OPTIMIZATION OF TREATMENT ASSIGNMENT WITH GENERALIZATION GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the task of optimizing treatment assignment based on predictions of the individual treatment effect (ITE), as found in applications such as personalized medicine or targeted advertising. We argue that traditional approaches do not provide rigorous guarantees for model selection, jeopardising expected gains of targeting treatment. For the first time we overcome this problem by maximizing directly future expected gains. Our proposal entails a ranking formulation of the Area Under the Uplift Curve ($AUUC$) for which we provide derivable surrogates and data-dependent generalization bounds based on local Rademacher complexity. Through careful experimental evaluation on real datasets we empirically demonstrate the tightness of our bounds and show their effectiveness for model selection.

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. Similarly in performance marketing one would prefer to target advertisement budget towards potential customers that would be more likely to be persuadable to purchase. We formalize this setting as a problem of optimizing treatment assignment and illustrate it in Figure 1. We focus on the case where data are available from prior experiments: it could be a pilot study using a randomized control trial with placebo for medicine or an A/B test for marketing (step 1). Such experiments are usually used to estimate the Average Treatment Effect (ATE) of treatment T on outcome Y: $ATE = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$. Moreover it is possible to learn an Individual Treatment Effect (ITE) predictor $ITE(x) = \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x\ ,\ T = 0]$ when covariates X are observed (Radcliffe (2007); Jaskowski & Jaroszewicz (2012)) (step 2). Such models are also known as *uplift models* in marketing literature and especially useful when treatment effect is heterogeneous. Practitioners use the predicted ITE to *rank* future instances and target treatment to the ones with the highest scores (step 3) (Devriendt et al. (2018)).

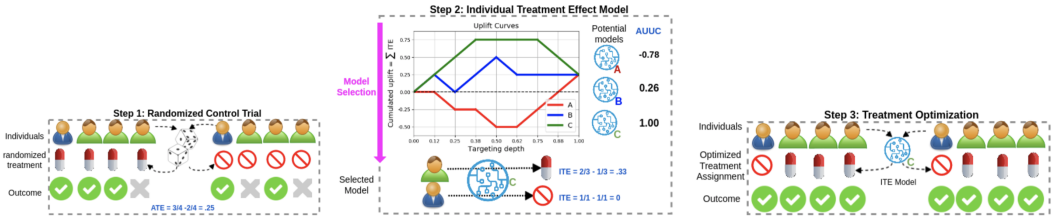

Figure 1: **The task of optimizing treatment assignment**. Step 1 starts with a randomized control trial allowing to estimate ATE. Then Step 2 consists in learning and evaluating several ITE models and selecting the best performing one by $AUUC$ on data gathered at Step 1. Finally at Step 3, the best model is used to target treatment on the next cohort of individuals. Our main contribution is to provide guarantees for selecting the best ITE model.

Model selection is a crucial step in the process as the quality of the model will be strongly influencing the gain of targeting treatment. Indeed, when a new cohort of individuals is available, the predictions of the model will be used to target treatment: highest scored individuals would get treatment (green

individuals in Figure 1) whilst the lowest scored ones would be excluded from treatment (blue individuals). The metric of choice to value the quality of a model is the Area Under the Uplift Curve ($AUUC$) (Rzepakowski & Jaroszewicz (2010)). This metric measures the cumulative uplift along individuals sorted by predicted ITE. A good model (with a high $AUUC$) scores higher those individuals for which the ITE is strong compared to ones for which the ITE is low.

We highlight a discrepancy between the documented practice of model selection by $AUUC$ and the generalization guarantees of traditional ITE models. Indeed, ITE models (as reviewed in (Gutierrez & Gérardy (2017))) optimize proxies of the $AUUC$, typically an outcome prediction accuracy. Now the Empirical Risk Minimization (ERM) principle provides generalization guarantees for the performance of models on unseen data *but only for the loss* that is optimized, which is a proxy and not $AUUC$ itself. Hence there is a risk that selecting models by $AUUC$, whereas they optimize another loss, might lead to weak or negative gains when targeting treatments on future instances of the problem. Moreover, practitioners typically select models or hyperparameters by cross-validation (Devriendt et al. (2018)), leading to costly procedure where models need to be learned and evaluated many times.

