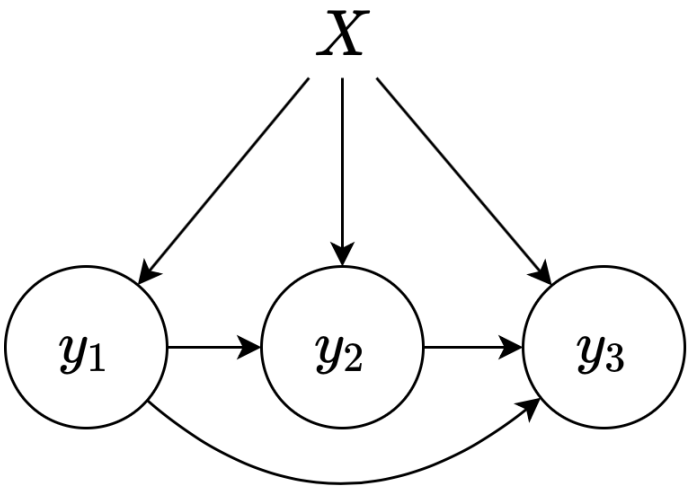


Figure 4.1: Diagram of classifier chains for $L = 3$.

We use the same idea for our problem in two steps. Let X_0 and X_1 be covariates of control and treated individuals respectively. At the beginning we train a first model μ_0 on control data:

$$\mu_0 = \mathbb{E}\left[Y = 1 \mid X_0\right] \tag{4.1}$$

then we use predictions $\hat{\mu}_0(X_1)$ as an extra feature for the model μ_1 learning on the treatment data, effectively injecting a dependency between the two datasets:

$$\mu_1 = \mathbb{E}\left[Y = 1 \mid X_1, \hat{\mu}_0(X_1)\right] \tag{4.2}$$

The diagram of the learning process of DDR is introduced in Figure 4.2.

In order to obtain uplift for each individual we compute the difference:

$$\hat{u}^{DDR}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}, \hat{\mu}_0(\mathbf{x})) - \hat{\mu}_0(\mathbf{x}) \tag{4.3}$$

Intuitively, the second model is learning the difference between the expected outcome in treatment and control, that is the uplift itself. Examination of the weights

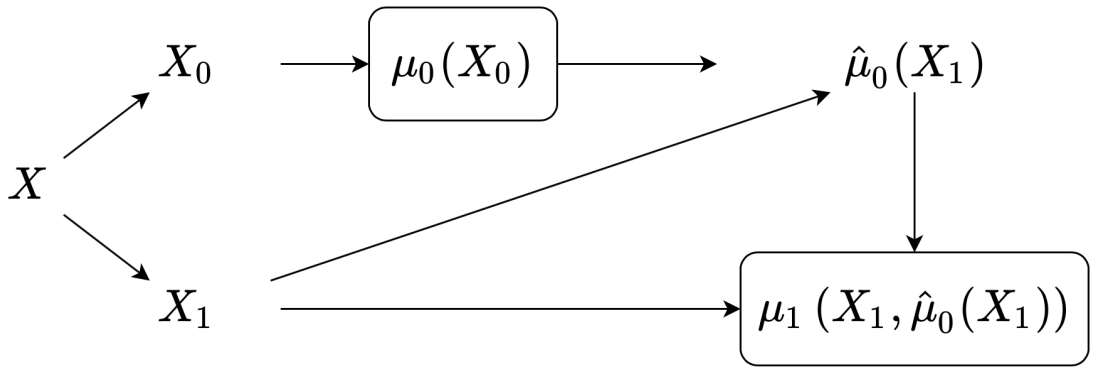


Figure 4.2: Diagram of the learning phase of Dependent Data Representation.

of this uplift model could also lead to interesting information on the role of different features in explaining the treatment outcome.

4.4 Shared Data Representation

Shared Data Representation (SDR) approach for uplift modeling is based on a popular implementation of the multi-task framework [26] and is model-agnostic. A predictor is learned on a modified features representation that allows to learn related tasks jointly and with a single loss. We specialize this approach considering predicting outcomes in control and treatment groups as the related tasks. [46] produced the method of similar design.

The general form of the model is given by

$$\mathbb{E}\left[Y \mid X = \mathbf{x}, T = t\right] = f(\langle \mathbf{w}_{common}, \mathbf{x} \rangle + \mathbb{1}_{[t=1]} \langle \mathbf{w}_1, \mathbf{x} \rangle + \mathbb{1}_{[t=0]} \langle \mathbf{w}_0, \mathbf{x} \rangle) \tag{4.4}$$

with f an arbitrary link function. Practically speaking we augment the dataset by stacking the original features with a conjunction of the treatment group indicator and the same features. Letting $\mathbf{X}_1$ and $\mathbf{X}_0$ be the matrices of covariates X_1 and X_0 respectively such that $\mathbf{X}_1 \cup \mathbf{X}_0 = \mathbf{X}$, we obtain the following shared learning representation:

$$\mathbf{X}_{train}^{SDR} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_1 & 0 \\ \mathbf{X}_0 & 0 & \mathbf{X}_0 \end{bmatrix}$$

and train classifier μ on resulting matrix.

So a single vector of weights $\mathbf{w}$ is learned jointly as

$$\mathbf{w} = [\mathbf{w}_{common}\ \mathbf{w}_1\ \mathbf{w}_0]$$

where $\mathbf{w}_{common}$ is a vector of weights that relate to the original features and $\mathbf{w}_1$ and $\mathbf{w}_0$ are corresponding to treatment/control conjunction features. Illustration of the learning phase of SDR is in Figure 4.3.

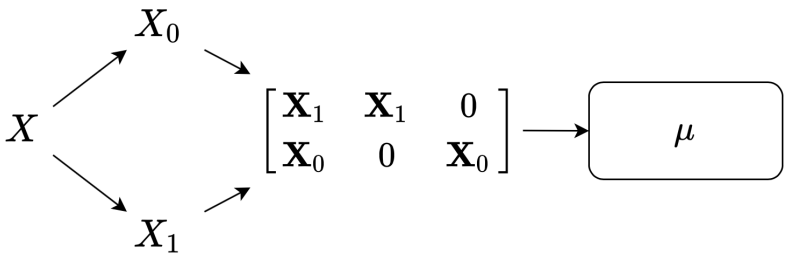


Figure 4.3: Diagram of the learning phase of Shared Data Representation.

At inference time we compute the difference between predicted probabilities using two representations of the individual features, corresponding to the counterfactual outcomes – as if new individual was treated or not:

$$\begin{aligned} \hat{u}^{SDR}(\mathbf{x}) &= \mu(\begin{bmatrix} \mathbf{x} & \mathbf{x} & 0 \end{bmatrix}) - \mu(\begin{bmatrix} \mathbf{x} & 0 & \mathbf{x} \end{bmatrix}) \\ &= \mathbb{E}\left[Y = 1 \mid \begin{bmatrix} \mathbf{x} & \mathbf{x} & 0 \end{bmatrix}\right] - \mathbb{E}\left[Y = 1 \mid \begin{bmatrix} \mathbf{x} & 0 & \mathbf{x} \end{bmatrix}\right] \end{aligned} \tag{4.5}$$

An advantage of this method is the possibility to assign different regularization penalties for $\mathbf{w}_{common}$ and $\mathbf{w}_1$ / $\mathbf{w}_0$. We define such penalties as λ_{common} and λ_{task} respectively. In this way it is possible to control the strength of the connection between the tasks. As reported in [26], it is equivalent to rescaling the conjunction features by $\sqrt{\frac{\lambda_{common}}{\lambda_{task}}}$.

Intuitively this model allows to learn a common set of weights for predicting the global average outcome whilst keeping enough capacity to express the peculiar influence of features in the treatment or control conditions.

4.5 Experiments

For the experiments, we have selected two real-world datasets. The first dataset is **Hillstrom** (see Section 2.3.2). We use the no-email vs women e-mail split with "visit" as outcome as in [83].

Our second dataset is **CRITEO-UPLIFT1**[1] which is constructed by assembling

[1]dataset is released at `http://research.criteo.com/outreach`. This version of dataset is used in due to the fact that this work preceded the work in Section 3, where the new version of dataset is presented.

data resulting from incrementality tests, a particular randomized trial procedure where a random part of the population is prevented from being targeted by advertising. It consists of 25M rows, each one representing a user with 12 features, a treatment indicator and 2 labels (visits and conversions).

For the experiments we firstly preprocess datasets, specifically we binarize categorical variables and normalize the features, for the classification we use Logistic Regression model from `Scikit-Learn` [80] Python library as it has fast learning and inference processes. Then we do each experiment in the following way: we do 50 stratified random train/test splits both for treatment and control groups with a ratio 70/30, during learning process we tune parameters of each model on a grid search. For DDR and SDR we use the regularization trick that we explained earlier, we tune additional regularization terms on a grid search as well.

As a performance measure we utilize Qini coefficient Q based on "joint, absolute" Qini curve [38] as metric corrects uplifts of selected individuals with respect to the number of individuals in treatment/control groups thanks to weighted term (see Section 2.1.2.1). To check statistical significance we use two-sample paired t-test at 5% confidence level (marked in bold in the tables when positive).

4.5.1 Choice of base classifier

Most uplift approaches use a base classifier. As one can see on Table 4.1, there are huge gaps in learning and inference time between different base classifier used in the same approach. So we can conclude that it make sense to use Generalized Linear Model (e.g. Logistic Regression) in the large-scale case.

	Learning time	Inference time
Logistic Regression	1x ± 0.06	1x ± 0.03
Random Forest (10 trees)	7.8x ± 0.27	35.7x ± 1.33
MultiLayer Perceptron (100 neurons)	319x ± 57.1	24.3x ± 0.72
SVM (Dual, RBF kernel)	2801x ± 101	4053x ± 141

Table 4.1: Relative learning and inference times of different base classifiers for Two-Models approach on Hillstrom dataset (single machine, 100 repetitions provide variance estimates).

4.5.2 Performance of Dependent Data Representation

We compare DDR with a Two-Models as first is an extention of the second, results are shown on Table 4.2. We use Hillstrom dataset with a "visit" outcome and cover three cases: firstly we compare approaches on a full dataset, then reduce control group randomly choosing 50% of it and for the last experiment we randomly choose 10% of control group to check how methods will perform with imbalanced data case. Indeed it is usually the case that the control group is kept to a minimum share so as not to hurt global treatment efficiency (e.g. ad revenue). As we can see, DDR significantly outperform Two-Models on imbalanced cases.

	BALANCED T/C	IMBALANCED T/C (50% OF C GROUP)	HIGHLY IMBALANCED T/C (10% OF C GROUP)
TWO-MODELS	0.06856	0.06292	0.03979
DDR	0.06866	**0.06444**	**0.04557**

Table 4.2: Performances of Two-Models and DDR approaches measured as mean Qini coefficient Q.

Different directions of DDR As DDR approach is based on a consecutive learning of two classifiers, there are two ways of learning − to fit first model on treatment group and then use output as a feature for the second one and fit it on a control part (we denote it as $T \to C$), or vice versa ($C \to T$).

Table 4.3 indicates that both approaches are comparable in the balanced case but $C \to T$ direction is preferable in other cases (at least with this dataset). Since the test set has more treated examples it makes sense that the stronger predictor obtained on this group by using information from predicted uplift on control performs best.

	BALANCED T/C	IMBALANCED T/C (50% OF C GROUP)	HIGHLY IMBALANCED T/C (10% OF C GROUP)
DDR ($T \to C$)	0.06895	0.06394	0.03979
DDR ($C \to T$)	0.06866	0.06444	0.04557

Table 4.3: Comparison of directions of learning in DDR approach (Qini coefficient Q).

Complexity of treatment effect with DDR To investigate complexity of the link between treatment and control group we use a dummy classifier (predicting the average within-group response) successively for one of treatment or control group while still using the regular model for the remaining group. Intuitively if the treatment effect is a constant, additive uplift then a simple re-calibration using a dummy model should be good enough. Conversely if there is a rich interaction between feature and treatment to explain outcome then a second, dummy classifier would perform poorly.

Table 4.4 indicates that the rich interaction hypothesis seems more plausible in this case, with maybe an even richer one in treated case.

	BALANCED T/C
DDR	0.06866
DDR (DUMMY FOR C GROUP)	0.04246
DDR (DUMMY FOR T GROUP)	0.01712

Table 4.4: Comparison between different variants of DDR approach.

4.5.3 Performance of Shared Data Representation

Here we compare SDR with Class Variable Transformation (CVT) approach because of similar nature of the uplift prediction. CVT model is learned with samples reweighting as in the original paper [57].

Table 4.5 indicates that SDR significantly outperforms CVT on imbalanced cases. Note that due to heavy down-sampling in the imbalanced cases it is not trivial to compare Qini coefficient Q values between columns.

	BALANCED T/C	IMBALANCED T/C (50% OF C GROUP)	HIGHLY IMBALANCED T/C (10% OF C GROUP)
CVT	0.06879	0.06450	0.05518
SDR	0.06967	**0.06945**	**0.08842**

Table 4.5: Performances of CVT and SDR approaches measured as mean Qini coefficient Q.

Usefullness of conjunction features In order to check usefulness of conjunctions features with SDR we compare it with a trivial variant in which we simply add an indicator variable for treatment instead of the whole feature set. This allows the model to learn only a simple re-calibration of the prediction for treated/control.

Table 4.6 indicates that it strongly degrades model performance.

	SDR (STANDARD)	SDR (T/C INDICATOR)
Q	**0.06967**	0.02706

Table 4.6: Comparison between variants of SDR in balanced treatment/control conditions.

Performance in imbalanced outcome condition We also compare SDR approach with CVT on CRITEO-UPLIFT1 dataset with conversion as outcome on a random sample of 50,000. Ratio between control and treatment group is 0.18 so it is highly imbalanced case as well but the outcome is also imbalanced with average level at only .00229.

