Uplift Modeling with Generalization Guarantees

Artem Betlei, Eustache Diemert, Massih-Reza Amini

Uplift Modeling with Generalization Guarantees

Artem Betlei[*,†], Eustache Diemert[*], Massih-Reza Amini[†]
[*] Criteo AI Lab, [†] UGA/CNRS LIG
Grenoble, France

Abstract

In this paper, we consider the task of ranking individuals based on the potential benefit of being "treated" (e.g. by a drug or exposure to recommendations or ads), referred to as Uplift Modeling in the literature. This application has gained a surge of interest in recent years and it is found in many applications such as personalized medicine, recommender systems or targeted advertising. In real life scenarios the capacity of models to rank individuals by potential benefit is measured by the Area Under the Uplift Curve (AUUC), a ranking metric related to the well known Area Under ROC Curve. In the case where the objective function, for learning model parameters, is different from AUUC, the capacity of the resulting system to generalize on AUUC is limited. To tackle this issue, we propose to learn a model that directly optimizes an upper bound on AUUC. To find such a model we first develop a generalization bound on AUUC and then derive from it a learning objective called AUUC-max, usable with linear and deep models. We empirically study the tightness of this generalization bound, its effectiveness for hyperparameters tuning and show 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.

1 Introduction

In many applications there is a need to target actions to specific portions of a population so as to maximize a global utility. For instance, in personalized

medicine one is interested in prescribing a treatment only to patients for whom it would be beneficial [23].

Similarly in performance marketing, one would prefer to target advertisement budget towards potential customers that would be more likely to be persuadable to purchase [15]. The term *uplift* designates the expected outcome difference between treated and untreated individuals and is related to treatment effectiveness. Practitioners are interested in models that predict an individual uplift given some observable characteristics of the individuals. Such predictions would then be used to design future treatment policies, usually under some budget constraints.

Recent reviews of the Uplift Modeling literature [14, 52] illustrate how such problems arise in a wide range of economic activities (credit scoring [35], catalog mailing, customer retention [34], insurance [18]), medicine studies (bone marrow transplant, tamoxifen prescription, hepatitis [23] and breast cancer treatment [27]) and social sciences evaluations (job training programs [21], psychology [26] and student growth [5]).

Illustrative example. Fig. 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 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) [39]. 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) [14, 16].

For a new cohort of individuals is available, the predictions of the model will be then used to target treatment: highest scored individuals would be treated (green individuals in Figure 1) whilst the lowest scored ones would be excluded from treatment (blue individuals). This strategy is useful as soon as treatment effect is heterogeneous (i.e. depending on observable covariates). Note that the prediction value itself is not of interest here but rather the ranking induced by the predictions.

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

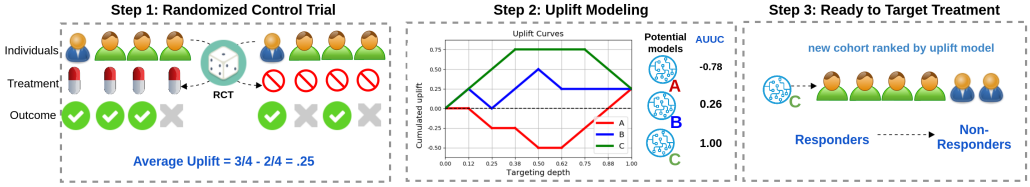

Figure 1: Typical Uplift Modeling pipeline schematized in three steps. Step 1 starts with a randomized control trial using an A/B test. Then Step 2 consists in learning and evaluating several uplift models and selecting the best performing one by AUUC on data gathered at Step 1. Finally at Step 3, the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

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). 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 [12].

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. 4). 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 [28]. These bounds are usually used for model selection or to analyze the model's generalization ability. Recent works in individual treatment effect estimation (see Section 2) and Uplift Modeling fields

propose to bound generalization error of Precision when estimating Heterogeneous Effect (PEHE) [42] 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 [49]. But as discussed above, these pointwise objective functions are not the most appropriate for AUUC.

Our 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.

(*a*) In Section 3, we propose the first generalization bound for AUUC using data-dependent concentration inequalities on dependent variables.

(*b*) In Section 3.3, 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.

(*c*) Section 4 reports a thorough performance evaluation against a range of competitive baselines on two real-world datasets.

2 Background

Notations. Let $\mathcal{X} \in \mathbb{R}^d$, $\mathbf{x} \in \mathcal{X}$ be a feature vector, and $Y \in \{0, 1\}$ be the outcome variable, indicating positive ($y = 1$) or negative ($y = 0$) outcome. Additionally, let the treatment variable $G \in \{T, C\}$ denote whether an individual receives treatment ($g = T$) or not ($g = C$). We assume a dataset from the RCT: $(\mathbf{x}_i, y_i, g_i) \overset{\text{iid}}{\sim} P_{\mathcal{X},Y,G}; \mathcal{X} \perp\!\!\!\perp G$. We define then $S_g = \{\mathbf{x}_i, y_i, g\}_{i=1\ldots n_g}$ as the particular subset of the training set S of size N, i.e. $S = S_T \sqcup S_C$ and $N = n_T + n_C$. Also let $\tilde{S}_g = \{\mathbf{x}_i, 1 - y_i, g\}_{i=1\ldots n_g}$ be the version of subset with reverted labels. We define $\bar{y}_g = \mathbb{E}[Y|G = g]$ and $\lambda_g = \bar{y}_g(1 - \bar{y}_g)$ the treatment dependent conditional mean and variance of Bernoulli outcome. Finally, let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ be the set of real-valued functions.