Considering the possible errors in model selection and inefficiencies of traditional ITE model formulations our proposition is to study generalization bounds for $AUUC$, from which we derive a learning objective that can be safely and efficiently used for model selection. Our main contributions are summarized as follows.

1. We propose the first generalization bounds for $AUUC$ using data-dependent concentration inequalities on dependent variables (Section 3).
2. We propose the first ITE model guaranteed to generalize for treatment optimization by deriving a surrogate of the ranking formulation of $AUUC$ (Section 4).
3. We perform a thorough empirical evaluation (Section 5) covering: i) tightness of different variants of the bounds, ii) its usefulness for model selection, iii) choice of different surrogates of the $AUUC$ loss, iv) performance on real datasets.

2 Related Works

In this section we will review existing ITE prediction approaches, evaluation metrics, and pitfalls of ITE prediction.

2.1 Evaluation metrics

At first glance one could question the usage of specific metrics for evaluation ITE models. After all, if we could observe the outcome of a given individual in both the treated and untreated case we could use a conventional metric such as mean squared error (MSE). Such a metric is formalized in (Shalit et al. (2017)) as a point-wise Precision when Estimating Heterogeneous Effect ($PEHE$): $\epsilon_{PEHE} = \int_{\chi}(\hat{u}(x) - u(x))^2 p(x)dx$. In practice it has undesirable properties: i) χ could be arbitrarily sparse, leading to estimating conditional expectations on very few data points or even ii) impossible if a given x is observed only in one of the control or treatment conditions.

At the same time one can estimate group-level treatment effect as $\hat{ATE}$ over the group of individuals. This idea underlies the Area Under the Uplift Curve ($AUUC$) (Rzepakowski & Jaroszewicz (2010)), which is popular method for evaluating ITE models in the literature. This metric is an extension of the Area Under the Lift Curve (AUL) (Tufféry (2011)). Each point on uplift curve corresponding to ITE model p is the difference in mean outcome rate (or lift) produced by model p of groups T (treatment, $T = 1$) and C (control, $T = 0$), at a particular threshold percentage of all examples. Such a curve is shown in Figure 1. (Kuusisto et al. (2014)) have defined $AUUC$ as a difference of AULs for treatment and control groups.

In this work we normalize $AUUC$ as in (Surry & Radcliffe (2011))[1], namely i) we subtract $AUUC$ of the *random* model and ii) we divide resulting area by area between $AUUC$ of the *ideal* model

[1] authors call it "Qini coefficient" by analogy to Gini coefficient

and of the random model, giving the following formula:

$$AUUC(p) = \frac{2\bar{y}_T AUL_T(p) - \bar{y}_T - 2\bar{y}_C AUL_C(p) + \bar{y}_C}{\bar{y}_T(1-\bar{y}_T) + \bar{y}_C(1-\bar{y}_C)} \tag{1}$$

where $\bar{y}_T, \bar{y}_C$ are average outcome rates of groups T and C respectively. Subtraction of $AUUC$ of the *random* model allows to disentangle what is the gain of using ITE model p compared to a random model. Also dividing by ideal model performance makes $AUUC$ scores more comparable between datasets.

2.2 ITE PREDICTION METHODS

The **Two Models (TM)** approach (Hansotia & Rukstales (2002)) is probably the most trivial method to predict ITE. It uses two separate probabilistic models, $P_T(Y = 1|X)$ for group T and $P_C(Y = 1|X)$ for group C. ITE then can be computed as:

$$\hat{p}^{TM}(x) = \hat{P}_T(Y = 1|X = x) - \hat{P}_C(Y = 1|X = x) \tag{2}$$

For this method, any prediction model with its outputs interpreted as probabilities can be used (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 (Radcliffe & Surry (2011)) for a detailed critic). This remark makes a general argument for using methods that combine knowledge of both parts of the dataset.

In an attempt to overcome this problem, (Betlei et al. (2018)) proposed the approach **Shared Data Representation (SDR)** and showed that a multi-task approach empirically performs well when the treatment is imbalanced. Another attempt in this direction by the same authors is **Dependent Data Representation (DDR)** which can be seen as a classifier cascade transferring knowledge from one treatment group to another.