Table 4.7 indicates that SDR again significantly outperforms CVT in this setting.

	CVT	SDR
Q	0.25680	**0.54228**

Table 4.7: Performances of CVT and SDR in highly imbalanced conditions for both treatment and outcome.

4.6 Summary

We proposed two new approaches for the uplift modeling problem based on dependent and shared data representations. Experiments show that they outperform current methods in imbalanced treatment conditions. In particular they allow to learn rich interaction between the features and treatment to explain response. Future research

would include learning more complex (highly non-linear) data representations permitting even richer interactions between features and treatment. Particular research is currently underway, although without meaningful results so far.

Chapter 5

AUUC Maximization with Generalization Guarantees

This chapter is based on published paper [18]: "Uplift Modeling with Generalization Guarantees" – Artem Betlei, Eustache Diemert, Massih-Reza Amini, published at KDD 2021.

5.1 Motivation

The problem with pointwise uplift prediction. Uplift modeling calls for a ranking objective in order to choose the top most responsive individuals as it is implemented in the AUUC metric. In the state-of-the-art, a large part of uplift modeling techniques resort to pointwise prediction, which consists in predicting accurate assessments of observations relevance by defining a pointwise learning objective, as a sum or average over individual samples in the dataset (overview in Section 2.2). However, two methods that perform equally at predicting scores may perform differently at predicting the ranking of samples.

This situation is also common in other tasks like classification where it has been shown that algorithms designed to minimize the error rate may not lead to the best possible Area Under the ROC Curve (AUC) as one may inadvertently degrade AUC whilst keeping a fixed error rate [32].

Moreover, the Empirical Risk Minimization (ERM) principle gives guarantees of generalization to unseen data *for the loss that is optimized*. Hence it cannot be summoned to obtain such guarantees if the pointwise loss and the metric of interest (i.e. AUUC) are not the same. Finally, the situation we describe happens in practice, as it can be observed in a simple experiment: when selecting model hyperparameters by loss one can have similar training losses that lead to very different AUUC (see Sec.

5.6). For these theoretical and empirical reasons we propose to learn an uplift model by optimizing a quantity that is a direct surrogate of the AUUC.

Importance of generalization bounds. Many studies in machine learning and data-mining now often incorporate generalization bounds in the design of learning algorithms [66]. These bounds are usually used for model selection or to analyze the model's generalization ability. Recent works in CATE prediction and uplift modeling fields propose to bound generalization error of Precision when estimating Heterogeneous Effect (PEHE) [94] and the deviation of a given pointwise estimator of the uplift with respect to a given loss function such as the least mean square error [104]. But as discussed above, these pointwise objective functions are not the most appropriate for AUUC.

5.2 Contributions

Considering the crucial role of treatment targeting in many applications, the need for models that optimize the metric of interest directly and the advances in the technical tools needed to study generalization properties of ranking models, we form the following research agenda: *i*) study generalization bounds for AUUC, *ii*) derive a learning objective and *iii*) experiment the corresponding empirical performance compared to traditional methods. Our main contributions in that respect are summarized as follows.

1. We propose the first generalization bound for AUUC using data-dependent concentration inequalities on dependent variables.

2. We present a ranking based algorithm, referred to as AUUC-max, directly maximizing a lower bound of the generalization error of AUUC, usable with different models, and that is efficient for hyperparameters tuning.

3. We report thorough performance evaluation against a range of competitive baselines on two real-world datasets.

5.3 Area Under the Uplift Curve

Formalization of AUUC. We chose the "separate, relative" uplift curve introduced in [38], their evaluations have concluded that this choice is robust to treatment

imbalance and captures well the intended usage of uplift models to target future treatments. We give a self-contained formula in Definition 1, corresponding to (Equations 10 and 16) of [38].

Definition 1 (Area Under the Uplift Curve)**.** *Let* $f(S_1, \frac{p}{100}n_1)$ *and* $f(S_0, \frac{p}{100}n_0)$ *be the first* p *percentages of* S_1 *and* S_0 *respectively when both ordered by prediction of model* f*. The empirical AUUC of the model* f *on* S_1 *and* S_0 *is given by:*

$$\widehat{AUUC}(f, S_1, S_0) = \int_0^1 V(f, x)dx \approx \sum_{p=1}^{100} V(f, \frac{p}{100}) \tag{5.1}$$

where

$$V(f, \frac{p}{100}) = \frac{1}{n_1} \sum_{i \in f(S_1, \frac{p}{100}n_1)} y_i - \frac{1}{n_0} \sum_{j \in f(S_0, \frac{p}{100}n_0)} y_j \tag{5.2}$$

5.4 On the Generalization Bound of AUUC and Learning Objective

In this section, we bound the difference between AUUC and its expectation and use this new bound to formulate a corresponding learning objective. For that purpose, we start by drawing a connection between AUUC and bipartite ranking risk (Section 5.4.1); and by means of Rademacher concentration inequalities build a generalization bound (Section 5.4.2). Then we define a principled optimization method with generalization guarantees for AUUC that leverages the bound as a robust learning objective (Section 5.4.3). Finally, we review related approaches and their merits as found in the literature.

5.4.1 Connection between AUUC and Bipartite Ranking Risk

From the connection between the Area under the ROC curve (AUC) and the bipartite ranking risk, we can show that AUUC is a weighted combination of ranking losses for the treatment and control responses. Formal version of the decomposition is provided in Proposition 2.

Proposition 2. *Let* $\widehat{AUUC}(f, S_1, S_0)$ *be the empirical area under uplift curve of the model* f *on the sets* S_1 *and* S_0*,* $\tilde{S}_0$ *is control set with reverted labels; and* $AUUC(f) = \mathbb{E}_{S_1,S_0}\left[\widehat{AUUC}(f, S_1, S_0)\right]$ *be its expectation. Then* $AUUC(f)$ *is related to ranking loss (Equation 5.4) as:*

$$AUUC(f) = \gamma - \left(\lambda_1 \mathbb{E}_{S_1}[\hat{R}(f, S_1)] + \lambda_0 \mathbb{E}_{\tilde{S}_0}[\hat{R}(f, \tilde{S}_0)]\right) \tag{5.3}$$

where

$$\hat{R}(f, S_t) \triangleq \frac{1}{n_t^+ n_t^-} \sum_{(\mathbf{x}_i,+1)\in S_t} \sum_{(\mathbf{x}_j,0)\in S_t} \mathbb{1}_{f(\mathbf{x}_i)\leq f(\mathbf{x}_j)} \tag{5.4}$$

is the empirical bipartite ranking risk, $t \in \{1, 0\}$*,* n_t^+, n_t^- *are the amounts of positives and negatives respectively in the set* S_t *(i.e.* $n_t = n_t^+ + n_t^-$*),* $\gamma = \mathbb{E}_{S_1,S_0}[\bar{y}_1 - \frac{(\bar{y}_1)^2}{2} - \frac{(\bar{y}_0)^2}{2}]$ *and* $\bar{y}_1, \bar{y}_0$ *are average treatment and control groups' outcomes respectively.*

Proof. From Definition 1:

$$\widehat{AUUC}(f, S_1, S_0) = \int_0^1 V(f, x)dx$$

[99, Equation 13] allows us to express $V(f, x)$ as a difference of *cumulative outcome rates* $F_f^{S_1}(x)$ and $F_f^{S_0}(x)$ (for the formal definition please refer to [99]) of collections S_1 and S_0 respectively, induced by model f:

$$V(f, x) = F_f^{S_1}(x) - F_f^{S_0}(x)$$

Hence,

$$\begin{aligned}\widehat{AUUC}(f, S_1, S_0) &= \int_0^1 V(f, x)dx = \int_0^1 \Big(F_f^{S_1}(x) - F_f^{S_0}(x)\Big)dx \\ &= \int_0^1 F_f^{S_1}(x)dx - \int_0^1 F_f^{S_0}(x)dx \end{aligned} \tag{5.5}$$

By the mean while, we have from [99, Equation 9] a connection between $F_f^{\mathcal{D}}(x)$ and Gini coefficient $Gini(f, \mathcal{D})$ – popular metric in binary classification indicated the ability of the model to discriminate between positive and negative classes and used frequently in credit scoring and direct marketing fields. So over the dataset $\mathcal{D}$ connection is:

$$Gini(f, \mathcal{D}) = \frac{2\int_0^1 F_f^{\mathcal{D}}(x)dx - \bar{y}_{\mathcal{D}}}{\bar{y}_{\mathcal{D}}(1 - \bar{y}_{\mathcal{D}})} \tag{5.6}$$

where $\bar{y}_{\mathcal{D}}$ is average outcome rate on $\mathcal{D}$. Note that the Gini coefficient is also related to the area under ROC curve as follows [100]:

$$Gini(f, \mathcal{D}) = 2AUC(f, \mathcal{D}) - 1 \tag{5.7}$$

From (5.6) and (5.7), it then comes :

$$\int_0^1 F_f^{\mathcal{D}}(x)dx = \bar{y}_{\mathcal{D}}(1 - \bar{y}_{\mathcal{D}}) \cdot AUC(f, \mathcal{D}) + \frac{(\bar{y}_{\mathcal{D}})^2}{2} \tag{5.8}$$

From (5.5) and (5.8) it comes :

$$\widehat{AUUC}(f, S_1, S_0) = \bar{y}_1(1-\bar{y}_1) \cdot AUC(f, S_1) \\ - \bar{y}_0(1-\bar{y}_0) \cdot AUC(f, S_0) + \frac{(\bar{y}_1)^2}{2} - \frac{(\bar{y}_0)^2}{2}$$

Now by reverting labels in S_0; i.e. $AUC(f, S_0) = (1 - AUC(f, \tilde{S}_0))$ we get

$$\begin{aligned}\widehat{AUUC}(f, S_1, S_0) &= \bar{y}_1(1-\bar{y}_1)AUC(f, S_1) \\ &\quad + \bar{y}_0(1-\bar{y}_0)\Big(1 - AUC(f, \tilde{S}_0)\Big) + \frac{(\bar{y}_1)^2}{2} - \frac{(\bar{y}_0)^2}{2} \\ &= \bar{y}_1(1-\bar{y}_1) \cdot AUC(f, S_1) \\ &\quad + \bar{y}_0(1-\bar{y}_0) \cdot AUC(f, \tilde{S}_0) + \frac{(\bar{y}_1)^2}{2} + \frac{(\bar{y}_0)^2}{2} - \bar{y}_0\end{aligned}$$

Using the connection between AUC and the empirical ranking loss $AUC(f, \mathcal{D}) = 1 - \hat{R}(f, \mathcal{D})$, we have :

$$\begin{aligned}\widehat{AUUC}(f, S_1, S_0) &= \bar{y}_1(1-\bar{y}_1) \cdot \Big(1 - \hat{R}(f, S_1)\Big) \\ &\quad + \bar{y}_0(1-\bar{y}_0) \cdot \Big(1 - \hat{R}(f, \tilde{S}_0)\Big) + \frac{(\bar{y}_1)^2}{2} + \frac{(\bar{y}_0)^2}{2} - \bar{y}_0 \\ &= \hat{\gamma}_{S_1,S_0} - \Big(\lambda_1 \hat{R}(f, S_1) + \lambda_0 \hat{R}(f, \tilde{S}_0)\Big)\end{aligned}$$

where, for sake of notation, we use group indices 1 and group 0 instead of datasets S_1 and S_0 in the upper indices of $\bar{y}$; and $\lambda_1 = \bar{y}_1(1-\bar{y}_1), \lambda_0 = \bar{y}_0(1-\bar{y}_0), \hat{\gamma}_{S_1,S_0} = \bar{y}_1 - \frac{(\bar{y}_1)^2}{2} - \frac{(\bar{y}_0)^2}{2}$.

By taking the expectations in both sides of equation we finally get :

$$\begin{aligned}AUUC(f) &= \mathbb{E}_{S_1,S_0}\Big[\widehat{AUUC}(f, S_1, S_0)\Big] \\ &= \gamma - \Big(\lambda_1 \mathbb{E}_{S_1}[\hat{R}(f, S_1)] + \lambda^0 \mathbb{E}_{\tilde{S}_0}[\hat{R}(f, \tilde{S}_0)]\Big)\end{aligned}$$

where, $\gamma = \mathbb{E}_{S_1,S_0}[\hat{\gamma}_{S_1,S_0}]$. □

5.4.2 Rademacher Generalization Bounds

Let us now consider the minimization problems of the pairwise ranking losses over the treatment and the control subsets (Equation 5.4), and the following dyadic transformation defined over each of the groups S_1 and $\tilde{S}_0$:

$$\mathcal{T}(S_t) = \Big\{(\mathbf{z} = (\mathbf{x}, \mathbf{x}'), \tilde{y}) \,\Big|\, ((\mathbf{x}, y), (\mathbf{x}', y')) \in S_t \times S_t \wedge y \neq y'\Big\} \tag{5.9}$$

where, $t \in \{1, 0\}$, $\tilde{y} = +1$ iff $y = +1$ and $y' = 0$ and $\tilde{y} = -1$ otherwise. Here we suppose that $\mathcal{T}(S_t)$ contains just one of the two pairs that can be formed by two examples of different classes. This transformation corresponds then to the set of $n_t^+ n_t^-$ pairs of observations in S_t that are from different classes.