How to evaluate models? We will review here the formal definition of AUUC along with basic concepts needed to frame the problem of maximizing AUUC. It is important to recall firstly that in Uplift Modeling one can never observe a given individual in both treated and untreated conditions which is also called *fundamental problem of causal inference*. Therefore metrics computing a difference to the true treatment effect can only work in a simulation setting, where it is possible to assign an individual to group T and C simultaneously and hence know both outcomes. A popular example of such a metric is Precision when Estimating Heterogeneous Effect (PEHE) [42]. However, in real-life scenarios when only one of the outcomes is observed, using Uplift Model f one can estimate cumulative uplift for group of k points as $V(f, k)$ (see Def. 1). This idea underlies the *Area Under the Uplift Curve* (AUUC) [39] is the most popular method for evaluating Uplift Model in the literature. Intuitively, with a good model, the best scored individuals should yield a high prediction. An uplift curve can verify this property: it ranks individual samples according to *predicted* uplift (see the X-axis of Fig. 1) and cumulatively sum the *observed* uplift (in Y-axis). The AUUC is then the area under this curve. On Fig. 1 - middle picture, Uplift Model C is better than A or B as its area is greater, denoting that cumulated uplift is greater when individuals are ranked according to model C. AUUC is thus useful when *i*) all individuals could not be "treated" (e.g. because of limited budget) and *ii*) data comes from real-world (no simulation). Intuitively, AUUC penalizes heavily ranking errors on highest scored individuals, which is reasonable for settings where treatment would be administered only to highest predictions in future cohorts.

Formalization of AUUC. The Uplift Modeling literature yields multiple variants of 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 Table 2 in [13] for a comprehensive picture of available alternatives. We chose the "separate, relative" uplift curve introduced in [13], 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 Def. 1, corresponding to (Eq. 10 and 16) of [13].

Definition 1 (Area Under the Uplift Curve)**.** *Let* $f(S_T, \frac{p}{100}n_T)$ *and* $f(S_C, \frac{p}{100}n_C)$ *be the first* p *percentages of* S_T *and* S_C *respectively when both ordered by prediction of model* f*. The empirical AUUC of the model* f *on* S_T *and* S_C *is*

given by:

$$\widehat{AUUC}(f, S_T, S_C) = \int_0^1 V(f, x)dx \approx \sum_{p=1}^{100} V(f, \frac{p}{100})$$

where

$$V(f, \frac{p}{100}) = \frac{1}{n_T} \sum_{i \in f(S_T, \frac{p}{100} n_T)} y_i - \frac{1}{n_C} \sum_{j \in f(S_C, \frac{p}{100} n_C)} y_j$$

Basic models. We now introduce popular uplift models to materialize the task. *Two Models (TM)* [19] is a trivial method to predict uplift. It uses two separate probabilistic models to predict outcome in treated or untreated conditions:

$$u^{TM}(\mathbf{x}) = P(y = 1|\mathbf{x}, g = T) - P(y = 1|\mathbf{x}, g = C) \quad (1)$$

and any prediction model can be used for the estimation of posteriors (typically logistic regression). We notice that when the average response is low and/or noisy there is the risk for the difference of predictions to be very noisy too and lead to arbitrary ranking of individuals overall (see [36] for a detailed critic). This remark makes a general argument for using methods that combine knowledge of both parts of the dataset. Multi-task approaches, e.g. *Shared Data Representation* (*SDR*) have been proposed in [8] to overcome this problem and showed better empirical performance when the treatment is imbalanced. *Class Variable Transformation (CVT)* [23] combines binary treatment and outcome in order to use a single classification model. For this purpose a new label and predictor are defined:

$$\begin{cases} Z = YG + (1 - Y)(1 - G) \\ u^{CVT}(\mathbf{x}) = 2P(z = 1|\mathbf{x}) - 1 \end{cases} \quad (2)$$

Similar label transformations could be traced back to Robinson [38] and extended to more general settings [32, 4]. Other related, productive lines of research have been *i*) the adaptation of split criteria of Decision Trees [36, 40, 43] for uplift prediction and; *ii*) deep representation learning approaches for the observational case that carefully match treated/control embedding distributions [42, 9, 29].

Relation with CATE. In causal inference one is often interested in estimating the Individual Treatment Effect (ITE), which corresponds to the individual difference of *potential outcomes* in the Neyman-Rubin causal framework [41]. Such a quantity is not observable but plays an important theoretical role. An empiric counter-part is the Conditional Average Treatment Effect (CATE) which measures or predicts the expected difference in outcome (Y) between being treated or not (T), conditionally to observed co-variates (X): $CATE(x) = \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]$. This quantity can be estimated from data and it is used to predict treatment benefits individually. Such predictions induce a natural ordering of individuals which in turn can be used to target future treatment to a portion of the population. In fact many usual Uplift Modeling methods learn the CATE and it is known that a perfect CATE model would maximize AUUC [49]. The interested reader can refer to [52] for a unified view of Uplift Modeling and CATE.

3 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 3.1); and by means of Rademacher concentration inequalities build a generalization bound (Section 3.2). Then we define a principled optimization method with generalization guarantees for AUUC that leverages the bound as a robust learning objective (Section 3.3). Finally, we review related approaches and their merits as found in the literature (Section 3.4).

3.1 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 1.