(Jaskowski & Jaroszewicz (2012)) proposed the **Class Variable Transformation (CVT)** technique that combines binary treatment and outcome in order to use a single classification model. For this purpose a new label :

$$Z = YT + (1 - T)(1 - Y), \tag{3}$$

is defined which gives a predictor of the form

$$\hat{p}^{CVT}(x) = 2\hat{P}(Z = 1|X = x) - 1. \tag{4}$$

Another productive line of research has been the adaptation of split criteria of **Decision Trees** (Radcliffe & Surry (2011); Rzepakowski & Jaroszewicz (2012); Sołtys Michałand Jaroszewicz & Rzepakowski (2015)) for ITE prediction.

2.3 METRIC MAXIMIZATION AND GENERALIZATION BOUNDS

Overall, even though the common methods described in previous section were shown to empirically perform well, they still miss optimization towards the optimal ordering. We now focus on two previous works that are more directly related to maximizing ranking performance metrics for ITE and that provide some generalization guarantees.

SVM for Differential Prediction (Kuusisto et al. (2014)) is probably the most similar work to our approach. Essentially authors propose to maximize $AUUC$ directly by expressing it as a weighted sum of two Areas Under ROC Curve (AUC) and maximizing it using a suitable Support Vector Machine (SVM) objective. We build upon their seminal work by deriving a generalization bound for $AUUC$ with a similar $AUUC$ decomposition. Moreover we explain how to optimize differentiable surrogates of the bound for a very large class of models including deep neural nets. We also experiment on much larger, real-world datasets than in the original study.

Another related work that proposes generalization bounds is (Shalit et al. (2017)). At a detailed level their study is sensibly different in that i) it tackles the observational case (and assumes no unobserved confounders) and ii) bounds $PEHE$ (an MSE on the individual treatment effect - see Section 2.1) which is quite different to $AUUC$, that is by essence a ranking metric. Another difference is that we propose data-dependent bounds that take into account the specifics of evaluation datasets and might thus be more informative about the difficulty of the practical problem (see also Section 4.2 where we explain how to use our generalization bound for model selection).

3 ANALYZES AND GENERALIZATION BOUNDS

Overall, our plan is to bound the difference between $AUUC$ and its expectation and use it to propose corresponding learning objective. For that purpose we start by drawing a connection between $AUUC$ and AUC (Section 3.1) and by means of Rademacher concentration inequalities (Section 3.3) build a bound. Then we explain how to estimate it in practice (Section 3.4). Finally we define a principled optimization method with generalization guarantees for $AUUC$ (Section 4).

3.1 CONNECTION BETWEEN AUUC AND AUC

Before beginning our analyzes, we suppose that labels in the group C are replaced by $(1-T)(1-Y)$ as in (Eq. 3) (this will avoid a minimax optimization later). Let S^T denote the subset of the training set S related to observations having been given a treatment and S^C denote the control group with the reverted labels. i.e. $S = S^T \sqcup S^C$. Let also $\bar{y}_T$, $\bar{y}_C$ be the average outcome rates of groups T and C (with changed labels) respectively. Furthermore, let $\lambda_T = \bar{y}_T(1-\bar{y}_T), \lambda_C = \bar{y}_C(1-\bar{y}_C)$ be the variance of outcome as a Bernoulli random variable in treatment and reverted label control respectively.

Proposition 1 *$AUUC$ is related to ranking loss (Eq. 6) as:*

$$AUUC(f, S^T, S^C) = 1 - (\alpha \hat{R}(f, S^T) + \beta \hat{R}(f, S^C)), \tag{5}$$

where

$$\hat{R}(f, S^g) \triangleq \frac{1}{n_+^g n_-^g} \sum_{(\mathbf{x}_i, +1) \in S^g} \sum_{(\mathbf{x}_j, 0) \in S^g} \mathbb{1}_{f(\mathbf{x}_i) < f(\mathbf{x}_j)} \tag{6}$$

is the empirical bipartite ranking risk, $g \in \{T, C\}$, $\alpha = \frac{2\lambda_T}{\lambda_T + \lambda_C}, \beta = \frac{2\lambda_C}{\lambda_T + \lambda_C}$.