From this definition and the class of functions, $\mathcal{H}$, defined as:

$$\mathcal{H} = \{h : \mathbf{z} = (\phi(\mathbf{x}), \phi(\mathbf{x}')) \mapsto f(\phi(\mathbf{x})) - f(\phi(\mathbf{x}')), f \in \mathcal{F}\}, \tag{5.10}$$

where, $\phi(\mathbf{x})$ is the feature representation associated to observation $\mathbf{x}$. The empirical loss (Equation 5.4) can then be rewritten as:

$$\hat{R}(h, \mathcal{T}(S_t)) = \frac{1}{n_t^+ n_t^-} \sum_{(\mathbf{z}, \tilde{y}) \in \mathcal{T}(S_t)} \mathbb{1}_{\tilde{y}h(\mathbf{z}) \leq 0}. \tag{5.11}$$

The loss defined in (Equation 5.11) is equivalent to a binary classification error over the pairs of examples in $\mathcal{T}(S_t)$. With this equivalence, one may expect to use efficient generalization bounds developed in binary classification. However, (Equation 5.11) is a sum over random dependent variables; as each training examples in S_t may be present in different pairs of examples in $\mathcal{T}(S_t)$, and the study of the consistency of the Empirical Risk Minimization principle cannot be carried out using classical tools; as the central i.i.d. assumption on which these tools are built on is transgressed. For this study, we consider $\mathcal{T}(S_t)$ as a dependency graph of random variables on its nodes, and similar to [102], we decompose it using the *exact proper fractional cover* of the graph proposed by [56] and defined as:

Definition 2. *Let $\mathcal{G} = (V, E)$ be a graph. $\mathcal{C} = \{(\mathcal{C}_j, \omega_j)\}_{j \in [J]}$, for some positive integer J, with $\mathcal{C}_j \subseteq V$ and $\omega_j \in [0; 1]$ is an exact proper fractional cover of $\mathcal{G}$, if:*

1. *it is* proper: *$\forall j$, $\mathcal{C}_j$ is an* independent set*, i.e., there is no connections between vertices in $\mathcal{C}_j$;*

2. *it is an* exact fractional cover *of $\mathcal{G}$: $\forall v \in V$, $\sum_{j : v \in \mathcal{C}_j} \omega_j = 1$.*

The weight $W(\mathcal{C})$ of $\mathcal{C}$ is given by: $W(\mathcal{C}) = \sum_{j \in [J]} \omega_j$ and the minimum weight $\chi^*(\mathcal{G}) = \min_{\mathcal{C} \in \mathcal{K}(\mathcal{G})} W(\mathcal{C})$ over the set $\mathcal{K}(\mathcal{G})$ of all exact proper fractional covers of $\mathcal{G}$ is the fractional chromatic number *of $\mathcal{G}$.*

Here, the weight $W(\mathcal{C})$ of $\mathcal{C}$ is given by $W(\mathcal{C}) = \sum_{k=1}^{J} \omega_k$ and the minimum weight, called the fractional chromatic number, and defined as $\chi^*(\mathcal{G}) = \min_{\mathcal{C} \in \mathcal{K}(\mathcal{G})} W(\mathcal{C})$ corresponds to the smallest number of subsets containing independent variables. A

trivial property that we rely on here is that for a dependency graph induced by a bipartite ranking problem we always have that $\chi^*(\mathcal{G})$ is equal to the minimal chromatic number which in turn is simply the cardinality of the largest class: $max(n_+, n_-)$.

For the sake of clarity we show an example on Figure 5.1, where a set of example S_t is composed of 2 positive ($\mathbf{x}_1^+, \mathbf{x}_2^+$ with output $y = 1$) and 3 negative ($\mathbf{x}_1'^-, \mathbf{x}_2'^-, \mathbf{x}_3'^-$: $y = 0$) examples; the left part depicts all the possible pairs of examples over which the ranking loss is estimated; in the right, the corresponding set $\mathcal{T}(S_t)$ and the induced dependency graph $\mathcal{G}$ between pairs of examples (where edges denote statistical dependence between pairs in $\mathcal{T}(S_t)$; the minimal coloring of $\mathcal{G}$ that are covers containing each independent pairs is, in this case, equal to the fractional chromatic number $\chi^*(\mathcal{G})$.

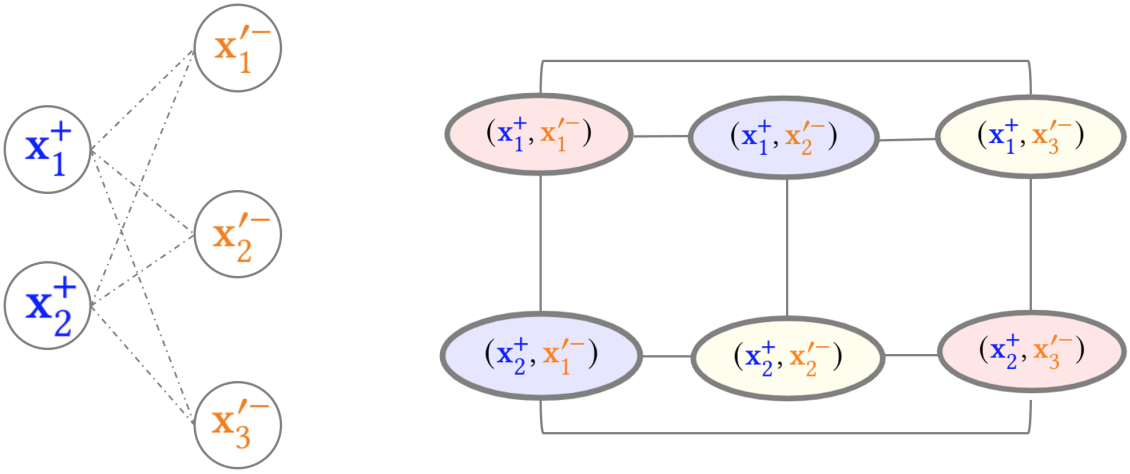


Figure 5.1: Dependency structure of a bipartite ranking problem composed of $n_t^+ = 2$ positive and $n_t^- = 3$ negative examples. (left: original data S_t and the composition of pairs shown in dashed; right: induced dependency graph $\mathcal{G}$; edges indicate dependencies between pairs in $\mathcal{T}(S_t)$, colors show covers that contain independent pairs, in this case we have $\chi^*(\mathcal{G}) = \max(n_t^+, n_t^-) = 3$).

From the definition of covers $\mathcal{C} = \{(\mathcal{C}_j, \omega_j\}_{j\in[J]}$ containing independent pairs, it is possible to adapt complexity terms, proposed to estimate the capacity of function classes in the i.i.d. setting, to the interdependent case [102]. The resulting capacity measure is defined as the weighted sum of complexity terms, each defined with respect to an element of $\mathcal{C}$. This capacity measure, denoted as fractional Rademacher complexity can be computed over the training set for a class of functions with bounded variance [85]; based on local Rademacher complexities [14] that have been found tight in practice. In this case, a strategy which consists in choosing a model with the best

generalization error tends to select functions with small variance in their predictions and a small bounded complexity that is computable on a training set.

Definition 3. ***The Local Fractional Rademacher Complexity***, $\mathfrak{R}_{S_t}(\mathcal{F}_r)$, *of the class of functions with bounded variance* $\mathcal{F}_r = \{f : X \mapsto \mathbb{R} : \mathbb{V}f \leq r\}$ *over the dyadic transformation,* $\mathcal{T}(S_t)$ *of size* $n_t^+ n_t^-$, *of the set* S_t, *is given by:*

$$\mathfrak{R}_{S_t}(\mathcal{F}_r) = \frac{1}{n_t^+ n_t^-} \mathbb{E}_\sigma \left[\sum_{j \in [J]} \omega_j \mathbb{E}_{X_{\mathcal{C}_j}} \left[\sup_{f \in \mathcal{F}_r} \sum_{i \in \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) \right] \right] \tag{5.12}$$

with $\sigma = (\sigma_1, \ldots, \sigma_{n_t^+ n_t^-})$ *being* $n_t^+ n_t^-$ *independent Rademacher variables verifying:* $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2; \forall i \in \{1, \ldots, n_t^+ n_t^-\}$.

From these statements, we can now present the first data-dependent generalization lower bound for AUUC.

Theorem 1. *Let* $S = \{\mathbf{x}_i, y_i\}_{i=1\ldots m} \in (\mathcal{X} \times \mathcal{Y})^m$ *be a dataset of* m *examples drawn i.i.d. according to a probability distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$, *and decomposable according to treatment* S_1 *and reverted label control* $\tilde{S}_0$ *subsets. Let* $\mathcal{T}(S_1)$ *and* $\mathcal{T}(\tilde{S}_0)$ *be the corresponding transformed sets. Then for any* $1 > \delta > 0$ *and* $0/1$ *loss* $\ell : \{-1, +1\} \times \mathbb{R} \to [0, 1]$, *with probability at least* $(1 - \delta)$ *the following lower bound holds for all* $f \in \mathcal{F}_r$*:*

$$\begin{aligned} AUUC(f) \geq \gamma &- \left(\lambda_1 \hat{R}_\ell(f, S_1) + \lambda_0 \hat{R}_\ell(f, \tilde{S}_0) \right) \\ &- \mathfrak{C}_\delta(\mathcal{F}_r, S_1, \tilde{S}_0) - \frac{25}{48} \left(\frac{\lambda_1}{n_1^+} + \frac{\lambda_0}{n_0^-} \right) \log \frac{2}{\delta} \end{aligned} \tag{5.13}$$

where,

$$\begin{aligned} \mathfrak{C}_\delta(\mathcal{F}_r, S_1, \tilde{S}_0) &= \lambda_1 \mathfrak{R}_{S_1}(\mathcal{F}_r) + \lambda_0 \mathfrak{R}_{\tilde{S}_0}(\mathcal{F}_r) \\ &+ \left(\frac{\frac{5}{2}\sqrt{\mathfrak{R}_{S_1}(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_1^+}} \lambda_1 + \frac{\frac{5}{2}\sqrt{\mathfrak{R}_{\tilde{S}_0}(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_0^-}} \lambda_0 \right) \sqrt{\log \frac{2}{\delta}} \end{aligned}$$

is defined with respect to local fractional Rademacher complexities of the class of functions $\mathcal{F}_r$ *estimated over the treatment and the control sets.*

Proof. From Proposition 2:

$$AUUC(f) = \gamma - \left(\lambda_1 \mathbb{E}_{S_1}[\hat{R}(f, S_1)] + \lambda_0 \mathbb{E}_{\tilde{S}_0}[\hat{R}(f, \tilde{S}_0)] \right)$$

From [85], we have the following upper bounds for each of the ranking losses hold with probability $1-\delta/2$:

$$\forall \mathcal{F}_r, \mathbb{E}_{S_1}[\hat{R}(f,S_1)] - \hat{R}(f,S_1) \leq \inf_{a_1>0}\left((1+a_1)\mathfrak{R}_{S_1}(\mathcal{F}_r) + \frac{5}{4}\sqrt{\frac{2r\log\frac{2}{\delta}}{n_1^+}} + \frac{25}{16}\left(\frac{1}{3}+\frac{1}{a_1}\right)\frac{\log\frac{2}{\delta}}{n_1^+}\right)$$

$$\forall \mathcal{F}_r, \mathbb{E}_{\tilde{S}_0}[\hat{R}(f,\tilde{S}_0)] - \hat{R}(f,\tilde{S}_0) \leq \inf_{a_0>0}\left((1+a_0)\mathfrak{R}_{\tilde{S}_0}(\mathcal{F}_r) + \frac{5}{4}\sqrt{\frac{2r\log\frac{2}{\delta}}{n_0^-}} + \frac{25}{16}\left(\frac{1}{3}+\frac{1}{a_0}\right)\frac{\log\frac{2}{\delta}}{n_0^-}\right)$$

The infinimums of the upper-bounds are reached for respectively

$$a_1 = \frac{5}{4}\sqrt{\frac{\log\frac{2}{\delta}}{n_1^+\mathfrak{R}_{S_1}(\mathcal{F}_r)}}, \qquad a_0 = \frac{5}{4}\sqrt{\frac{\log\frac{2}{\delta}}{n_0^-\mathfrak{R}_{\tilde{S}_0}(\mathcal{F}_r)}}$$

By plugging back these values into the upper-bounds the result follows from the union bound.

□

Note that the convergence rate of the bound is governed by least represented class in both treatment and reverted control subsets. To the best of our knowledge this is the first data-dependent generalization bound proposed for AUUC.

5.4.3 AUUC-max Learning Objective

From Theorem 1, we can formulate an optimization problem for the expected value of AUUC as follows:

$$\underset{f\in\mathcal{F}_r}{\operatorname{argmax}}\, AUUC(f) \equiv \underset{\theta,r}{\operatorname{argmin}}\Big(\lambda_1\hat{R}(f_\theta,S_1) + \lambda_0\hat{R}(f_\theta,\tilde{S}_0) + \mathfrak{C}_\delta(\mathcal{F}_r,S_1,\tilde{S}_0)\Big) \tag{5.14}$$

where θ are parameters of the model.

There are two remarks that we can make at this point. First, both terms $\hat{R}(f_\theta,S_1)$ and $\hat{R}(f_\theta,\tilde{S}_0)$ in (5.14) are defined over the instantaneous ranking loss $\mathbb{1}_{\tilde{y}(f(\mathbf{x})-f(\mathbf{x}'))\leq 0}$ and in practice we need a differentiable surrogate over these losses so that the minimization problem can be solved using standard optimization techniques. Second, the local fractional Rademacher complexities $\mathfrak{R}_{S_1}(\mathcal{F}_r)$ and $\mathfrak{R}_{S_0}(\mathcal{F}_r)$ that appear in

$\mathfrak{C}_\delta(\mathcal{F}_r, S_1, \tilde{S}_0)$ should be estimated for some fixed class of functions $\mathcal{F}_r$ with a well suited value of r.

For the first point, we propose to use differentiable surrogates of the instantaneous ranking loss [105], such as

$$s_{log}(z) = \frac{\ln(1+e^{-z})}{\ln(2)} \quad \text{and} \quad s_{poly}(z) = (-(z-\mu))^p \mathbb{1}_{z<\mu}$$

Note that $s_{log}(z)$ upper-bounds the indicator function $\mathbb{1}_{z\leq 0}$. This is also the case for $s_{poly}(z)$ with $\mu = 1$ and $p = 3$.