Proposition 1. *Let* $\widehat{AUUC}(f, S_T, S_C)$ *be the empirical area under uplift curve of the model* f *on the sets* S_T *and* S_C*; and* $AUUC(f) = \mathbb{E}_{S_T,S_C}\left[\widehat{AUUC}(f, S_T, S_C)\right]$ *be its expectation. Then* $AUUC(f)$ *is related to ranking loss (Eq. 4) as:*

$$AUUC(f) = \gamma - \Big(\lambda_T \mathbb{E}_{S_T}[\hat{R}(f, S_T)] + \lambda_C \mathbb{E}_{\tilde{S}_C}[\hat{R}(f, \tilde{S}_C)]\Big) \tag{3}$$

where

$$\hat{R}(f, S_g) = \frac{1}{n_g^+ n_g^-} \sum_{(\mathbf{x}_i,+1)\in S_g} \sum_{(\mathbf{x}_j,0)\in S_g} \mathbb{1}_{f(\mathbf{x}_i)\le f(\mathbf{x}_j)} \tag{4}$$

is the empirical bipartite ranking risk, $g \in \{T, C\}$*,* n_g^+, n_g^- *are the amounts of positives and negatives respectively in the set* S_g *(i.e.* $n_g = n_g^+ + n_g^-$*), and* $\gamma = \mathbb{E}_{S_T,S_C}[\bar{y}_T - \frac{(\bar{y}_T)^2}{2} - \frac{(\bar{y}_C)^2}{2}]$*.*

Proof. From Definition 1:

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

[45, Eq. 13] allows us to express $V(f, x)$ as a difference of *cumulative outcome rates* $F_f^{S_T}(x)$ and $F_f^{S_C}(x)$ (for the formal definition please refer to [45]) of collections S_T and S_C respectively, induced by model f:

$$V(f, x) = F_f^{S_T}(x) - F_f^{S_C}(x)$$

Hence,

$$\begin{aligned}\widehat{AUUC}(f, S_T, S_C) &= \int_0^1 V(f, x)dx = \int_0^1 \Big(F_f^{S_T}(x) - F_f^{S_C}(x)\Big)dx \\ &= \int_0^1 F_f^{S_T}(x)dx - \int_0^1 F_f^{S_C}(x)dx \end{aligned} \tag{5}$$

By the mean while, we have from [45, Eq. 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:

$$Gini(f, \mathcal{D}) = \frac{2\int_0^1 F_f^{\mathcal{D}}(x)dx - \bar{y}_{\mathcal{D}}}{\bar{y}_{\mathcal{D}}(1 - \bar{y}_{\mathcal{D}})} \tag{6}$$

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 [46]:

$$Gini(f, \mathcal{D}) = 2AUC(f, \mathcal{D}) - 1 \tag{7}$$

From (6) and (7), it then comes :

$$\int_0^1 F_f^{\mathcal{D}}(x)dx = \bar{y}_{\mathcal{D}}(1 - \bar{y}_{\mathcal{D}}) \cdot AUC(f, \mathcal{D}) + \frac{(\bar{y}_{\mathcal{D}})^2}{2} \tag{8}$$

From (5) and (8) it comes :

$$\begin{aligned}\widehat{AUUC}(f, S_T, S_C) &= \bar{y}_T(1 - \bar{y}_T) \cdot AUC(f, S_T) \\ &- \bar{y}_C(1 - \bar{y}_C) \cdot AUC(f, S_C) + \frac{(\bar{y}_T)^2}{2} - \frac{(\bar{y}_C)^2}{2}\end{aligned}$$

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

$$\begin{aligned}&\widehat{AUUC}(f, S_T, S_C) = \bar{y}_T(1 - \bar{y}_T)AUC(f, S_T) \\ &+ \bar{y}_C(1 - \bar{y}_C)\Big(1 - AUC(f, \tilde{S}_C)\Big) + \frac{(\bar{y}_T)^2}{2} - \frac{(\bar{y}_C)^2}{2} \\ &= \bar{y}_T(1 - \bar{y}_T) \cdot AUC(f, S_T) \\ &+ \bar{y}_C(1 - \bar{y}_C) \cdot AUC(f, \tilde{S}_C) + \frac{(\bar{y}_T)^2}{2} + \frac{(\bar{y}_C)^2}{2} - \bar{y}_C\end{aligned}$$

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

$$\begin{aligned}&\widehat{AUUC}(f, S_T, S_C) = \bar{y}_T(1 - \bar{y}_T) \cdot \Big(1 - \hat{R}(f, S_T)\Big) \\ &+ \bar{y}_C(1 - \bar{y}_C) \cdot \Big(1 - \hat{R}(f, \tilde{S}_C)\Big) + \frac{(\bar{y}_T)^2}{2} + \frac{(\bar{y}_C)^2}{2} - \bar{y}_C \\ &= \hat{\gamma}_{S_T, S_C} - \Big(\lambda_T \hat{R}(f, S_T) + \lambda^C \hat{R}(f, \tilde{S}_C)\Big)\end{aligned}$$

where, for sake of notation, we use group T and group C instead of datasets S_T and S_C in the upper indices of $\bar{y}$; and $\lambda_T = \bar{y}_T(1 - \bar{y}_T), \lambda_C = \bar{y}_C(1 - \bar{y}_C), \hat{\gamma}_{S_T, S_C} = \bar{y}_T - \frac{(\bar{y}_T)^2}{2} - \frac{(\bar{y}_C)^2}{2}$.

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

$$AUUC(f) = \mathbb{E}_{S_T,S_C}\left[\widehat{AUUC}(f, S_T, S_C)\right] \\ = \gamma - \left(\lambda_T \mathbb{E}_{S_T}[\hat{R}(f, S_T)] + \lambda^C \mathbb{E}_{\tilde{S}_C}[\hat{R}(f, \tilde{S}_C)]\right)$$

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

3.2 Rademacher Generalization Bounds

Let us now consider the minimization problems of the pairwise ranking losses over the treatment and the control subsets (Eq. 4), and the following dyadic transformation defined over each of the groups S_T and $\tilde{S}_C$:

$$\mathcal{T}(S_g) = \left\{(\mathbf{z} = (\mathbf{x}, \mathbf{x}'), \tilde{y}) \,\middle|\, ((\mathbf{x}, y), (\mathbf{x}', y')) \in S_g \times S_g \wedge y \neq y'\right\} \tag{9}$$

where, $g \in \{T, C\}$, $\tilde{y} = +1$ iff $y = +1$ and $y' = 0$ and $\tilde{y} = -1$ otherwise. Here we suppose that $\mathcal{T}(S_g)$ 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 $n_g^+ n_g^-$ pairs of observations in S_g that are from different classes.