The proof is based on the derivation of the expression of $AUL(f, S^g)$ from $AUC(f, S^g)$ (Tufféry (2011)); and the equality $AUC(f, S^g) = 1 - \hat{R}(f, S^g)$.

3.2 LEARNING OBJECTIVE

Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ be the set of real-valued functions, we suppose that observations of the groups T and C are identically and independently distributed according to some distribution $\mathcal{D}^g; g \in \{T, C\}$. From (Eq. 5), the learning objective is hence to find $f \in \mathcal{F}$ in such a way that

$$AUUC(f) = \mathbb{E}_{S^T, S^C}\left[AUUC(f, S^T, S^C)\right] = 1 - (\alpha \mathbb{E}_{S^T}[\hat{R}(f, S^T)] + \beta \mathbb{E}_{S^C}[\hat{R}(f, S^C)]) \tag{7}$$

is as large as possible. From this expression, the problem then casts into controlling

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}_+^g, \mathbf{x}' \sim \mathcal{D}_-^g}(f(\mathbf{x}) < f(\mathbf{x}')), \tag{8}$$

in both groups $g \in \{T, C\}$; $\mathcal{D}_+^g$ (resp. $\mathcal{D}_-^g$) is the conditional distribution of the preferred or positive (resp. non-preferred or negative) examples of the group S^g.

3.3 RADEMACHER GENERALIZATION BOUNDS

Let us now consider the minimization problems of the pairwise ranking losses over the treatment and the control subsets, and the following dyadic transformation defined over each of the groups S^T and S^C :

$$\mathcal{T}(S^g) = \left(\begin{cases} (\mathbf{z}_j = (\phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'})), & \tilde{y}_j = +1 \quad \text{if } y_i = +1 \text{ and } y_{i'} = 0 \\ (\mathbf{z}_j = (\phi(\mathbf{x}_{i'}), \phi(\mathbf{x}_i)), & \tilde{y}_j = -1 \quad \text{elsewhere} \end{cases} \right)_{j=(i'-1)n_-^g + i}, \tag{9}$$

where $g \in \{T, C\}$; $\phi(\mathbf{x}) \in \mathbb{R}^p$ is the feature representation associated to observation $\mathbf{x}$, and n_-^g (resp. n_+^g) is the number of negative (resp. positive) observations in S^g and $[J] = \{1, \ldots, J\}$ denotes the set of first J integers. With the class of functions

$$\mathcal{H} = \{h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}; (\phi(\mathbf{x}^y), \phi(\mathbf{x}^{y'})) \mapsto f(\phi(\mathbf{x}^y)) - f(\phi(\mathbf{x}^{y'})), f \in \mathcal{F}\}, \tag{10}$$

the empirical loss (Eq. 6) can be rewritten as :

$$\hat{\mathcal{L}}(\mathcal{T}(S), h) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{\tilde{y}_j h(\mathbf{z}_j)\leq 0}. \tag{11}$$

We now state our main result which is a lower bound of the $AUUC$, stated in Theorem 1.

Theorem 1 *Let $S = (\mathbf{x}_i, y_i)_{i=1}^m \in (\mathcal{X}\times\mathcal{Y})^m$ 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 changed label control S^C subsets. Let $\mathcal{T}(S^T)$ and $\mathcal{T}(S^C)$ be the corresponding transformed set. 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 $f \in \mathcal{F}_r$:*

$$AUUC(f) \geq 1 - \underbrace{\left(\alpha\hat{R}_\ell(f, S^T) + \beta\hat{R}_\ell(f, S^C)\right)}_{empirical\ term} - \underbrace{\mathfrak{C}_\delta(\mathfrak{R}_T(\mathcal{F}_r), \mathfrak{R}_C(\mathcal{F}_r))}_{complexity\ term} - \underbrace{\frac{25}{48}\left(\frac{\alpha}{n_+^T} + \frac{\beta}{n_+^C}\right)\log\frac{2}{\delta}}_{data\ term},$$

$$\mathfrak{C}_\delta(\mathfrak{R}_T(\mathcal{F}_r), \mathfrak{R}_C(\mathcal{F}_r)) = (\alpha\mathfrak{R}_T(\mathcal{F}_r) + \beta\mathfrak{R}_C(\mathcal{F}_r)) + \left(\frac{\frac{5}{2}\sqrt{\mathfrak{R}_T(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_+^T}}\alpha + \frac{\frac{5}{2}\sqrt{\mathfrak{R}_C(\mathcal{F}_r)} + \frac{5}{4}\sqrt{2r}}{\sqrt{n_+^C}}\beta\right)\sqrt{\log\frac{2}{\delta}}$$