For the second point, we propose to upper bound both local Rademacher complexities $\mathfrak{R}_{S_1}(\mathcal{F}_r)$ and $\mathfrak{R}_{S_0}(\mathcal{F}_r)$ following Proposition 3.

Proposition 3. *Let S_t be a sample of size n_t with n_t^+ samples with positive labels and such that $\forall \mathbf{x} \in S_t \ \|\phi(\mathbf{x})\| \leq R$. Let $\mathcal{F}_r = \{\phi(\mathbf{x}) \mapsto \mathbf{w}^\top \phi(\mathbf{x}) : \|\mathbf{w}\| \leq \Lambda; f \in \mathcal{F} : \mathbb{V}f \leq r\}$, be the class of linear functions with bounded variance and bounded norm over the weights. Then for any $1 > \delta > 0$, the empirical local fractional Rademacher complexity of $\mathcal{F}_r$ over the set of pairs $\mathcal{T}(S_t)$ of size $n_t^+ n_t^-$, can be bounded with probability at least $1 - \frac{\delta}{2}$ by:*

$$\mathfrak{R}_{S_t}(\mathcal{F}_r) \leq \sqrt{\frac{R^2\Lambda^2}{n_t^+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_t^+}} \tag{5.15}$$

Proof.

$$\begin{aligned}
\mathfrak{R}_{S_t}(\mathcal{F}_r) &= \frac{1}{n_t^+ n_t^-} \sum_{j\in[J]} \mathbb{E}_{X_{\mathcal{C}_j}} |\mathcal{C}_j| \left[\frac{1}{|\mathcal{C}_j|} \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}_r} \sum_{i\in\mathcal{C}_j} \sigma_i f(\mathbf{x}_i) \right] \right] \\
&= \frac{1}{n_t^+ n_t^-} \sum_{j\in[J]} |\mathcal{C}_j| \underbrace{\mathbb{E}_{X_{\mathcal{C}_j}} \left[\hat{\mathfrak{R}}_{\mathcal{C}_j}(\mathcal{F}_r) \right]}_{\mathfrak{R}_{\mathcal{C}_j}(\mathcal{F}_r)} \\
&\overset{\text{[74, Equation 3.14]}}{\leq} \frac{1}{n_t^+ n_t^-} \sum_{k=1}^{n_t^-} n_t^+ \left(\hat{\mathfrak{R}}_{\mathcal{C}_j}(\mathcal{F}_r) + \sqrt{\frac{\log\frac{2}{\delta}}{2n_t^+}} \right) \\
&\overset{\text{[74, Theorem 4.3]}}{\leq} \frac{1}{n_t^+ n_t^-} \sum_{k=1}^{n_t^-} n_t^+ \left(\sqrt{\frac{R^2\Lambda^2}{n_t^+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_t^+}} \right) = \sqrt{\frac{R^2\Lambda^2}{n_t^+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_t^+}}.
\end{aligned}$$

□

Finally, we apply Cauchy-Swartz and Popoviciu's inequalities to bound the variance of any function $f \in \mathcal{F}_r$, $\mathbb{V}f$, by $r = \Lambda^2 R^2$ (see Section 5.8). Noting that R

is a constant depending on the set of feature representations we can transform the optimization problem in (θ, r) in (Equation 5.14) to a problem in $(\mathbf{w}, \Lambda)$. Furthermore, the constraint on the weights Λ can be considered in practice as a max-norm regularizer [97] and taken as a hyperparameter of the model.

From these settings, and the definition of a given surrogate loss $s : \mathbb{R} \to \mathbb{R}_+$ over the instantaneous ranking loss, the version of the optimization problem (5.14) that we consider is given in (Equation 5.16). In the following, we refer to the derived algorithm as AUUC-max. At the high level we decompose the optimization problem in $(\mathbf{w}, \Lambda)$ of (Equation 5.16) by choosing a grid of values for Λ and make use of the generalization guarantees of the bound to select the best model $\mathbf{w}^*$, that corresponds to the maximum lower bound value. Note that AUUC-max is working with both linear and deep models, as we derive (Equation 16) using feature representations $\phi(\mathbf{x})$.

AUUC-max optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \hat{\mathcal{L}}_{\mathbf{w}}(S_1, \tilde{S}_0) = & \frac{1}{n_1^+ n_1^-} \sum_{\{\mathbf{x}_i,+1\}\in S_1} \sum_{\{\mathbf{x}_j,0\}\in S_1} s(\mathbf{w}^\top \phi(\mathbf{x}_i) - \mathbf{w}^\top \phi(\mathbf{x}_j)) \\ & + \frac{1}{n_0^+ n_0^-} \sum_{\{\mathbf{x}_k,+1\}\in \tilde{S}_0} \sum_{\{\mathbf{x}_l,0\}\in \tilde{S}_0} s(\mathbf{w}^\top \phi(\mathbf{x}_k) - \mathbf{w}^\top \phi(\mathbf{x}_l)) + \mathfrak{C}_\delta(\mathcal{F}_{\Lambda^2 R^2}, S_1, \tilde{S}_0) \\ & \text{subject to} \quad \|\mathbf{w}\| \leq \Lambda \end{aligned} \tag{5.16}$$

Theoretically, a joint or alternate optimization over $(\mathbf{w}, \Lambda)$ is also possible. Interestingly, a small grid of Λs is sufficient in practice to obtain competitive performance (see Section 5.6).

Note that the usual practice for uplift models is to iterate over hyperparameters grids (e.g. for optimization and regularization) and select the best model by estimating the mean empirical AUUC over a k-fold cross-validation: this implies an inner "for" loop in place of our lower bound computation and consequently additional calculations.

5.5 Related work

In this section, we review some related works that address the problems of AUUC maximization and the generalization study of uplift and CATE.

SVM for Differential Prediction [64] proposes to maximize AUUC directly by expressing it as a weighted sum of two AUCs and maximizing it using a Support Vector Machine method (see Section 2.2.3). Our work bears similarity to their seminal work by borrowing the idea of decomposing AUUC into a weighted sum of AUCs. We further propose to optimize differentiable surrogates of the objective in the case of imbalanced treatment, and provide an algorithm allowing to maximize AUUC using linear or deep models with generalization guarantees as well as an efficient hyperparameter tuning procedure.

Promoted Cumulative Gain [38] draw a list-wise learning to rank formulation of AUUC and use the LambdaMART [25] algorithm to optimize it, alleviating the need for derivable surrogates at the price of more complex models.

Generalization bounds. The work of [94] provide a bound for the PEHE metric (so usable for simulation settings) and pioneered the use of generalization bounds for CATE. More closely to our work, [104] proposed a generalization bound for uplift prediction. However, the main differences with our approach is that the upper-bound of AUUC proposed in [104] is a MSE-like proxy that is applicable in the case where the variables Y and T are never observed together whereas we bound AUUC directly without such hypothesis. Further, the definition of the proxy objective proposed in [104] assumes that samples are i.i.d., whilst in our study the equivalence between the ranking objective (5.4) and the classification error over the pairs of examples (5.11) gives rise to the consideration of dependent samples that calls for specific concentration inequalities, namely *local fractional* Rademacher theory, that ensures fast convergence rates [14]. Finally, from an optimization side the approach developed in [104] leads to a mini-max optimization problem, that is avoided in AUUC-max by using the "revert label in control" trick.

5.6 Experimental evaluation

We conducted an number of experiments aimed at evaluating the merits of pairwise ranking and the proposed approach for AUUC maximization.

Experimental Setting. We use two open, real-life datasets. First one is Hillstrom (see Section 2.3.2), for which we used no-email vs women e-mail split as a binary

treatment and "visit" as an outcome as in [83]. Second one CRITEO-UPLIFTv2 (see Section 3), for which we picked "visit" as an outcome.

To compare algorithms[1] each dataset was split into train (70%) and test (30%) sets. Then, 5-fold cross-validation was used on train set for hyperparameters tuning before retraining the best model on the whole train set. Hyperparameters grids for the all algorithms are of similar size and values can be found in Section 5.8, as well as the details about used prediction models. Finally, algorithms are compared by AUUC on test set, using an empirical Bernstein bound [72] to compute a 95% confidence test set bound on the expectation of AUUC. More details are provided in Section 5.8.

Evidence of generalization problem with pointwise objectives. We perform the following experiment to highlight the problem of AUUC generalization with learning models that optimize a pointwise objective. As baseline model, we consider *Class Variable Transformation* (CVT) [57] introduced in Section 2.2, which is also based on label reverting as our approach, but that optimizes a pointwise log-loss objective, on Hillstrom dataset. Experiments are conducted by varying the regularization parameter L2 of CVT and AUUC-max and computing the correlation, R, between the corresponding training loss and test loss (Figure 5.2 top) and between the training loss and AUUC on the test set (Figure 5.2 bottom). Results indicate that i) both algorithms generalize in terms of their internal objective (top row) ii) CVT training loss does not correlate with test AUUC and many points with a similar train loss give very different test performance (bottom left) iii) AUUC-max training loss is mildly correlated to test AUUC and shows better performance across different regularization parameters (bottom right).

Tightness of Local Fractional Rademacher bounds. We also examine the choice of local fractional Rademacher complexity in the generalization bound. For that purpose we compute the generalization error on the *Hillstrom* dataset for different variants of Theorem 1: using the local fractional Rademacher concentration inequality on bipartite ranking risk (our proposition, in blue on Figure 5.3) or [4, 101] (in orange) or [44] (in green). We observe that our bound makes an average error of 0.015, which is much tighter than the alternatives. This result illustrates the benefit of a variance based data-dependent analysis framework that we propose for AUUC.

[1] For research purpose we will release the code.

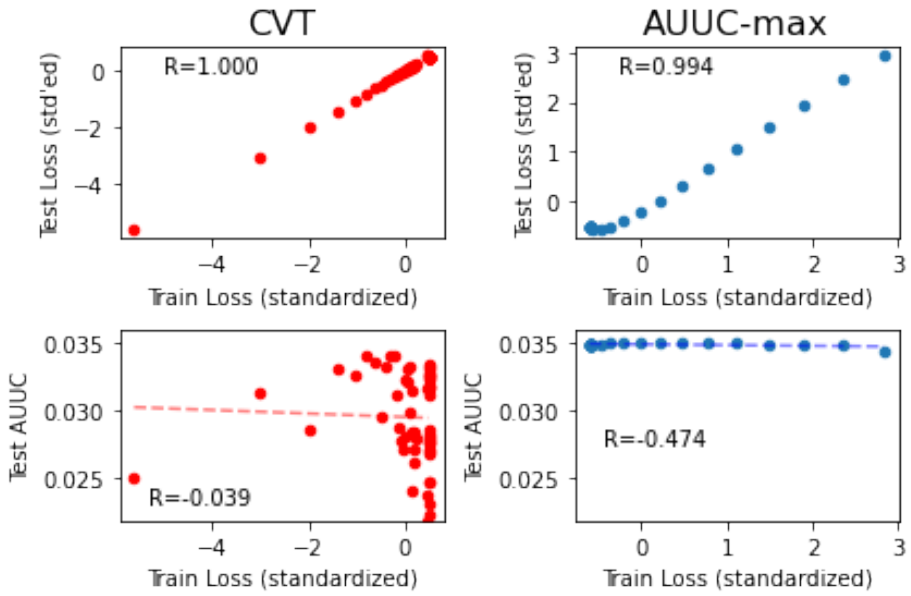


Figure 5.2: AUUC generalization problem with a pointwise objective on the Hillstrom dataset. CVT optimizing pointwise log-loss objective (left), AUUC-max optimizing (Equation 5.16) (right). R is the correlation coefficient.

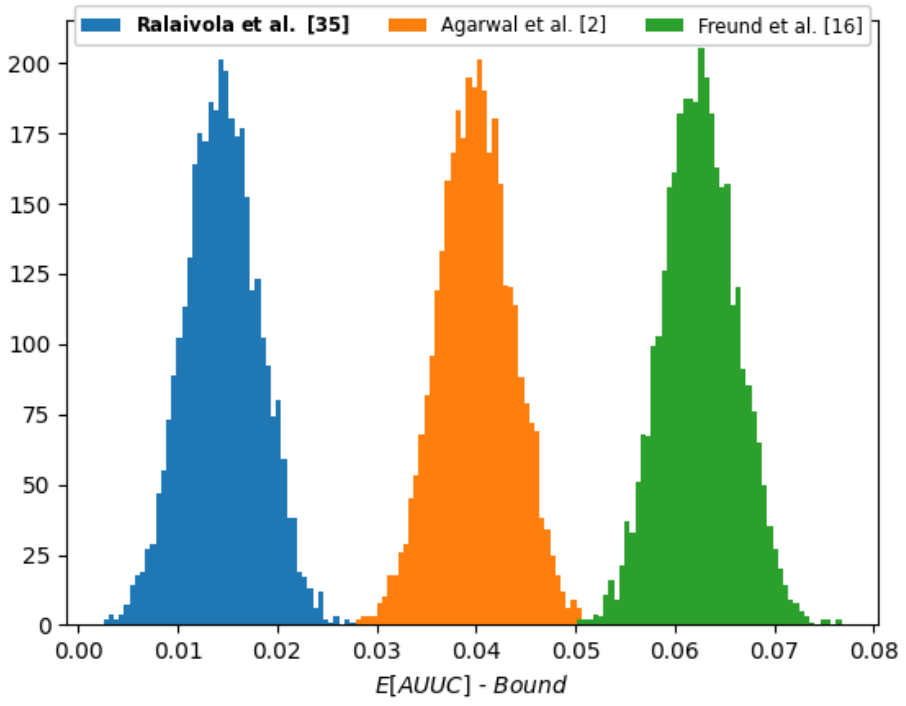


Figure 5.3: **AUUC bound tightness depending on inner bipartite ranking risk bounding technique** (closer to 0 is better).