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

$$\mathcal{H} = \{h : \mathbf{z} = (\phi(\mathbf{x}), \phi(\mathbf{x}')) \mapsto f(\phi(\mathbf{x})) - f(\phi(\mathbf{x}')), f \in \mathcal{F}\}, \tag{10}$$

where, $\phi(\mathbf{x})$ is the feature representation associated to observation $\mathbf{x}$. The empirical loss (Eq. 4) can then be rewritten as:

$$\hat{R}(h, \mathcal{T}(S_g)) = \frac{1}{n_g^+ n_g^-} \sum_{(\mathbf{z},\tilde{y}) \in \mathcal{T}(S_g)} \mathbb{1}_{\tilde{y}h(\mathbf{z}) \leq 0}. \tag{11}$$

The loss defined in (Eq. 11) is equivalent to a binary classification error over the pairs of examples in $\mathcal{T}(S_g)$. With this equivalence, one may expect to use efficient generalization bounds developed in binary classification. However, (Eq. 11) is a sum over random dependent variables; as each training examples in S_g may be present in different pairs of examples in $\mathcal{T}(S_g)$, 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 $\mathcal{T}(S_g)$ as a dependency graph of random variables on its nodes, and similar to [48], we decompose it using the *exact proper fractional cover* of the graph proposed by [22] 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 Fig. 2, where a set of example S_g 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 $\mathcal{T}(S_g)$ and the induced dependency graph $\mathcal{G}$ between pairs of examples (where edges denote statistical dependence between pairs in $\mathcal{T}(S_g)$; the minimal coloring of $\mathcal{G}$ that are covers containing each independent pairs is, in this case, equal to the fractional chromatic number $\chi^*(\mathcal{G})$.

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 [48]. 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 [37]; based on local Rademacher complexities [7] 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.

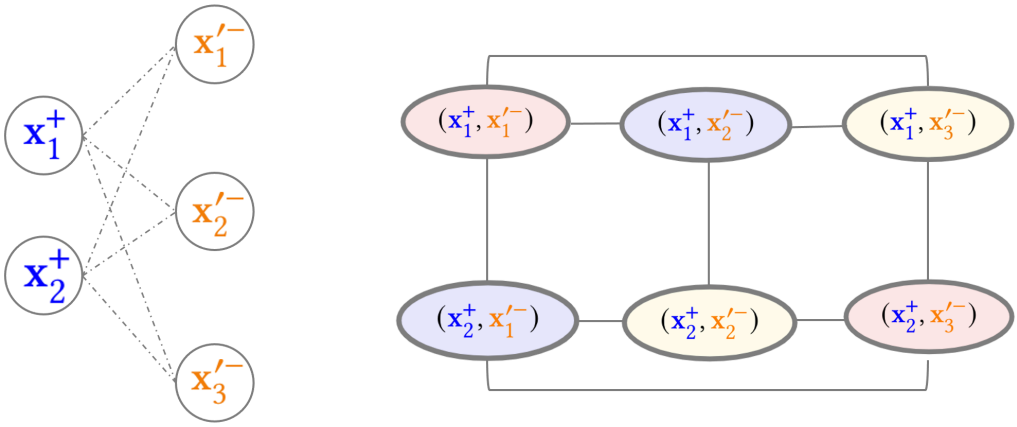

Figure 2: Dependency structure of a bipartite ranking problem composed of $n_g^+ = 2$ positive and $n_g^- = 3$ negative examples. (left: original data S_g and the composition of pairs shown in dashed; right: induced dependency graph $\mathcal{G}$; edges indicate dependencies between pairs in $\mathcal{T}(S_g)$,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

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

$$\mathfrak{R}_{S_g}(\mathcal{F}_r) = \frac{1}{n_g^+ n_g^-} \mathbb{E}_\sigma \left[\sum_{j \in [J]} \omega_j \mathbb{E}_{X_{\mathcal{C}_j}} \left[\sup_{f \in \mathcal{F}_r} \sum_{i \in \mathcal{C}_j} \sigma_i f(\mathbf{x}_i) \right] \right] \tag{12}$$

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

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_T *and reverted label control* $\tilde{S}_C$ *subsets. Let* $\mathcal{T}(S_T)$ *and* $\mathcal{T}(\tilde{S}_C)$ *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* $f \in \mathcal{F}_r$*:*

$$AUUC(f) \geq \gamma - \Big(\lambda_T \hat{R}_\ell(f, S_T) + \lambda_C \hat{R}_\ell(f, \tilde{S}_C)\Big) - \mathfrak{C}_\delta(\mathcal{F}_r, S_T, \tilde{S}_C) - \frac{25}{48}\left(\frac{\lambda_T}{n_T^+} + \frac{\lambda_C}{n_C^+}\right)\log\frac{2}{\delta}$$

where, $\mathfrak{C}_\delta(\mathcal{F}_r, S_T, \tilde{S}_C) = \lambda_T \mathfrak{R}_{S_T}(\mathcal{F}_r) + \lambda_C \mathfrak{R}_{\tilde{S}_C}(\mathcal{F}_r) + \left(\frac{\frac{5}{2}\sqrt{\mathfrak{R}_{S_T}(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_T^+}}\lambda_T + \frac{\frac{5}{2}\sqrt{\mathfrak{R}_{\tilde{S}_C}(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_C^-}}\lambda_C\right)\sqrt{\log\frac{2}{\delta}}$ *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 1:

$$AUUC(f) = \gamma - \Big(\lambda_T \mathbb{E}_{S_T}[\hat{R}(f, S_T)] + \lambda_C \mathbb{E}_{\tilde{S}_C}[\hat{R}(f, \tilde{S}_C)]\Big) \tag{13}$$