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

The proof is based on the generalization upper bounds of the ranking losses, $\hat{R}_\ell(f, S^T)$ and $\hat{R}_\ell(f, S^C)$, proposed in (Ralaivola & Amini (2015)). The result is then deduced from the union bound after finding the optimal constants that appear in the infimums of these generalization bounds.

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

3.4 Computation of the bound for $\mathfrak{R}_S(\mathcal{F}_r)$

As one can see, our lower bound for $AUUC$ consists of three main terms, namely empirical, complexity and data terms. To use it in practice we first need to clarify how to estimate local fractional Rademacher complexities $\mathfrak{R}_T(\mathcal{F}_r)$ and $\mathfrak{R}_C(\mathcal{F}_r)$ that appear in $\mathfrak{C}_\delta(\mathfrak{R}_T(\mathcal{F}_r), \mathfrak{R}_C(\mathcal{F}_r))$. A solution is to upper bound $\mathfrak{R}_S(\mathcal{F}_r)$ by a term that includes its empirical counterpart. This allows to directly estimate a lower bound for $AUUC$ from training data.

Proposition 2 *Let $S \subseteq \{x : \|\mathbf{x}\| \leq R\}$ be a sample of size N with n^+ positive labels and let $\mathcal{F}_r = \{\mathbf{x} \mapsto \mathbf{w}\cdot\mathbf{x} : \|\mathbf{w}\| \leq \Lambda; f \in \mathcal{F} : \mathbb{V}f \leq r\}$, be the class of linear functions with bounded variance and bounded norm over the weights. Then, the local fractional empirical Rademacher complexity of $\mathcal{F}_r$ can be bounded with probability $1 - \frac{\delta}{2}$ by:*

$$\mathfrak{R}_S(\mathcal{F}_r) \leq \sqrt{\frac{R^2\Lambda^2}{n_+}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n_+}} \tag{12}$$

4 Proposed approach

We propose to extend the class variable transformation approach due to its convenient properties. Firstly, under reverting label in control group we manage to avoid a minimax optimization problem of maximizing weighted difference of $AUC(f, S^T)$ and $AUC(f, S^C)$ (see proof of Proposition 1 for details) and use instead Expression 5. Secondly, according to (Eq. 4), ranking of data points by their ITE score is equivalent to ranking them by probability predictions of the model.

4.1 AUUC Maximization

We formulate an optimization problem for the empirical value of $AUUC$ as follows:

$$\arg\max_\theta AUUC \equiv \arg\min_\theta \left(\alpha\hat{R}(f_\theta, S^T) + \beta\hat{R}(f_\theta, S^C)\right), \tag{13}$$

Both terms $\hat{R}(f_\theta, S^T)$ and $\hat{R}(f_\theta, S^C)$ in equation 13 are non-differentiable functions as they are defined over the instantaneous ranking loss $\mathbb{1}_{f(\mathbf{x}_i)<f(\mathbf{x}_j)}$. However we can use differentiable surrogates of the latter (Yan et al. (2003)), such as $s_{log}(\mathbf{x}_i, \mathbf{x}_j) = \ln\left(1 + e^{-(\mathbf{x}_i - \mathbf{x}_j)}\right) / \ln(2)$, $s_{sigmoid}(\mathbf{x}_i, \mathbf{x}_j) = 1/(1 + e^{-k(\mathbf{x}_i - \mathbf{x}_j)})$, $s_{poly}(\mathbf{x}_i, \mathbf{x}_j) = (-(\mathbf{x}_i - \mathbf{x}_j - \mu))^p \, \mathbb{1}_{\mathbf{x}_i - \mathbf{x}_j < \mu}$.