Influence of the constraint on Λ. This term controls the bound tightness and also model regularization in the AUUC-max learning objective. A smaller Λ gives a tighter bound but also a more constrained model, up to some point where it is too constrained to be useful. In practice there is a region where both are near optimal as can be observed for the Hillstrom dataset on Figure 5.5 in Section 5.8.

Tuning parameters by bound is efficient. Following [9] we compare our method applied to linear model with hyperparameters chosen by bound (that is original AUCmax) versus chosen by cross-validation (+CV) in Table 5.8 in Section 5.8. Models tuned by either methods are practically equivalent (up to the 4th digit) whilst the bound method yields computation savings in $\mathcal{O}(k)$ where k is the number of folds. We observed similar behavior when using deep models.

AUUC-max is competitive in practice. Table 5.1 contains quantitative performance results of AUUC-max and a large selection of competitive baselines on *Hillstrom*. Firstly we remark that, in line with previous studies [39, 38, 64, 57], it is difficult to observe statistically significant results on this task. Nonetheless, small increases in AUUC can lead to important gains in the application [83]. We note that AUUC-max (deep, s_{log}) and AUUC-max (linear, s_{poly}) ranks 1st and 3rd respectively, indicating that our method is competitive both in performance and training time, which is in the last column of Table 5.1 (time is indicated relative to TM).

Additionally, Figure 5.4 presents uplift curves of the top ranked methods on the first 30% of population on *Hillstrom*. It is often the case in practice that we want to target only a small portion of the population for efficiency or budget constraints. One can see that bipartite ranking-based techniques (AUUC-max and SVM-DP) produce the highest cumulative uplifts on this threshold, which is an additional evidence of usefulness of bipartite ranking methods in uplift modeling. Figure of the full uplift curves for all methods are provided in Section 5.8.

For evaluation on the larger *CRITEO-UPLIFT v2* collection we select best performing methods on *Hillstrom* that can be trained reasonably fast. Results in Table 5.2 show very little variability and we find that no method performing significantly better than another, as on Hillstrom, though AUUC-max (deep, s_{log}) ranks 2nd.

Table 5.1: *Hillstrom*: comparison of baselines and AUUC-max. Top-2 results are in bold. †: original implementation of algorithm on LIBSVM was used.

Model	Train AUUC	Test AUUC	# params	Time
TM (Equation 2.16)	.03240	.02860 ± .00326	46	1.00x
CVT (Equation 2.25)	.03171	.02752 ± .00324	23	0.53x
SVM-DP [64]	.03273	.02957 ± .00321	23	0.02x †
DDR [17]	.03218	.02842 ± .00325	47	1.10x
SDR [17]	.03299	**.02958** ± .00327	67	2.44x
TARNet [94]	.03292	.02863 ± .00325	34,882	11.60x
GANITE [107]	.02563	.02900 ± .00326	7,045	1.12x
AUUC-max (linear, s_{poly})	.03239	.02912 ± .00326	23	0.37x
AUUC-max (deep, s_{log})	.03246	**.02999** ± .00325	15,469	1.34x

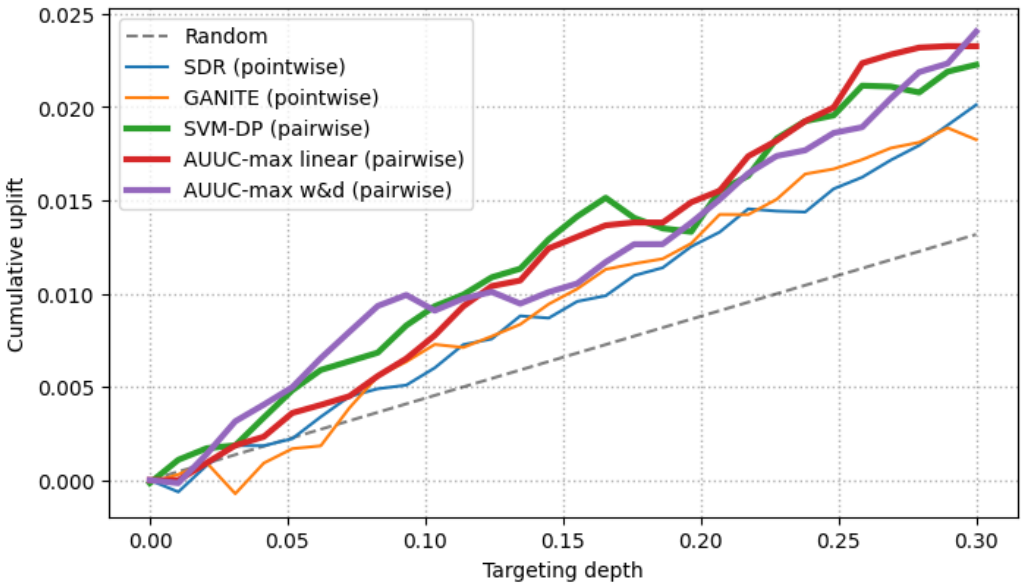


Figure 5.4: Uplift curves for the first 30% of population on Hillstrom. (higher is better)

Table 5.2: *CRITEO-UPLIFT v2*: comparison of baselines and AUUC-max. Top-2 results are in bold.

Model	Train AUUC	Test AUUC
TM (Equation 2.16)	.00925	.00922 ± .00001
SVM-DP [64]	.00928	**.00925** ± .00002
DDR [17]	.00925	.00920 ± .00001
SDR [17]	.00926	.00923 ± .00001
AUUC-max (linear, s_{poly})	.00925	.00921 ± .00001
AUUC-max (deep, s_{log})	.00927	**.00924** ± .00001

5.7 Summary

We propose the first, data-dependent generalization lower bound for the uplift modeling metric, AUUC, used in numerous practical cases. Then we derive a robust learning objective that optimizes a derivable surrogate of the AUUC lower bound. Our method alleviates the need of cross-validation for choosing regularization and optimization parameters, as we empirically show. As a result we highlight its simplicity and computational benefits. Experiments show that our method is competitive with the most relevant baselines from the literature, all methods being properly and fairly tuned. An exciting area for future works would be to compare Proposition 3 with the novel techniques of bounding $\mathfrak{R}_{S_t}(\mathcal{F}_r)$ for deep networks [13]. Another promising direction is about to adapt our bound to the other uplift models (e.g. SDR or TARNet). As a final word we expect that thanks to the availability of a powerful learning objective suited for deep models we could witness much progress in the field in the future, especially as researchers take advantage of recent advances in neural architecture search developed for other models and apply it to uplift modeling.

5.8 Additional details and experiments

Bounding variance of f

Let us remind function f from Proposition 3:

$$f(\phi(\mathbf{x})) = w^\top \phi(\mathbf{x}), \text{where } \|w\| \leq \Lambda, \|\phi(\mathbf{x})\| \leq R.$$

We need to proof that $\mathbb{V}(f) \leq r = \Lambda^2 R^2$.

Proof. Firstly we use Cauchy-Schwartz inequality for $f(\phi(\mathbf{x}))$:

$$|w^\top \phi(\mathbf{x})| \leq \|w\| \cdot \|\phi(\mathbf{x})\| \leq \Lambda R,$$

so now $-\Lambda R \leq w^\top \phi(\mathbf{x}) \leq \Lambda R$.

We apply then Popoviciu inequality on variances:

$$\mathbb{V}(f(\phi(\mathbf{x}))) = \mathbb{V}(w^\top \phi(\mathbf{x})) \leq \frac{(\Lambda R + \Lambda R)^2}{4} = (\Lambda R)^2 = r.$$

□

Experimental Setup details

Implementation details. Technically we implemented all surrogate losses and methods (except SVM-DP for which we used original code implemented on LIBSVM codebase) in Tensorflow framework [3]. For the optimization, Adam algorithm was used with step decay to update the learning rate.

Prediction models. For the TM, CVT, DDR and SDR methods we applied logistic regression as a prediction model. As was reported on TARNet paper, feed-forward neural network with fully-connected exponential-linear layers was used. For the deep model of AUUC-max we used feed-forward neural network with Wide & Deep architecture [30] which is focused on training linear model and deep neural network jointly in order to profit simultaneously from memorization and generalization.

Hyperparameters. For SVM-DP we found best parameter C on the range [1e-3,1e-2,1e-1,1e0,1e1,1e2,1e3]. For the other algorithms we applied random search through 50 and 30 parameters combinations for Hillstrom and CRITEO-UPLIFTv2 respectively, grids of the hyperparameters for the datasets are provided in Tables 5.3,5.4,5.5 and Tables 5.6,5.7 respectively.

Generalization problem with AUUC proxies experiment (Figure 5.2). The regularization parameter is L2 for both CVT and AUUC-max; values are 30 equally spaced points between $[0, 1]$. The dataset used is Hillstrom. We experienced similar behavior with other baselines such as TM.

Evaluation of the generalization bound (Figure 5.3). To assess the tightness of our bound, we depict the distribution of the differences between the true AUUC ($= \mathbb{E}[AUUC]$) and the lower bound computed on the Hillstrom dataset. For that purpose, we learn an AUUC-max model and record the train and test AUUCs. $\mathbb{E}[AUUC]$ is estimated from the upper bound of an Empirical Bernstein inequality [72] on the test sets obtained from 3,500 random train/test splits, giving a precision greater or equal than .001 with probability $> .99$. The distribution of the generalization error modeled by the bound is then simply the difference between train and test AUUCs.

Surrogates. For the surrogate s_{poly} for AUUC-max we used additional hyperparameters μ and p on the ranges of [0.1,0.3,0.5,0.7,1] and [2,3] respectively, according to the recomendations of [105]. We report the best performing surrogates in Tables 5.1 and 5.2.

Hardware information. All experiments were run on a Linux machine with 32 CPUs (Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz), with 2 threads per core, and 120Gb of RAM, with parallelising across 16 CPUs.

Table 5.3: Hyperparameters grid for TM, CVT, DDR and SDR on Hillstrom data

Parameter	TM & CVT & DDR & SDR
batch size	[128,512,1024]
learning rate	[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3,1e-2,5e-2,1e-1]
l_2 reg. term	[0,1e-6,1e-5,1e-4,1e-3,1e-2]

Test set bound

We derived test set bound on AUUC in order to get tight confidence intervals using only one train/test split. As a building block we used the test set bound for U-statistic [81] which is based on empirical Bernstein bound [72], then we constructed

Table 5.4: Hyperparameters grids for TARNet and GANITE on Hillstrom data

Parameter	TARNet
batch size	[128,512,1024]
learning rate	[1e-5,5e-5,1e-4,5e-4,1e-3]
l_2 reg. term	[0,1e-6,1e-5,1e-4,1e-3,1e-2]
# layers	[2, 3, 4]
# neurons	[32, 64, 128]

Parameter	GANITE
batch size	[128,512,1024]
learning rate	[1e-5, 1e-4, 1e-3]
# epochs	[50, 100, 500]
α	[1, 10, 100, 1000]
h_dim	[50, 100, 500]

Table 5.5: Hyperparameters grids for AUUC-max on Hillstrom data

Parameter	AUUC-max (linear)	AUUC-max (deep)
batch size	[256,512,1024]	[128,256,512,1024]
learning rate	[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3,1e-2]	[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3,1e-2]
Λ	[1e-2,5e-2,1e-1,5e-1,1e0,5e0,1e1,5e1,1e2]	[1e-2,5e-2,1e-1,5e-1,1e0,5e0,1e1,5e1,1e2]
l_2 reg. term	-	[0, 1e-5, 1e-3]
# layers	-	[2, 3, 4]
# neurons	-	[32, 64, 128]

Table 5.6: Hyperparameters grid for baselines on Criteo-UPLIFT v2 data

Parameter	TM & DDR & SDR
batch size	[128,512,1024]
learning rate	[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3,1e-2,5e-2,1e-1]
l_2 reg. term	[0,1e-6,1e-5,1e-4,1e-3,1e-2]

a union bound similarly to the our main result in Theorem 1. With probability at

Table 5.7: Hyperparameters grids for AUUC-max on Criteo-UPLIFT v2 data

Parameter	AUUC-max (linear)	AUUC-max (deep)
batch size	[512,1024,2048]	[512,1024,2048]
learning rate	[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3,1e-2]	[1e-5,5e-5,1e-4,5e-4,1e-3,5e-3,1e-2]
Λ	[1e-2,5e-2,1e-1,5e-1,1e0,5e0,1e1,5e1,1e2]	[1e-2,5e-2,1e-1,5e-1,1e0,5e0,1e1,5e1,1e2]
l_2 reg. term	-	[0, 1e-5, 1e-3]
# layers	-	[2, 3, 4]
# neurons	-	[32, 64, 128]

least $(1-\delta)$:

$$\begin{aligned} AUUC(f) \leq \widehat{AUUC}(f, S_{test_1}, S_{test_0}) \\ + \lambda_1 \left(\sqrt{\frac{4\hat{\Sigma}^2(S_{test_1}) \log \frac{8}{\delta}}{n_1}} + \frac{10}{n_1} \log \frac{8}{\delta} \right) \\ + \lambda_0 \left(\sqrt{\frac{4\hat{\Sigma}^2(S_{test_0}) \log \frac{8}{\delta}}{n_0}} + \frac{10}{n_0} \log \frac{8}{\delta} \right), \end{aligned} \tag{5.17}$$

where $\hat{\Sigma}^2(S_{test_1})$ is empirical variance of ranking loss for the treatment subset of test set, similarly for the control subset.