From [37], we have the following upper bounds for each of the ranking losses hold with probability $1 - \delta/2$:

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

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

The infinimums of the upper-bounds are reached for respectively

$$a_T = \frac{5}{4}\sqrt{\frac{\log\frac{2}{\delta}}{n_T^+\mathfrak{R}_{S_T}(\mathcal{F}_r)}}, \qquad a_C = \frac{5}{4}\sqrt{\frac{\log\frac{2}{\delta}}{n_C^-\mathfrak{R}_{\tilde{S}_C}(\mathcal{F}_r)}}$$

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.

3.3 AUUC-max Learning Objective

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

$$\begin{aligned}&\underset{f\in\mathcal{F}_r}{\operatorname{argmax}}\, AUUC(f) \equiv \\ &\underset{\theta,r}{\operatorname{argmin}}\Big(\lambda_T \hat{R}(f_\theta, S_T) + \lambda_C \hat{R}(f_\theta, \tilde{S}_C) + \mathfrak{C}_\delta(\mathcal{F}_r, S_T, \tilde{S}_C)\Big)\end{aligned} \tag{14}$$

where θ are parameters of the model.

There are two remarks that we can make at this point. First, both terms $\hat{R}(f_\theta, S_T)$ and $\hat{R}(f_\theta, \tilde{S}_C)$ in (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 $\mathfrak{R}_{S_T}(\mathcal{F}_r)$ and $\mathfrak{R}_{S_C}(\mathcal{F}_r)$ that appear in $\mathfrak{C}_\delta(\mathcal{F}_r, S_T, \tilde{S}_C)$ 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, such as $s_{log}(z) = \ln(1+e^{-z}) / \ln(2)$ and $s_{poly}(z) = (-(z-\mu))^p \mathbb{1}_{z<\mu}$ [50]. 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 $\mathfrak{R}_{S_T}(\mathcal{F}_r)$ and $\mathfrak{R}_{S_C}(\mathcal{F}_r)$ following Proposition 2.

Proposition 2. *Let S_g be a sample of size n_g with n_g^+ samples with positive labels and such that $\forall \mathbf{x} \in S_g \ \|\phi(\mathbf{x})\| \leq R$. 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 $\mathcal{T}(S_g)$ of size $n_g^+ n_g^-$, can be bounded with probability at least $1 - \frac{\delta}{2}$ by:*

$$\mathfrak{R}_{S_g}(\mathcal{F}_r) \leq \sqrt{\frac{R^2\Lambda^2}{n_g^+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_g^+}} \tag{15}$$

Proof.

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

$$= \frac{1}{n_g^+ n_g^-} \sum_{j\in[J]} |\mathcal{C}_j| \underbrace{\mathbb{E}_{X_{\mathcal{C}_j}} \left[\hat{\mathfrak{R}}_{\mathcal{C}_j}(\mathcal{F}_r) \right]}_{\mathfrak{R}_{\mathcal{C}_j}(\mathcal{F}_r)}$$

$$\overset{\text{[31, Eq. 3.14]}}{\leq} \frac{1}{n_g^+ n_g^-} \sum_{k=1}^{n_g^-} n_g^+ \left(\hat{\mathfrak{R}}_{\mathcal{C}_j}(\mathcal{F}_r) + \sqrt{\frac{\log\frac{2}{\delta}}{2n_g^+}} \right)$$

$$\overset{\text{[31, Th. 4.3]}}{\leq} \frac{1}{n_g^+ n_g^-} \sum_{k=1}^{n_g^-} n_g^+ \left(\sqrt{\frac{R^2\Lambda^2}{n_g^+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_g^+}} \right) = \sqrt{\frac{R^2\Lambda^2}{n_g^+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_g^+}}.$$

□

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 appendix A). Noting that R is a constant depending on the set of feature representations we can transform the optimization problem in (θ, r) in (Eq. 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 [44] 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 (14) that we consider is given in (Eq. 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 (Eq. 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 (Eq. 16) using feature representations $\phi(\mathbf{x})$.

AUUC-max optimization problem

$$\min_{\mathbf{w}} \hat{\mathcal{L}}_{\mathbf{w}}(S_T, \tilde{S}_C) = \frac{1}{n_T^+ n_T^-} \sum_{\{\mathbf{x}_i,+1\}\in S_T} \sum_{\{\mathbf{x}_j,0\}\in S_T} s(\mathbf{w}^\top \phi(\mathbf{x}_i) - \mathbf{w}^\top \phi(\mathbf{x}_j)$$
$$+ \frac{1}{n_C^+ n_C^-} \sum_{\{\mathbf{x}_k,+1\}\in \tilde{S}_C} \sum_{\{\mathbf{x}_l,0\}\in \tilde{S}_C} s(\mathbf{w}^\top \phi(\mathbf{x}_k) - \mathbf{w}^\top \phi(\mathbf{x}_l)) + \mathfrak{C}_\delta(\mathcal{F}_{\Lambda^2 R^2}, S_T, \tilde{S}_C) \quad (16)$$
$$\text{subject to} \quad \|\mathbf{w}\| \leq \Lambda$$

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 4).

Note that the usual practice for Uplift Models (see Fig. 1) 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.