Remark that s_{log} and s_{poly} with a proper choice of hyperparameters e.g. $(\mu = 1, p = 3)$ upper bound the indicator function.

Optimization problem equation 13 could be rewritten as:

$$\arg\max_\theta AUUC \equiv \arg\min_\theta \left(\alpha \hat{R}_s(f_\theta, S^T) + \beta \hat{R}_s(f_\theta, S^C)\right), s \in \{s_{log}, s_{sigmoid}, s_{poly}\} \quad (14)$$

which is now differentiable. We called an algorithm solving optimization problem equation 14 as "$AUUC$-max".

4.2 Model Selection using Rademacher bound

Selecting models or hyperparameters by their lower bound, as estimated on the training set at learning time, is guaranteed to generalize to unseen data when the final metric of choice is $AUUC$. Practically speaking this means one could avoid using internal cross-validation on the training set and save large amounts of computation. We study empirically this property in Section 5.

5 Experimental setup and results

We conducted a number of experiments aimed at evaluating how the proposed bound on the $AUUC$ can help to learn an efficient ITE model. To this end, we first present an empirical evidence on the tightness of our bound compared to other generalization bounds proposed for ranking (Section 5.2), as well as its usefulness in model selection processes (Section 5.3). Finally we compare performance of the proposed method with the other ITE prediction approaches (Section 5.4). Technically we implemented all methods and surrogate losses in Keras framework (Chollet et al. (2015)). For all models we have batch size of 1000, run 200 epochs of learning with early stopping by loss on validation (with patience of 30 epochs) and use Adam optimizer with step decay to update the learning rate.

5.1 Benchmark

Our benchmark consists of two open source, real-life datasets that happen to pertain both to the digital marketing application. **Hillstrom Email Marketing data** (Hillstrom (2008)) contains results of an e-mail campaign for an Internet based retailer. **Criteo-UPLIFT2** (Diemert Eustache, Betlei Artem et al. (2018)) is a large scale dataset constructed from incrementality A/B tests, a particular protocol where a random part of the population is prevented from being targeted by advertising. For the speed of experiments we pick a random subsample of size 1M.

5.2 Tightness of the proposed bound

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 $AUUC$s. The true $AUUC$ is estimated from the upper bound of an Empirical Bernstein inequality (Maurer & Pontil (2009)) on the test sets obtained from 30,000 random train/test splits, giving a precision greater or equal than $.001$ (with probability $> .99$); this is sufficient to confidently compare it to the proposed bound estimated on the training data coming from the splits. In Figure 2, we observe that the difference between the estimated bound for $\mathbb{E}[AUUC]$ and our Rademacher bound is close to 0.1, which is quite tight considering that $AUUC \in [-1; 1]$. For the sake of illustrating the tightness of this bound, we have computed other lower-bounds using different AUC bounds proposed in the literature. In Figure 2, we call our bound as "rademacher" vs variants using Corollary 18 and Theorem 27 from (Agarwal et al. (2005)) as "agarwal" and "freund" respectively (last one is a generalization of the result proposed in (Freund et al. (2003))). This results suggest that the lower-bounds obtained with this

strategy are at least 2 times looser, and illustrate the benefit of data-dependent approach as the one we propose in this work.

5.3 USEFULNESS OF THE BOUND FOR MODEL SELECTION

To check the ability of selecting effective ITE models using the proposed Rademacher lower-bound of (Eq. 12), we follow the protocol of (Langford & Shawe-Taylor (2002)). According to the assumptions of Proposition 2, we consider the class of linear functions with bounded variance and bounded norm on their weights, and select the hyperparameters of the model, namely L_2 regularization term and the initial learning rate from the $[0, 1\mathrm{e}^{-6}, 1\mathrm{e}^{-4}]$ and $[5\mathrm{e}^{-4}, 1\mathrm{e}^{-3}]$ respectively (we find ranges of such potential hyperparameters experimentally). We select the best model then by either, the usual 5-fold cross-validation (stratified by treatment variable to be able to compare metric through the folds) on training set, or from the lower-bound of $AUUC$ as estimated during training. We repeated the procedure on 100 random train/test splits (also stratified by treatment) of Hillstrom data. Finally we report the mean test $AUUC$s of both model selection techniques ($AUUC^{test}_{CV}$ and $AUUC^{test}_{Bound}$ respectively) on Table 1.