Effectiveness of AUUC-max for hyperparameters tuning

Table 5.8: *Hillstrom*: comparison of different parameter tuning techniques for AUUC-max. Training time is indicated relative to the AUUC-max (linear, s_{log}) + CV

Model	Train AUUC	Test AUUC	Time
AUUC-max (linear, s_{log})	.03230	.02878 ± .00325	0.27x
AUUC-max (linear, s_{log}) + CV	.03235	.02918 ± .00326	1.00x
AUUC-max (linear, s_{poly})	.03239	.02912 ± .00326	0.22x
AUUC-max (linear, s_{poly}) + CV	.03240	.02934 ± .00326	0.94x

Comparison of AUUC-max with PCG

Table 5.9: *Hillstrom*: comparison of AUUC-max with PCG. Result of PCG is taken from [38], Table 11.

Model	Test AUUC
PCG	.03055 ± N/A
AUUC-max (linear, s_{poly})	.02958 ± .00326
AUUC-max (deep, s_{log})	.03069 ± .00326

Influence of Λ

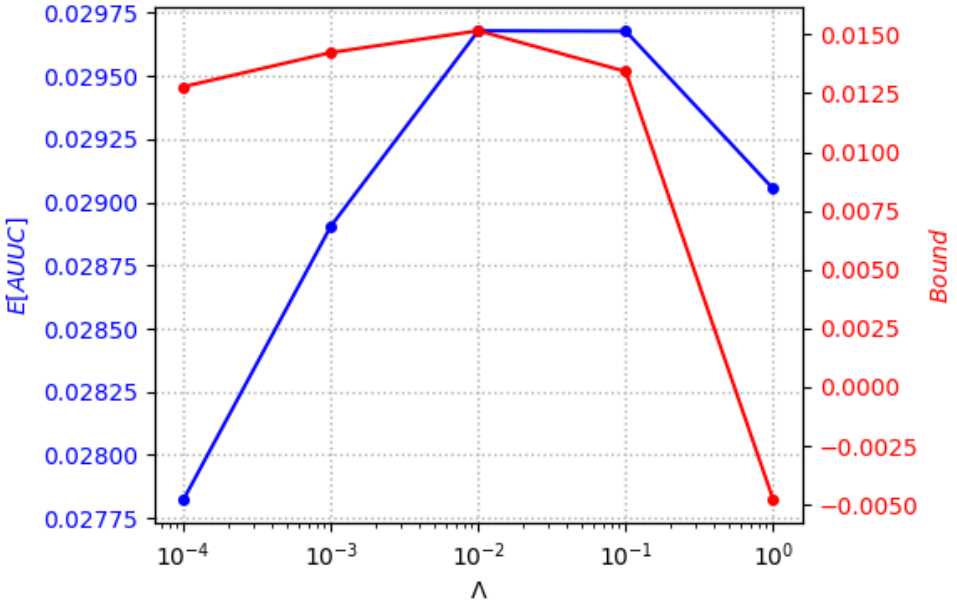


Figure 5.5: Influence of Λ on bound tightness and AUUC-max model performance.

Uplift curves on Hillstrom

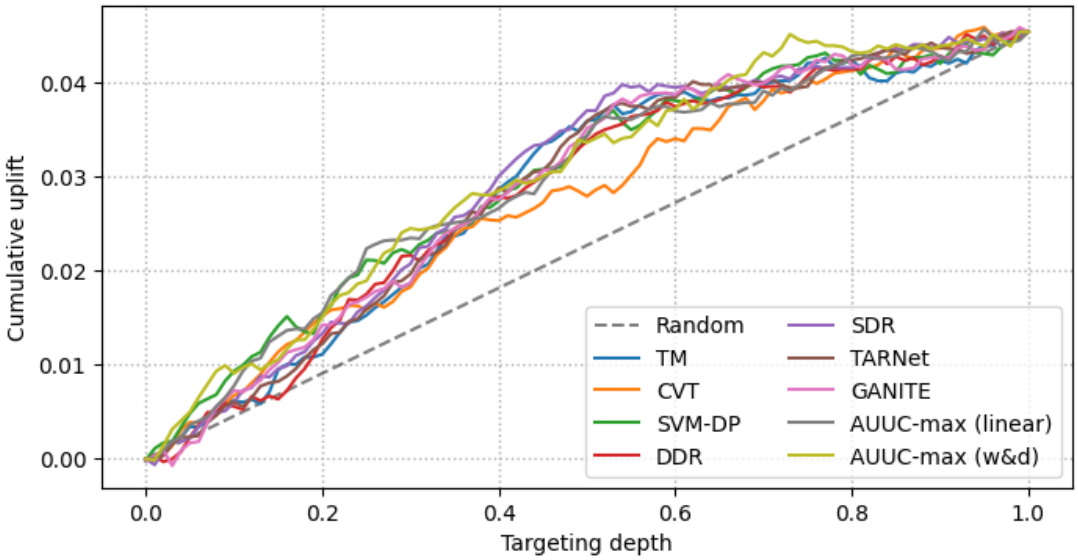


Figure 5.6: Uplift curves on Hillstrom. (higher is better)

Chapter 6

Differentially Private Uplift Modeling from Aggregated Data

This chapter is based on published paper [19]: "Differentially Private Individual Treatment Effect Estimation from Aggregated Data" – Artem Betlei, Théophane Gregoir, Thibaud Rahier, Aloïs Bissuel, Eustache Diemert, Massih-Reza Amini, accepted to Privacy Preserving Machine Learning, virtual ACM CCS 2021 workshop.

6.1 Motivation

Many of data-driven domains 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 [1] 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 uplift modeling and CATE prediction methods which fully leverage the trade-off between privacy and utility.

6.2 Related work

6.2.1 Learning from aggregated data

Learning individual-level behavior from aggregated-level data has long been known as *ecological inference* problem. Plenty of presented methods [60] use aggregated data, avoiding the problem of *ecological fallacy*, where the inferences drawn from aggregate level drastically differs from the ground truth at the individual level.

Besides, the most relevant level of aggregation has not yet been completely determined by research community as the term "aggregated data" has been referring to different frameworks: label similarities with complete access to features [111], aggregated labels with complete access to features [20] or aggregated labels with aggregated features [21]. Here, both features and labels will be considered as sensitive and therefore 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 in a query framework.

6.2.2 Differential Privacy

Differential privacy [42] 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 [27] 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 [2, 77].

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

6.3 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. For ϵ-ADUM, we identify and illustrate a bias-variance decomposition of the popular metric in CATE prediction, namely the Precision in Estimation of Heterogeneous Effects (PEHE) metric, 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.

6.4 Aggregated Data Uplift Model and its bias-variance trade-off

6.4.1 Preliminaries

6.4.1.1 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 $f^1, f^0 : \mathcal{K} \to \mathbb{R}$ and a real random variable ξ (independent of X) with $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] = \sigma^2$, such that

$$Y = Tf^1(X) \ + \ (1-T)f^0(X) \ + \xi. \tag{6.1}$$

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

$$u(\mathbf{x}) = f^1(\mathbf{x}) - f^0(\mathbf{x}). \tag{6.2}$$

Finally, we assume we have access to $\mathcal{D} = \{(\mathbf{x}_i, t_i, y_i)\}_{1 \leq i \leq n}$, 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.

6.4.1.2 Space partitioning

For a fixed positive integer p we define $\Pi_p(\mathcal{K}) := \left\{\pi \in \{1, \dots, p\}^{\mathcal{K}} \ : \ \pi \text{ surjective}\right\}$, 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 $G_\pi^{(1)}, \dots, G_\pi^{(p)}$ disjoint subsets of $\mathcal{K}$ such that $\bigcup_{1 \leq j \leq p} G_\pi^{(j)} = \mathcal{K}$. For a given $\mathbf{x} \in \mathcal{K}$, we denote $G_\pi(\mathbf{x}) = \pi^{-1}(\{\pi(\mathbf{x})\})$, the component of π which contains $\mathbf{x}$. For any $G \subset \mathcal{K}$ we denote $|G|^{\mathcal{D}} = \sum_{i \in \mathcal{D}} \mathbb{1}_{\mathbf{x}_i \in G}$, *i.e.* the number of points of $\mathcal{D}$ for which the feature vector $\mathbf{x}_i$ belongs to G.

6.4.2 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 $\mathbf{x} \in \mathcal{K}$ by the average treatment effect in the group $G_\pi(\mathbf{x})$. More formally, we define $\hat{u} : \mathcal{K} \to \mathbb{R}$ the function which to all $\mathbf{x} \in \mathcal{K}$ assigns:

$$\hat{u}_\pi(\mathbf{x}) = \hat{f}_\pi^1(\mathbf{x}) - \hat{f}_\pi^0(\mathbf{x}), \tag{6.3}$$

where $\hat{f}^1_\pi$ and $\hat{f}^0_\pi$ refer respectively to aggregated-data based models of the treatment and control response functions, *i.e.*:

$$\hat{f}^1_\pi(\mathbf{x}) = \frac{1}{|G_\pi(\mathbf{x})|^1} \sum_{i:\mathbf{x}_i \in G_\pi(\mathbf{x})} y_i t_i,$$
$$\hat{f}^0_\pi(\mathbf{x}) = \frac{1}{|G_\pi(\mathbf{x})|^0} \sum_{i:\mathbf{x}_i \in G_\pi(\mathbf{x})} y_i (1 - t_i).$$

$\hat{f}^1_\pi$ and $\hat{f}^0_\pi$ are piecewise constant functions defined using only aggregated information and would therefore be computable from an aggregate reporting API [61] thanks to `SUM` and `COUNT` queries.

We remind that the fundamental problem of causal inference (see Section 1.2.2) causes uplift modeling to be a very unique machine learning task, where the ground truth is unknown. By using aggregated data models for both the treatment and control positive outcome functions, we partially circumvent the fundamental problem of causal inference: as long as there are points from both the treatment and control groups in any given component $G_\pi(\mathbf{x})$ of π, the average treatment effect in $G_\pi(\mathbf{x})$ is consistently estimated by $\hat{u}(x)$.

Remark Besides, outside of the RCT setting, additionally assuming $\{\pi(X)\}$ is a valid adjustment set [78] for (T, Y) — e.g. in the case where $X \perp\!\!\!\perp T \mid \pi(X)$ 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 work.

6.4.2.1 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 $\mathbf{x} \in \mathcal{K}$ we define:

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

where $\mathbb{E}_\mathcal{D}[\hat{f}_\pi(\mathbf{x})]$ can be interpreted as the best possible $\pi-$piecewise constant approximation of f. In other words, $\mathbb{E}_\mathcal{D}[\hat{f}_\pi(\mathbf{x})]$ is the best approximation reachable by a $\pi-$ADM for f.

The (squared) bias term captures how well can f be approached by a piecewise constant function on π: the smaller the variations of f inside each of the subsets of

$\mathcal{K}$ defined by π, the lower the bias term. 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: The bigger the number of data points of $\mathcal{D}$ inside each subset of $\mathcal{K}$ defined by π, the better the approximation of aggregated target function by $\hat{f}_\pi$ (Law of Large Numbers). It should typically **increase when** $|\pi| = p$ **increases**.

Adapting the bias-variance decomposition of the mean squared error to the Precision in Estimation of Heterogeneous Effects (PEHE) metric (see Section 2.1) leads to the following

Proposition 4. *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:*

$$\begin{aligned}\epsilon_{PEHE}(\hat{u}_\pi) \leq 2\mathbb{E}_X \left[\text{Bias}^2\left(\hat{f}_\pi^{\mathcal{T}}(X)\right) + \text{Bias}^2\left(\hat{f}_\pi^{\mathcal{C}}(X)\right)\right] \\ + 2\mathbb{E}_X \left[\text{Var}\left(\hat{f}_\pi^{\mathcal{T}}(X)\right) + \text{Var}\left(\hat{f}_\pi^{\mathcal{C}}(X)\right)\right].\end{aligned}$$

Proof.

$$\begin{aligned}\epsilon_{PEHE}(\hat{u}_\pi) &= \mathbb{E}_{X,\mathcal{D}}\left[\left(u(X) - \hat{u}_\pi(X)\right)^2\right] \\ &= \mathbb{E}_{X,\mathcal{D}}\left[\left(\left(f^1(X) - f^0(X)\right) - \left(\hat{f}_\pi^1(X) - \hat{f}_\pi^0(X)\right)\right)^2\right] \\ &= \mathbb{E}_{X,\mathcal{D}}\left[\left(\left(f^1(X) - \hat{f}_\pi^1(X)\right) + \left(\hat{f}_\pi^0(X) - f^0(X)\right)\right)^2\right] \\ &\leq 2\mathbb{E}_{X,\mathcal{T}}\left[\left(f^1(X) - \hat{f}_\pi^1(X)\right)^2\right] + 2\mathbb{E}_{X,\mathcal{C}}\left[\left(f^0(X) - \hat{f}_\pi^0(X)\right)^2\right].\end{aligned}$$

We decompose now the inner part of the first term that corresponds to treatment population (second term, corresponding to control population, can be decomposed

analogically). For every $\mathbf{x}$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}\left[\left(f^1(\mathbf{x}) - \hat{f}^1_\pi(\mathbf{x})\right)^2\right] &= \mathbb{E}_{\mathcal{T}}\left[\left(f^1(\mathbf{x}) - \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})] + \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})] - \hat{f}^1(\mathbf{x})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{T}}\left[\underbrace{\left(f^1(\mathbf{x}) - \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})]\right)^2}_{\text{Independent of } \mathcal{T}}\right] + \mathbb{E}_{\mathcal{T}}\left[\left(\mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})] - \hat{f}^1(\mathbf{x})\right)^2\right] \\
&\quad + 2\mathbb{E}_{\mathcal{T}}\left[\underbrace{\left(f^1(\mathbf{x}) - \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})]\right)}_{\text{Independent of } \mathcal{T}} \cdot \left(\mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})] - \hat{f}^1(\mathbf{x})\right)\right] \\
&= \underbrace{\left(f^1(\mathbf{x}) - \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})]\right)^2}_{\text{Bias}^2\left(\hat{f}^{\mathcal{T}}_\pi(\mathbf{x})\right)} + \underbrace{\mathbb{E}_{\mathcal{T}}\left[\left(\mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})] - \hat{f}^1(\mathbf{x})\right)^2\right]}_{\text{Var}\left(\hat{f}^{\mathcal{T}}_\pi(\mathbf{x})\right)} \\
&\quad + 2\left(f^1(\mathbf{x}) - \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})]\right) \cdot \underbrace{\left(\mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})] - \mathbb{E}_{\mathcal{T}}[\hat{f}^1_\pi(\mathbf{x})]\right)}_{=\,0} \\
&= \text{Bias}^2\left(\hat{f}^{\mathcal{T}}_\pi(\mathbf{x})\right) + \text{Var}\left(\hat{f}^{\mathcal{T}}_\pi(\mathbf{x})\right).
\end{aligned}$$