3.4 Related Works

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 [27] proposes to maximize AUUC directly by expressing it as a weighted sum of two AUCs and maximizing it using a Support Vector Machine method. 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 [13] draw a list-wise learning to rank formulation of AUUC and use the LambdaMART [10] algorithm to optimize it, alleviating the need for derivable surrogates at the price of more complex models.

Representation learning for CATE prediction. Important work has been published recently using a broad family of methods to perform CATE prediction: TARNet, CFRNet [42], CRN [9], CEVAE [29], BNN [24],

GANITE [51], DeepMatch [25]. Most of the methods are designed for observational data and revolve around ways to lessen the covariate shift between $P_{X|G=1}$ and $P_{X|G=0}$ due to treatment selection bias. However, several approaches (TARNet, GANITE) perform well also in randomized case. TARNet represents deep architecture where group-dependent sets of layers (for groups T and C) follow the shared set of layers (for all data). GANITE is a variant of generative adversarial network that consists of two blocks: counterfactual block imputes counterfactual outcome, then ITE block learn distribution of ITE having access to both factual and counterfactual outcomes. Nevertheless, all of these methods aim to estimate CATE and not focus on finding optimal ranking.

Generalization bounds. The work of [42] 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, [49] proposed a generalization bound for uplift prediction. However, the main differences with our approach is that the upper-bound of AUUC proposed in [49] is a MSE-like proxy that is applicable in the case where the variables Y and G are never observed together whereas we bound AUUC directly without such hypothesis. Further, the definition of the proxy objective proposed in [49] assumes that samples are i.i.d., whilst in our study the equivalence between the ranking objective (4) and the classification error over the pairs of examples (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 [7]. Finally, from an optimization side the approach developed in [49] leads to a mini-max optimization problem, that is avoided in AUUC-max by using the "revert label in control" trick.

4 Experimental Results

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. *Hillstrom* [20] contains results of an e-mail campaign for an Internet retailer. Treatment (receiving women's merchandise e-mails) is balanced and independent of co-variates; outcome is visiting the retailer website. *Criteo-UPLIFT v2* [15] is a large scale dataset donated by the AdTech company Criteo as part of a contribution to the AdKDD'18 workshop. It is constructed from incre-

mentality A/B tests, a special procedure where a portion of users that could be exposed to advertising banners is prevented to see them. Treatment (seeing ads) is therefore severely unbalanced as it incurs a loss of business to the advertising platform. Note that this dataset is much larger-scale than Hillstrom with 13M samples. We report results for the "visit" outcome which has a very low positive ratio and average uplift compared to Hillstrom. This fact explains the lower AUUC numbers that are observed in the former (Table 1).

Table 1: Benchmark datasets

Data set	Hillstrom	Criteo-UPLIFT2
Size	42,693	13,979,592
Features	22	12
Group T ratio	0.49905	0.85
Positive class ratio	0.12883	0.047
Pos. class ratio in group T	0.1514	0.04854
Pos. class ratio in group C	0.10617	0.03820
Average uplift	0.04523	0.01034

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 appendix B, as well as the details about used prediction models. Finally, algorithms are compared by AUUC on test set, using an empirical Bernstein bound [30] to compute a 95% confidence test set bound on the expectation of AUUC. More details are provided in appendix C.

Evidence of the 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) [23] introduced in Section 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 (Fig. 3 top) and between the

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

training loss and AUUC on the test set (Fig. 3 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).

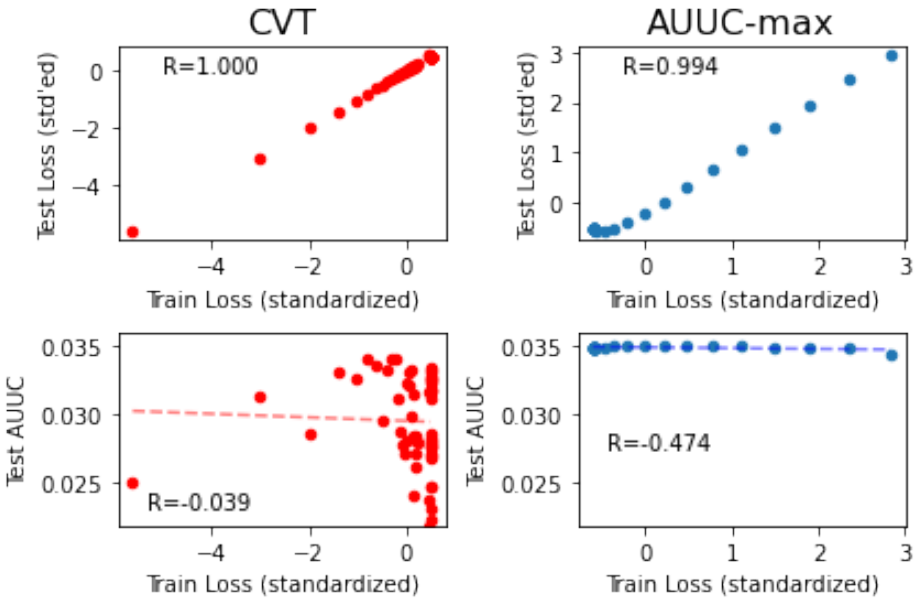

Figure 3: AUUC generalization problem with a pointwise objective on the Hillstrom dataset. CVT optimizing pointwise log-loss objective (left), AUUC-max optimizing (Eq. 16) - right. R is the correlation coefficient.

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 Th. 1: using the local fractional Rademacher concentration inequality on bipartite ranking risk (our proposition, in blue on Fig. 4) or [2, 47] (in orange) or [17] (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.