Table 1: **Model selection**: generalization bound vs 5-fold cross-validation on Hillstrom dataset. Note how close the selection by bound is to the usual cross-validation technique.

	Mean Test $AUUC \pm 2$ std
Selection by Bound	0.0657 ± 0.0248
Selection by CV	0.0665 ± 0.0245

For the model selection experiments, we observe on Table 1 that both methods perform comparably, even though the variance of the $AUUC$ metric is large. The mean difference takes place at the 3rd digit, suggesting that the model selection procedure using the bound is effective. In this case, the computation savings are of the order of number of folds used in cross-validation, usually > 5 in practice. Remark that the bound on $AUUC$ is directly estimated on the whole training data and that there is no hold-out validation set as in cross-validation. For clarity, in Figure 3 we also show a distribution of the gaps between $AUUC^{test}_{CV}$ and $AUUC^{test}_{Bound}$ over data splits.

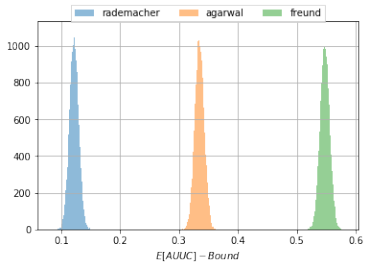

Figure 2: **Gap between $\mathbb{E}[AUUC]$ and different versions of the bound** (closer to 0 is better) on Hillstrom dataset. Note the tightness of the Rademacher version compared to (Agarwal et al. (2005)).

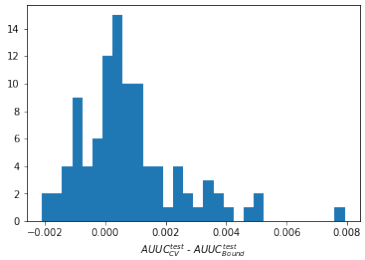

Figure 3: **Model Selection**: distribution of performance gaps between $AUUC$s of models selected by cross-validation vs by proposed bound on Hillstrom dataset (negative values are in favour of bound method and vice versa). Note suboptimality is $< 1\mathrm{e}^{-3}$ in most cases.

5.4 COMPARISON OF THE METHODS PERFORMANCE

For comparing the performance of the methods, we use a linear model and a multi-layer perceptron with 2 layers and 64 units for each layer with ReLU activations as the base classifiers. As baselines, we consider TM and CVT which are the most popular ITE models proposed in the literature. We keep the same range of potential regularization terms for all methods as $[0, 1\mathrm{e}^{-6}, 1\mathrm{e}^{-4}]$. For TM and CVT, the ranges of initial learning rates are $[1\mathrm{e}^{-1}, 5\mathrm{e}^{-1}]$ and $[5\mathrm{e}^{-3}, 1\mathrm{e}^{-2}]$ respectively (also found experimentally), for $AUUC$-max this range is the same as in Section 5.3. Also, for the $AUUC$-max we evaluate s_{log} and $s_{poly}(\mu = 0.1, p = 3)$ (according to practical suggestions of

(Yan et al. (2003))) as the surrogates. To check statistical significance we apply one-sided Mann-Whitney rank test at 95% confidence level (marked in bold in the tables when positive). Tables 2 and 3 contain quantitative results of the approaches performance on Hillstrom and CU2-BD datasets datasets respectively.

Table 2: **Performance**: comparison of $AUUC$-max vs baselines on Hillstrom dataset. Note we outperform baselines significantly.

	Mean Test $AUUC \pm 2$ std	Mean Test $AUUC \pm 2$ std
Base Classifier	Logistic Regression	Multi-Layer Perceptron
Two Models	.0619 ± .0273	.0653 ± .0239
Class Variable Transformation	.0617 ± .0267	.0626 ± .0255
$AUUC$-max (CV, s_{log})	.0650 ± .0249	–
$AUUC$-max (Bound, s_{log})	.0648 ± .0248	.0655 ±.0242
$AUUC$-max (CV, s_{poly})	**.0665 ± .0245**	–
$AUUC$-max (Bound, s_{poly})	**.0657 ± .0248**	.0656 ± .0249

Table 3: **Performance**: comparison of $AUUC$-max vs baselines on CU2-BD. Note we compete with the baselines by only using model selection by bound.