Now, using expectancy over X for the derived term:

$$\mathbb{E}_{X,\mathcal{T}}\left[\left(f^1(X) - \hat{f}^1_\pi(X)\right)^2\right] = \mathbb{E}_X\left[\text{Bias}^2\left(\hat{f}^{\mathcal{T}}_\pi(X)\right)\right] + \mathbb{E}_X\left[\text{Var}\left(\hat{f}^{\mathcal{T}}_\pi(X)\right)\right].$$

Applying the same derivation for control population term and integrating the result in the first inequality, we finally get:

$$\begin{aligned}
\epsilon_{PEHE}(\hat{u}_\pi) \leq\ & 2\mathbb{E}_X\left[\text{Bias}^2\left(\hat{f}^{\mathcal{T}}_\pi(X)\right) + \text{Bias}^2\left(\hat{f}^{\mathcal{C}}_\pi(X)\right)\right] \\
& + 2\mathbb{E}_X\left[\text{Var}\left(\hat{f}^{\mathcal{T}}_\pi(X)\right) + \text{Var}\left(\hat{f}^{\mathcal{C}}_\pi(X)\right)\right].
\end{aligned}$$

□

6.4.3 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 [73]. 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($(\mathbf{x}_i, t_i, y_i)_{i\in[1,n]}, \pi \in \Pi_p(\mathcal{K}), D_y > 0, \epsilon > 0$):
  for $k \in [1, p]$ **do**
    for $t \in \{0, 1\}$ **do**
      $E_{k,t} = (y_i \mid \pi(\mathbf{x}_i) = k,\ t_i = t)$
      $C_{k,t} = \text{COUNT}(E_{k,t}) + \text{Lap}(\frac{2}{\epsilon})$ ▷ $\frac{\epsilon}{2}$-DP count
      $S_{k,t} = \text{SUM}(E_{k,t}) + \text{Lap}(2\frac{D_y}{\epsilon})$ ▷ $\frac{\epsilon}{2}$-DP sum
      $\widehat{y}_{k,t} = \frac{S_{k,t}}{C_{k,t}}$ ▷ ϵ-DP mean
    end for
    $\widehat{u}_k = \widehat{y}_{k,1} - \widehat{y}_{k,0}$ ▷ ϵ-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

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

6.4.4.1 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^1 and f^0 are respectively L_1 and L_0 Lipschitz on $\mathcal{K}$ that we suppose uni-dimensional ($d = 1$) of diameter D_x. Moreover, we denote $\beta_\pi = \max_{G,G'\in\pi} \{\frac{\text{diam}(G)}{\text{diam}(G')}\}$, and make the assumptions that every group $G \in \pi$ is equally populated with respect to $\mathcal{T}$ and $\mathcal{C}$, *i.e.* $\forall G \in \pi$, $|G|^{\mathcal{T}} = |G|^{\mathcal{C}}$.

6.4.4.2 PEHE bounding for $\epsilon-$ADUM

Corollary 1. *For a given $\Delta \in (0, \frac{1}{2}]$, 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 $\geq 1-\Delta$*:*

$$\begin{aligned}\epsilon_{PEHE}(\hat{u}_\pi) \leq\ & 2(L_0^2+L_1^2)D_x^2\beta_\pi^2\mathbf{p}^{-2} && \textit{ADUM Bias}\\ &+4\left(2\sigma^2+(L_0^2+L_1^2)D_x^2\right)\frac{\mathbf{p}}{\mathbf{n}} && \textit{ADUM Variance}\\ &+\kappa D_y^2\frac{\mathbf{p^2}}{\mathbf{n^2}\boldsymbol{\epsilon^2}} && \epsilon-\textit{DP term.}\end{aligned} \tag{6.4}$$

We provide the sketch of proof here (assuming general function $\hat{f}_\pi$ – we then imply the same reasoning for $\hat{f}_\pi^{\mathcal{T}}$ and $\hat{f}_\pi^{\mathcal{C}}$).

Proof. 1. Firstly we bound Bias and Variance terms from Proposition 4 respectively using the data constants such as D_x, L, β_π, n and number of groups p:

$$\mathbb{E}_X\left[\text{Bias}^2(\hat{f}_\pi(X))\right] \leq L^2D_x^2\beta_\pi^2p^{-2}. \tag{6.5}$$

$$\mathbb{E}_X\left[\text{Var}(\hat{f}_\pi(X))\right] \leq \left(\sigma^2+L^2D_x^2\right)\frac{p}{n}. \tag{6.6}$$

2. We apply the following auxiliary lemma: for $x\in[-\frac{1}{2},\frac{1}{2}]$:

$$1-x+\frac{2}{3}x^2 \leq \frac{1}{1+x} \leq 1-x+2x^2.$$

3. The new, ϵ-DP versions of Bias and Variance terms from Proposition 4 are defined:

$$\begin{aligned}\text{Bias}(\hat{f}_\pi^\epsilon(\mathbf{x})) &= f(\mathbf{x})-\mathbb{E}_{\mathcal{D},N_S,N_C}\left[\hat{f}_\pi^\epsilon(\mathbf{x})\right]\\ \text{Var}(\hat{f}_\pi^\epsilon(\mathbf{x})) &= \mathbb{E}_{\mathcal{D},N_S,N_C}\left[\left(\hat{f}_\pi^\epsilon(\mathbf{x})-\mathbb{E}_{\mathcal{D},N_S,N_C}\left[\hat{f}_\pi^\epsilon(\mathbf{x})\right]\right)^2\right]\end{aligned}$$

4. For the `COUNT` query in Algorithm 1, we use lemma from step 2 having $\frac{N_C^\epsilon}{C_k}\in[-\frac{1}{2},\frac{1}{2}]$:

$$\frac{1}{C_k}-\frac{N_C^\epsilon}{C_k^2}+\frac{2(N_C^\epsilon)^2}{3C_k^3} \leq \frac{1}{C_k}\cdot\frac{1}{1+\frac{N_C^\epsilon}{C_k}} \leq \frac{1}{C_k}-\frac{N_C^\epsilon}{C_k^2}+2\frac{(N_C^\epsilon)^2}{C_k^3}$$

where N_C^ϵ denotes Laplacian noise of parameter $\frac{2}{\epsilon}$.

5. Using step 5, both $\text{Bias}^2(\hat{f}_\pi^\epsilon(\mathbf{x}))$ and $\text{Var}(\hat{f}_\pi^\epsilon(\mathbf{x}))$ are bounded:

$$\begin{aligned}\text{Bias}^2(\hat{f}_\pi^\epsilon(\mathbf{x})) &\leq \text{Bias}^2(\hat{f}_\pi(\mathbf{x}))+\kappa_{bias}D_y^2\frac{p^2}{n^2\epsilon^2}\\ \text{Var}(\hat{f}_\pi^\epsilon(\mathbf{x})) &\leq \text{Var}(\hat{f}_\pi(\mathbf{x}))+\kappa_{var}D_y^2\frac{p^2}{n^2\epsilon^2}\end{aligned}$$

6. Finally, ϵ-DP version of Proposition 4 is created (assuming $\hat{u}_\pi$ as ϵ-ADUM):

$$\begin{aligned}\epsilon_{PEHE}(\hat{u}_\pi) &\leq 2\mathbb{E}_X\left[\text{Bias}^2\left(\hat{f}_\pi^{\epsilon\mathcal{T}}(X)\right) + \text{Bias}^2\left(\hat{f}_\pi^{\epsilon\mathcal{C}}(X)\right)\right] \\ &+ 2\mathbb{E}_X\left[\text{Var}\left(\hat{f}_\pi^{\epsilon\mathcal{T}}(X)\right) + \text{Var}\left(\hat{f}_\pi^{\epsilon\mathcal{C}}(X)\right)\right] \\ &\leq 2(L_0^2 + L_1^2)D_x^2\beta_\pi^2 p^{-2} \\ &+ 4\left(2\sigma^2 + (L_0^2 + L_1^2)D_x^2\right)\frac{p}{n} \\ &+ \kappa D_y^2 \frac{p^2}{n^2\epsilon^2}.\end{aligned}$$

□

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 [55], 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 (6.4) has the following asymptotic variations with respect to ϵ and n:

- when ϵ is small compared to $\sqrt{p/n}$, (6.4) is dominated by its first and last terms and $p^{opt} = \Theta(n\epsilon)$,
- when ϵ is large compared to $\sqrt{p/n}$, (6.4) 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.

6.5 Experimental evaluation

6.5.1 Synthetic data

6.5.1.1 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.

6.5.1.2 Performance comparison

Here, ϵ-ADUM is compared with a *Two-Models* (TM) [51] uplift modeling method, formed by two ϵ-differentially private linear regressions [109] 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 6.1, 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 [1]). 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 $\epsilon \geq 2$.

6.5.1.3 Bias-variance trade-off illustration

The bias-variance trade-off introduced in Corollary 1 is illustrated experimentally in Figure 6.2. 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 (6.4)) before increasing due to a penalizing variance (second term of (6.4)) and the rising impact of the privacy-induced noise addition (third term in (6.4)) – 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 6.4.4).

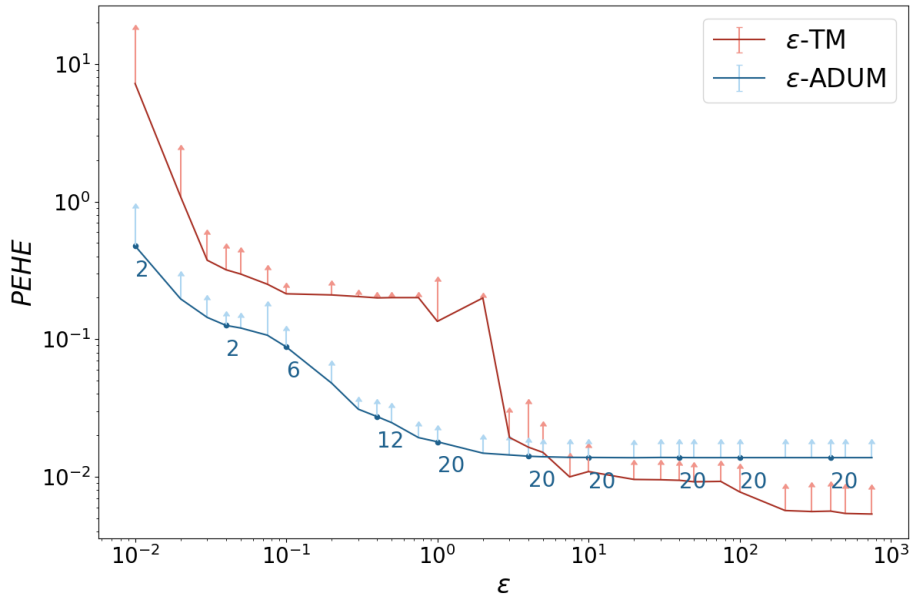


Figure 6.1: Comparison of test PEHE (lower is better) for ϵ-TM and ϵ-ADUM over 20 random train/test splits selecting 20000 points. Arrows represent standard deviations and the tuned number of groups for ϵ-ADUM is annotated in blue. For this experiment, $\sigma = 1$.

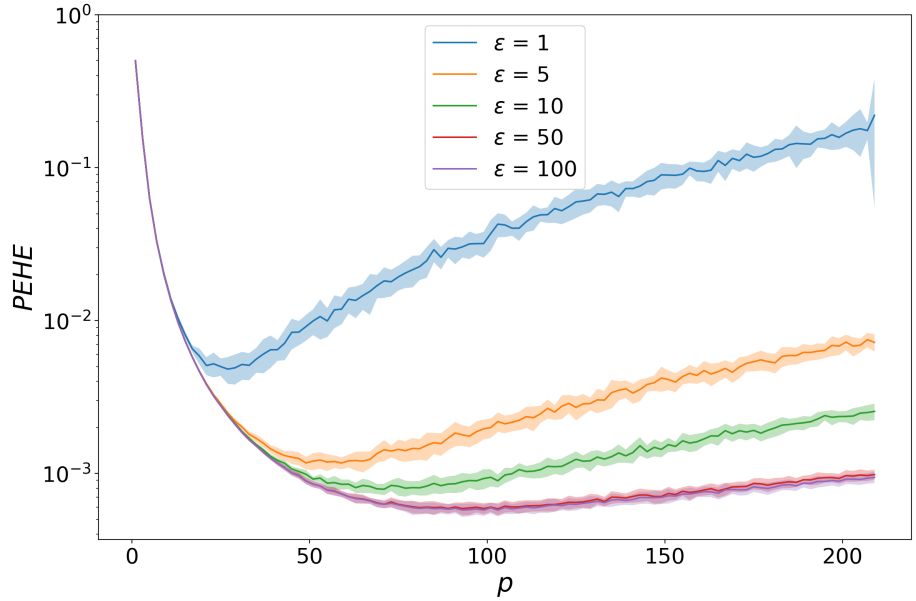


Figure 6.2: Test PEHE (lower is better) over 20 random train/test splits selecting 20000 points, illustrating the ϵ-ADUM bias-variance trade-off with respect to the number of groups p for 5 selected ϵ. For this experiment, $\sigma = 0.1$.