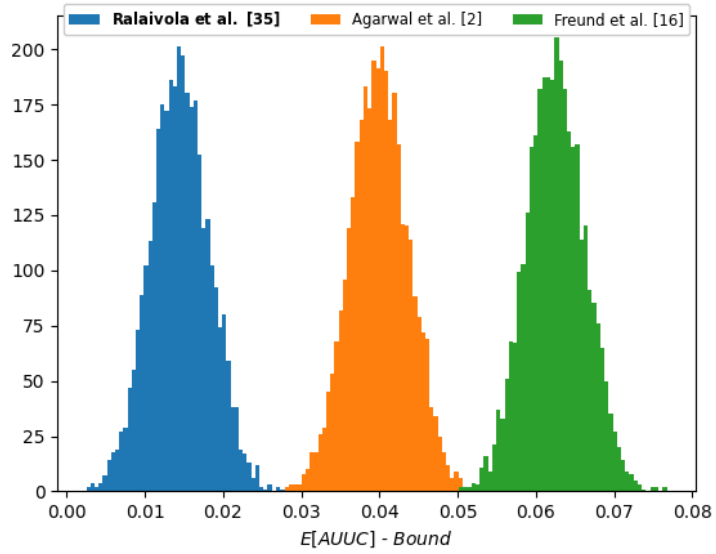

Figure 4: **AUUC bound tightness depending on inner bipartite ranking risk bounding technique** (closer to 0 is better).

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 Fig. 6 in appendix D.

Tuning parameters by bound is efficient. Following [3] we compare our method applied to linear model with hyperparameters chosen by bound (that is original AUUC-max) versus chosen by cross-validation (+CV) in Table 9 in appendix E. 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 2 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 [15, 13, 27, 23], it is difficult to observe statistically significant results on this task. Nonetheless, small increases in AUUC can lead to important gains in the application [36]. 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 2 (time is indicated relative to TM).

Additionally, Figure 5 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 appendix F.

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 3 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.

5 Conclusion and future works

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.

Table 2: *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 (Eq. 1)	.03240	.02860 ± .00326	46	1.00x
CVT (Eq. 2)	.03171	.02752 ± .00324	23	0.53x
SVM-DP [27]	.03273	.02957 ± .00321	23	0.02x †
DDR [8]	.03218	.02842 ± .00325	47	1.10x
SDR [8]	.03299	**.02958** ± .00327	67	2.44x
TARNet [42]	.03292	.02863 ± .00325	34,882	11.60x
GANITE [51]	.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

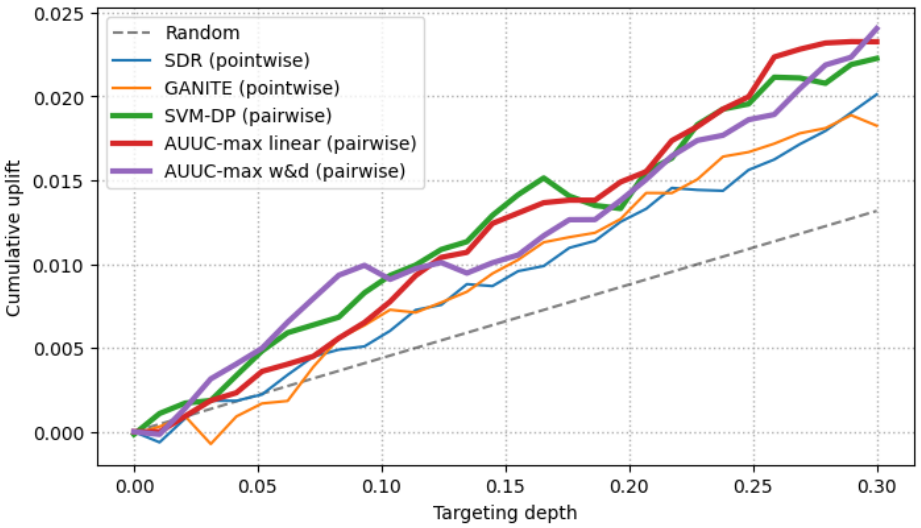

Figure 5: Uplift curves for the first 30% of population on Hillstrom. (higher is better)

Table 3: *Criteo UPLIFT v2*: comparison of baselines and AUUC-max. Top-2 results are in bold.

Model	Train AUUC	Test AUUC
TM (Eq. 1)	.00925	.00922 ± .00001
SVM-DP [27]	.00928	**.00925** ± .00002
DDR [8]	.00925	.00920 ± .00001
SDR [8]	.00926	.00923 ± .00001
AUUC-max (linear, s_{poly})	.00925	.00921 ± .00001
AUUC-max (deep, s_{log})	.00927	**.00924** ± .00001

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 2 with the novel techniques of bounding $\mathfrak{R}_{S_g}(\mathcal{F}_r)$ for deep networks [6]. 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 .

References

[1] 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.

[2] 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.

[3] 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.

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

[5] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*, 2019.

[6] 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.

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

[8] 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.

[9] Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

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

[11] 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.

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

[13] Floris Devriendt, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. *arXiv preprint arXiv:2002.05897*, 2020.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Leo Guelman, Montserrat Guillén, and Ana María Pérez Marín. Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study. *UB Riskcenter Working Paper Series, 2014/06*, 2014.

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

[20] 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.

[21] Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

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

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

[24] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[25] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077, 2020.

[26] 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.

[27] 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.

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

[29] 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.

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

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

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

[33] 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.

[34] 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.

[35] Nicholas J Radcliffe and Patrick D Surry. Differential response analysis: Modeling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI. Edinburgh, Scotland*, 1999.

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

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

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

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

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

[41] Jasjeet S Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32, 2008.

[42] 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.

[43] Szymon Sołtys Michałand Jaroszewicz and Piotr Rzepakowski. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, 29(6):1531–1559, 2015.

[44] 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.