	Mean Test $AUUC \pm 2$ std
Base Classifier	Logistic Regression
Two Models	.0264 ± .0204
Class Variable Transformation	.0261 ± .0182
$AUUC$-max (Bound, s_{poly})	.0251 ± .0164

As we can see on Table 2 (Hillstrom) , $AUUC$-max with s_{poly} significantly outperform both baselines, even using model selection by Rademacher lower bound. One of the explanations could be that for the $AUUC$-max there is a strong correlation between the loss function and the metric. Instead, using TM and CVT one can only minimize proxy functions which sometimes cannot result in the high $AUUC$. In order to reinforce such an argument, we draw typical validation loss and $AUUC$ (Figure 5). Plot shows a high correlation, suggesting the loss is behaving according to theory.

Besides, as an illustration of the potential of the method to be applied to other types of models we can observe that Multi-Layer Perceptrons (MLP) can be trained with good performance. To reduce computation time we didn't perform variants with cross-validation and rely on bound model selection only. We hypothetize that the chosen architecture might limit the potential of the method in this case as no significant results were observed.

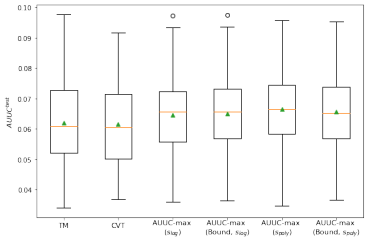

Figure 4: **Performance**: distribution of $AUUC$s on Hillstrom dataset as presented in Table 2.

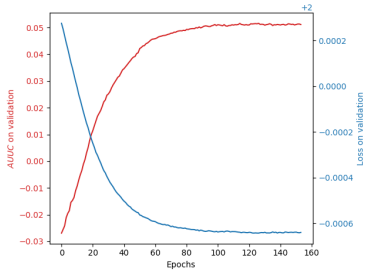

Figure 5: **Optimization**: illustration of the correlation between the optimized loss and the $AUUC$ on a test set. Note the smoothness and behavior over epochs.

We note that in Table 3 (CU2-BD) all models perform indistinguishably, perhaps due to the down-sampling we chose for reducing computation time.

6 Conclusion and Future Works

In this paper, we proposed a first data-dependent generalization lower bound for the popular ITE prediction metric, $AUUC$. We investigate tightness of the proposed bound and find that the Rademacher version is tighter than possible alternatives. Then we come up with an efficient model selection strategy that consists in estimating such a bound only on training set at learning time. We empirically show that it finds models and hyperparameters as good as those found by cross-validation. As a result we highlight its computational benefits. Further, we formulated a method to directly maximize this metric which is usable with most machine learning models, including neural networks. Experiments on two large collections show that our method compares favorably to relevant baselines from the literature.

REFERENCES

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.

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*, pp. 47–57. Springer, 2018.

François Chollet et al. Keras. `https://keras.io`, 2015.

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.

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

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.

Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*, pp. 1–13, 2017.

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

Kevin Hillstrom. The MineThatData e-mail analytics and data mining challenge, 2008.

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

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*, pp. 50–65. Springer, 2014.

John Langford and John Shawe-Taylor. Pac-bayes & margins. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pp. 423–430, 2002.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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.

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

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

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

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

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*, pp. 3076–3085. JMLR. org, 2017.

Szymon Sołtys Michałand Jaroszewicz and Piotr Rzepakowski. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, 29(6):1531–1559, 2015. ISSN 13845810. doi: 10.1007/s10618-014-0383-9.

Patrick D Surry and Nicholas J Radcliffe. Quality measures for uplift models. *submitted to KDD2011*, 2011. URL `http://www.stochasticsolutions.com/pdf/kdd2011late.pdf`.

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

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)*, pp. 848–855, 2003.