6.5.2 Real data

We perform real data experiments based on CRITEO-UPLIFTv2 dataset (see Section 3). Results are reported for the "visit" binary outcome, hence ϵ-differentially private logistic regressions [27] are used as prediction models in an ϵ-TM method. Besides, Area Under the Uplift Curve (AUUC) (see Section 2.1.2.1) built on "separate, relative" uplift curve [38] is applied as validation metric.

As presented in Section 6.4.2, 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 [98] $-$ 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 6.3).

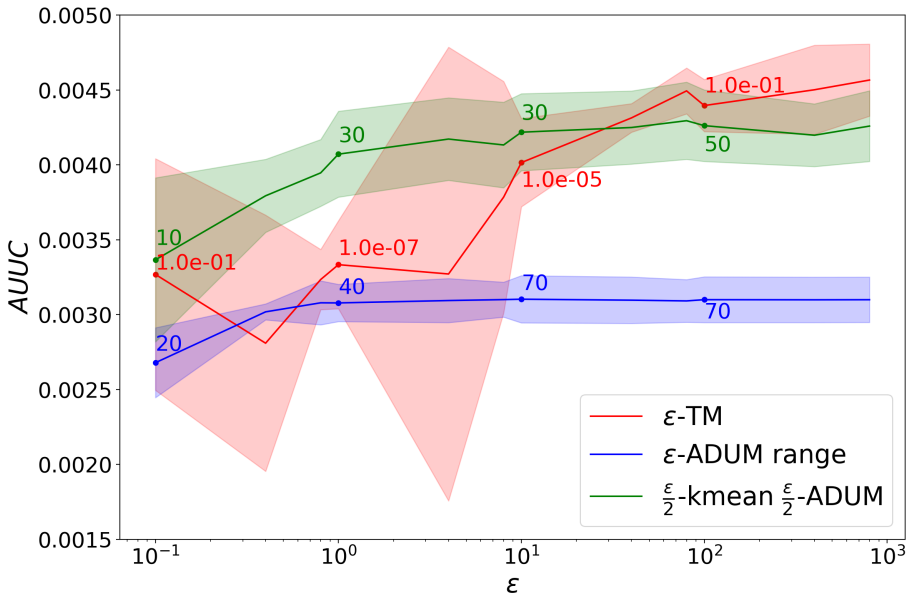


Figure 6.3: Comparison of test AUUC (higher is better) between individually-trained ϵ-TM and two variations of ϵ-ADUM over 4 random train/test splits randomly selecting 1M points from CRITEO-UPLIFTv2. The tuned number of groups for ϵ-ADUM is annotated in blue and green while the tuned regularization parameter C is in red for ϵ-TM.

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.

6.6 Summary

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 guided by the number of aggregation groups. 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. To go further, supplementary experiments highlighting the impact of partition design on ϵ-ADUM performances could help data providers build the most relevant partitions for uplift modeling.

Chapter 7

Conclusions and Future Perspectives

This dissertation has studied closely related problems of uplift modeling and CATE prediction, which are both machine learning-based techniques for treatment effect prediction at the individual level, which have become one of the main trends in a variety of application areas where personalization is key. Our research has mainly focused on specific problems encountered when working with uplift modeling in the field of online advertising, such as the large scale of data, imbalance treatment conditions, generalization, and privacy preservation.

Summary of Contributions

In **Chapter 3**, we released a publicly available large-scale dataset collected from several randomized control trials, scaling up previously available datasets by order of 2. We provided details on the data collection and performed series of sanity checks to validate the use of this data for tasks of interest. We also formalized how uplift modeling can be performed with this data, along with the relevant evaluation metrics. Then, we proposed synthetic response surfaces and heterogeneous treatment assignment providing a general set-up for CATE prediction and reported experiments to validate key characteristics of the dataset leveraging its size to evaluate and compare − with high statistical significance − a selection of baseline uplift modeling and CATE prediction methods.

In **Chapter 4**, we assumed an imbalanced treatment assignment scenario and formulated two new model-agnostic, data representation-based methods inspired by cascade and multi-task learning paradigms, applying the common idea of sharing the

knowledge between treatment and control populations. Experimental results over the several large-scale collections showed the benefits of the proposed approaches.

Then, in **Chapter 5** we covered the problem of direct optimization of the Area Under the Uplift Curve (AUUC), a popular metric in the uplift modeling. Utilizing the connection between uplift modeling and pairwise ranking we provided the first data-dependent generalization lower bound for the AUUC and introduced the corresponding objective of bound optimization, usable with linear and deep models. We empirically studied the tightness of this proposed bound, its effectiveness for hyperparameters tuning, and showed the efficiency of the proposed learning objective compared to a wide range of competitive baselines on two classical uplift modeling benchmarks using real-world datasets.

Finally, in **Chapter 6**, we considered the problem of learning uplift models with access to both labels and features only through aggregated queries upon groups. The interest in this problem was motivated by the recent increase of privacy constraints in different domains. We introduced ϵ-Aggregated Data Uplift Model (ϵ-ADUM), a differentially private method to learn uplift models from data aggregated over a given partition of the feature space. Then we identified a bias-variance decomposition of the well-known metric in the field, namely the Precision in Estimation of Heterogeneous Effects (PEHE) under ϵ-ADUM setup, and highlighted the role of underlying partition size in the privacy-utility trade-off. Series of experiments highlighted the bias-variance trade-off and confirmed theoretical derivations concerning the optimal number of groups. Along with this, we showed, running experiments on synthetic and real data sets, that group-based uplift models are competitive to baselines with full access to the data, suggesting that aggregation does not significantly penalize uplift modeling while guaranteeing privacy protection.

Future Perspectives

For Dependent Data Representation (Section 4.3) and Shared Data Representation (Section 4.4), reasonable direction of future work includes adapting the methods to be usable with deep neural networks, providing non-linear data representations in order to learn richer interactions between features and treatment, similar to TARNet [94]. Particular research is currently underway, although without meaningful results so far – we attribute this to the excessive prone of deep models to overfitting, especially in uplift modeling. Meanwhile, techniques that are similar to the desired deep version of

Shared Data Representation are introduced in several recent papers [35, 36], proving the rationality of the idea.

For the work in Section 5, first area for future research would be to enhance the current bound for $\mathfrak{R}_{S_t}(\mathcal{F}_r)$ (Proposition 3) in order to make bound for AUUC tighter and to relax additional assumptions used in proposed version. As a first step, one can apply novel approach of similar bounding for deep models from [13]. In addition, we are interested in adapting our learning objective AUUC-max to other models in the field, such as Shared Data Representation or TARNet – the idea is to combine rich data representations with the ability to directly optimize AUUC.

Regarding the work in Section 6, a sensible branch of further work is improving ϵ-ADUM and theory behind for more realistic cases, omitting the simplified assumptions of uni-dimensional data or equal sizes of treated and control units inside the groups G. Another line of research concerns an adaptation of the method to other partition techniques for multi-dimensional cases.

Bibliography

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

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

[4] Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr):393–425, 2005.

[5] Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201. PMLR, 2019.

[6] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[7] Ahmed M Alaa and Mihaela van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

[8] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[9] Amiran Ambroladze, Emilio Parrado-Hernández, and John S Shawe-taylor. Tighter pac-bayes bounds. In *Advances in neural information processing systems*, pages 9–16, 2007.

[10] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

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

[12] Susan Athey, Stefan Wager, et al. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 78, 2017.

[13] Andrew R Barron and Jason M Klusowski. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*, 2019.

[14] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[15] Mouloud Belbahri, Alejandro Murua, Olivier Gandouet, and Vahid Partovi Nia. Qini-based uplift regression. *arXiv preprint arXiv:1911.12474*, 2019.

[16] Ron Berman. Beyond the Last Touch : Attribution in Online Advertising. *Preliminary Version*, 2013.

[17] 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, 2018.

[18] Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift modeling with generalization guarantees. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 55–65, 2021.

[19] Artem Betlei, Théophane Gregoir, Thibaud Rahier, Aloïs Bissuel, Eustache Diemert, and Massih-Reza Amini. Differentially Private Individual Treatment Effect Estimation from Aggregated Data. working paper or preprint, September 2021.

[20] 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.

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

[22] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[24] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[25] Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81, 2010.

[26] Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and Scalable Response Prediction for Display Advertising. *ACM Trans. Intell. Syst. Technol.*, 5(4):61:1—-61:34, dec 2014.

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

[28] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalml: Python package for causal machine learning, 2020.

[29] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[30] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

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

[32] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, pages 313–320, 2004.

[33] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[34] Criteo. Criteo releases industry largest ever dataset for machine learning to academic community. *NewsWire*, 2015.

[35] Alicia Curth and Mihaela Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

[36] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *arXiv preprint arXiv:2106.03765*, 2021.

[37] Floris Devriendt, Darie Moldovan, and Wouter Verbeke. 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, 2018.

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

[39] Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM, 2018.

[40] 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.

[41] Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the ADKDD'17*, pages 1–6. 2017.

[42] 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.

[43] Carlos Fernandez, Foster Provost, Jesse Anderton, Benjamin Carterette, and Praveen Chandar. Methods for individual treatment assignment: An application and comparison for playlist generation. *arXiv preprint arXiv:2004.11532*, 2020.

[44] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

[45] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[46] Samuel M Gross and Robert Tibshirani. Data shared lasso: A novel tool to discover uplift. *Computational statistics & data analysis*, 101:226–235, 2016.

[47] X5 Retail Group. X5 RetailHero dataset. `https://storage.yandexcloud.net/datasouls-ods/materials/9c6913e5/retailhero-uplift.zip`, December 2019.

[48] Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems*, 72:24–32, 2015.

[49] Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. Uplift random forests. *Cybernetics and Systems*, 46(3-4):230–248, 2015.

[50] Pierre Gutierrez and Jean-Yves Gérardy. Causal Inference and Uplift Modeling A review of the literature. 67:1–13, 2016.

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

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

[53] Kevin Hillstrom. The MineThatData e-mail analytics and data mining challenge. http://www.minethatdata.com/Kevin_Hillstrom_MineThatData_E-MailAnalytics_DataMiningChallenge_2008.03.20.csv, March 2008.

[54] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[55] 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.

[56] S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures and Algorithms*, 24(3):234–248, 2004.

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

[58] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384, 2005.

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

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

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

[62] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[63] 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.

[64] Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support vector machines for differential prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer, 2014.

[65] Lily Yi-Ting Lai. *Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers*. PhD thesis, School of Computing Science-Simon Fraser University, 2006.

[66] John Langford. Tutorial on practical prediction theory for classification. *Journal of machine learning research*, 6(Mar):273–306, 2005.

[67] Damien Lefortier, Adith Swaminathan, Xiaotao Gu, Thorsten Joachims, and Maarten de Rijke. Large-scale validation of counterfactual learning methods: A test-bed. *arXiv preprint arXiv:1612.00367*, 2016.

[68] Chenchen Li, Xiang Yan, Xiaotie Deng, Yuan Qi, Wei Chu, Le Song, Junlong Qiao, Jianshan He, and Junwu Xiong. Reinforcement learning for uplift modeling. *arXiv preprint arXiv:1811.10158*, 2018.

[69] Victor SY Lo. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86, 2002.

[70] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

[71] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[72] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

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

[74] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[75] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2020.

[76] Diego Olaya, Kristof Coussement, and Wouter Verbeke. A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery*, 34(2):273–308, 2020.

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

[78] Judea. Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, 2000.

[79] Judea Pearl. *Causality*. Cambridge university press, 2009.

[80] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[81] Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical bernstein inequalities for u-statistics. In *Neural Information Processing Systems (NIPS)*, number 23, pages 1903–1911, 2010.

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

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

[84] 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.

[85] Liva Ralaivola and Massih-Reza Amini. Entropy-based concentration inequalities for dependent variables. In *International Conference on Machine Learning*, pages 2436–2444, 2015.

[86] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, jun 2011.

[87] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

[88] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

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

[91] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

[92] Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *International Conference on Machine Learning*, pages 8398–8407. PMLR, 2020.

[93] Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

[94] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[95] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019.

[96] Michał Sołtys, Szymon Jaroszewicz, and Piotr Rzepakowski. Ensemble methods for uplift modeling. *Data mining and knowledge discovery*, 29(6):1531–1559, 2015.

[97] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

[99] Patrick D Surry and Nicholas J Radcliffe. Quality measures for uplift models. `http://www.stochasticsolutions.com/pdf/kdd2011late.pdf`, 2011.

[100] Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

[101] Nicolas Usunier, Massih Amini, and Patrick Gallinari. A data-dependent generalisation error bound for the AUC. In *ICML'05 workshop on ROC Analysis in Machine Learning*, page 8, Bonn, Germany, August 2005.

[102] Nicolas Usunier, Massih R Amini, and Patrick Gallinari. Generalization error bounds for classifiers trained with interdependent data. In *Advances in neural information processing systems*, pages 1369–1376, 2006.

[103] 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, jul 2018.

[104] Ikko Yamane, Florian Yger, Jamal Atif, and Masashi Sugiyama. Uplift modeling from separate labels. In *Advances in Neural Information Processing Systems*, pages 9927–9937, 2018.

[105] Lian Yan, Robert H Dodier, Michael Mozer, and Richard H Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 848–855, 2003.

[106] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

[107] 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.

[108] Lukasz Zaniewicz and Szymon Jaroszewicz. Support Vector Machines for Uplift Modeling. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, ICDMW '13, pages 131–138, Washington, DC, USA, 2013. IEEE Computer Society.

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

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

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

[112] Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift Modeling with Multiple Treatments and General Response Types. 2017.

[113] Zhenyu Zhao and Totte Harinen. Uplift modeling for multiple treatments with cost optimization. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 422–431. IEEE, 2019.