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

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

A Bounding variance of f

Let us remind function f from Proposition 2:

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

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:

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

□

B 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 [1]. 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 [11] 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-UPLIFT v2 respectively, grids of the hyperparameters for the datasets are provided in Tables 4,5,6 and Tables 7,8 respectively.

Table 4: 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]

Generalization problem with AUUC proxies experiment (Fig. 3). 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 (Fig. 4). 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 [30] 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 polynomial 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 [50]. We report the best performing surrogates in Tables 2 and 3.

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: 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 6: 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 7: 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]

C 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 [33] which is based on empirical Bernstein bound [30], then we constructed a union bound similarly to the our main result in Th. 1. With

Table 8: 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]

probability at least $(1-\delta)$:

$$
\begin{aligned}
AUUC(f) \leq \widehat{AUUC}(f, S^T_{test}, S^C_{test}) \\
+ \lambda_T \left(\sqrt{\frac{4\hat{\Sigma}^2(S^T_{test}) \log \frac{8}{\delta}}{n_T}} + \frac{10}{n_T} \log \frac{8}{\delta} \right) \\
+ \lambda_C \left(\sqrt{\frac{4\hat{\Sigma}^2(S^C_{test}) \log \frac{8}{\delta}}{n_C}} + \frac{10}{n_C} \log \frac{8}{\delta} \right),
\end{aligned}
$$

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

D Influence of Λ

E Effectiveness of AUUC-max for hyperparameters tuning

F Uplift curves on Hillstrom

G Comparison of AUUC-max with PCG

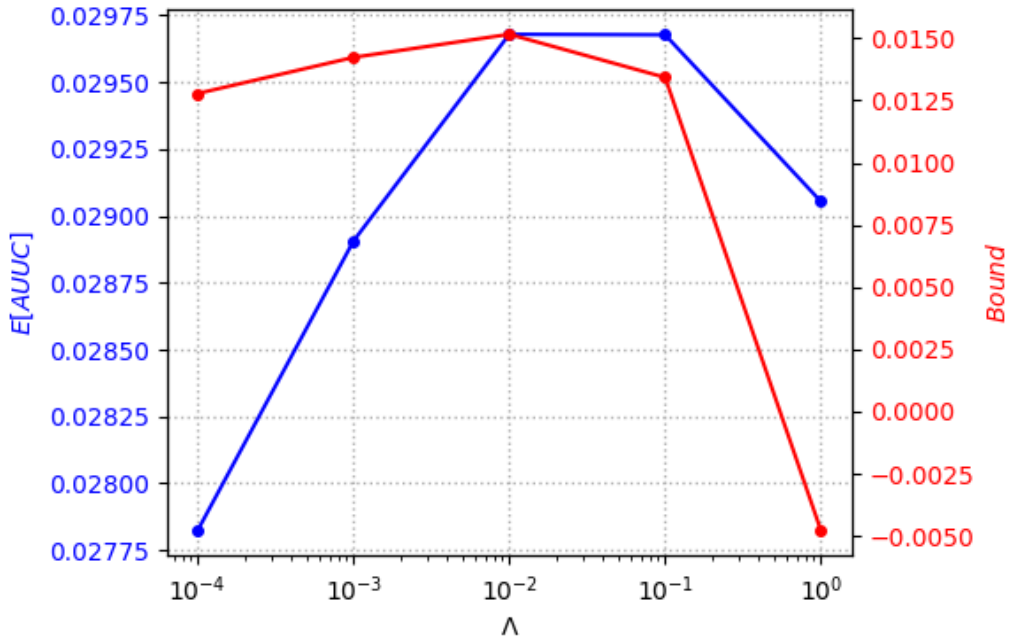

Figure 6: Influence of Λ on bound tightness and AUUC-max model performance.

Table 9: *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

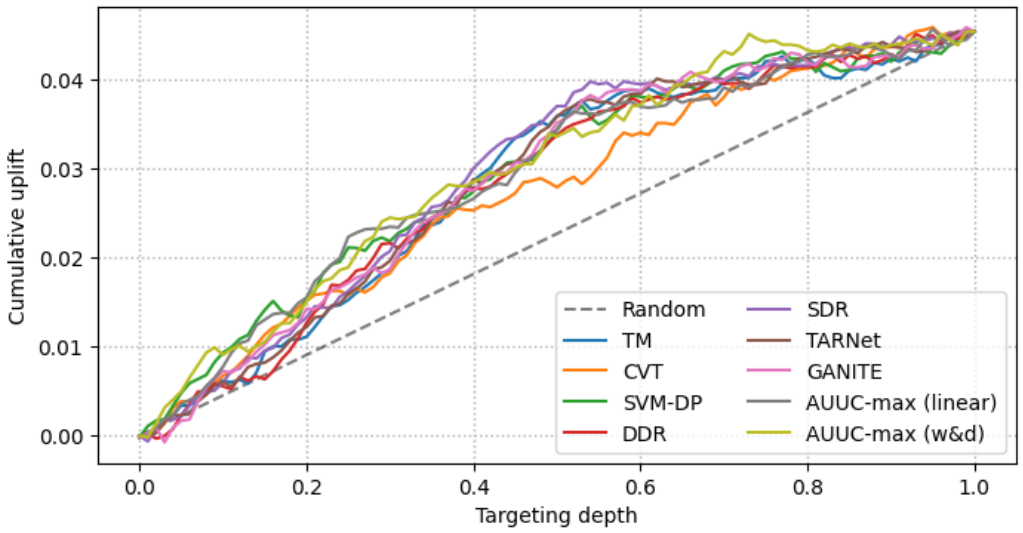

Figure 7: Uplift curves on Hillstrom. (higher is better)

Table 10: *Hillstrom*: comparison of AUUC-max with PCG. Result of PCG is taken from [13], Table 11.

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