CURLS: Causal Rule Learning for Subgroups with Significant Treatment Effect

Jiehui Zhou
zhoujiehui@zju.edu.cn
State Key Lab of CAD&CG, Zhejiang University
Hangzhou, Zhejiang, China
DAMO Academy, Alibaba Group
Hangzhou, Zhejiang, China

Linxiao Yang
linxiao.ylx@alibaba-inc.com
DAMO Academy, Alibaba Group
Hangzhou, Zhejiang, China

Xingyu Liu
liu_xingyu@zju.edu.cn
State Key Lab of CAD&CG, Zhejiang University
Hangzhou, Zhejiang, China

Xinyue Gu
guxinyue.gxy@alibaba-inc.com
DAMO Academy, Alibaba Group
Hangzhou, Zhejiang, China

Liang Sun*
liang.sun@alibaba-inc.com
DAMO Academy, Alibaba Group
Hangzhou, Zhejiang, China

Wei Chen*
chenvis@zju.edu.cn
State Key Lab of CAD&CG, Zhejiang University
Hangzhou, Zhejiang, China

ABSTRACT

In causal inference, estimating heterogeneous treatment effects (HTE) is critical for identifying how different subgroups respond to interventions, with broad applications in fields such as precision medicine and personalized advertising. Although HTE estimation methods aim to improve accuracy, how to provide explicit subgroup descriptions remains unclear, hindering data interpretation and strategic intervention management. In this paper, we propose *CURLS*, a novel rule learning method leveraging HTE, which can effectively describe subgroups with significant treatment effects. Specifically, we frame causal rule learning as a discrete optimization problem, finely balancing treatment effect with variance and considering the rule interpretability. We design an iterative procedure based on the minorize-maximization algorithm and solve a submodular lower bound as an approximation for the original. Quantitative experiments and qualitative case studies verify that compared with state-of-the-art methods, *CURLS* can find subgroups where the estimated and true effects are 16.1% and 13.8% higher and the variance is 12.0% smaller, while maintaining similar or better estimation accuracy and rule interpretability. Code is available at https://osf.io/zwp2k/.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; **Rule learning**; **Optimization algorithms**.

*Wei Chen and Liang Sun are corresponding authors.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671951

KEYWORDS

causal inference, rule learning, subgroup discovery, data heterogeneity, submodular optimization

ACM Reference Format:
Jiehui Zhou, Linxiao Yang, Xingyu Liu, Xinyue Gu, Liang Sun, and Wei Chen. 2024. CURLS: Causal Rule Learning for Subgroups with Significant Treatment Effect. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671951

1 INTRODUCTION

Causal inference is a data analysis process aiming at conclusions about whether and to what extent treatments affect outcomes [46]. Data heterogeneity needs to be taken into account when estimating treatment effects, as the effect of the same treatment often varies across subgroups. Discovering those subgroups with large effects and low variance (hereinafter referred to as significant treatment effect) compared to the overall population is widely used in domains such as healthcare [48], marketing [54], and public administration [20]. For example, marketers would like to find customer groups where advertising is more effective in driving purchases. Since randomized controlled trials (RCTs), the gold standard for causal inference, are not always feasible due to cost or ethical concerns, there is a strong need to uncover subgroups with significant treatment effects from observational data.

Existing causal models that consider the data heterogeneity, such as propensity score-based methods [18], double machine learning [12], meta-learners [33], entropy balancing [25] and tree-based recursive partitioning [1, 15], can estimate the effect of the treatment on the outcome on subgroups or individuals. However, most of these methods try to reduce the confounding bias in the estimation rather than directly learning subgroups with significant effects. Researchers have also explored rule learning [14, 61] and subgroup discovery [21, 53], utilizing easy-to-understand rules to describe subgroups with intriguing patterns. Unfortunately, most rule learning methods are oriented towards correlations rather than causality, which may lead to imprecise estimations in selection bias-affected

interventions. Thus, enhancing the treatment effect interpretability in the context of data heterogeneity is still underexplored.

We combine the strong analytical power of heterogeneous treatment effect (HTE) estimation methods with the high interpretability of rule learning to facilitate the identification and interpretation of subgroups with significant treatment effects. However, two challenges must be addressed. First, the trade-off between multiple objectives and constraints is not trivial. Since treatment effect estimation is a statistical inference problem, in addition to requiring a large effect value, it is also necessary to ensure that the uncertainty of the estimates is small (as shown in Rule 1 in Fig. 1). Also, the length and overlap of the rules are important constraints in order to facilitate user interpretation. Second, recognizing useful subgroups from a large number of potential candidates is difficult. Subgroups can be obtained by combining different attributes and values, which is often exponential and requires an efficient solution.

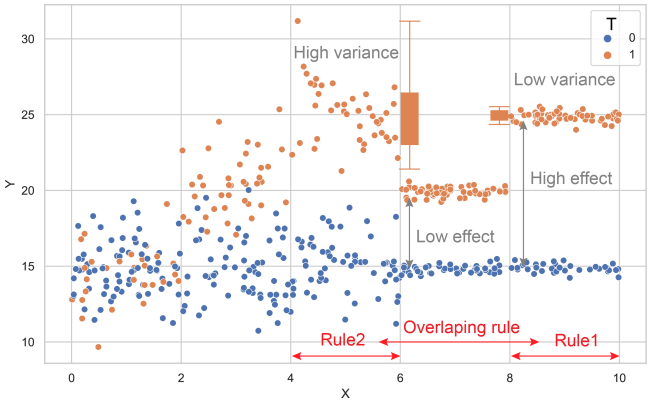

Figure 1: An illustrative toy example. There is only one covariate X, and the change in Y can be informally thought of as the effect. The subgroup corresponding to Rule1 has a high effect and low variance, which the users expect to find.

We propose *CURLS*, a causal rule learning method for identifying subgroups with significant treatment effects. To address the first challenge, we formally define causal rules, which consist of subgroups described by conjunctive normal forms (CNFs) and the corresponding effects estimated by inverse probability weighting (IPW) [28]. Then, mining subgroups with large treatment effects and low variance can be modeled as a discrete optimization problem. For the second challenge, we prove the existence of an approximate submodular lower bound for the optimization objective and design a solution based on the minorize-maximization (MM) algorithm and submodular optimization. Comprehensive quantitative experiments and qualitative case studies demonstrate the effectiveness of *CURLS*. In summary, our contributions are as follows:

- We pioneer the incorporation of rule learning into causal inference, aiming to delineate subgroups with significant treatment effects through rule-based descriptions. Specifically, we formulate this as an optimization problem, considering the trade-off between effects and variance and rule set size and overlap constraints.
- We propose an efficient optimization algorithm that iteratively maximizes the submodular lower bound of the original problem, which cuts down the original exponential search space.
- We conduct both quantitative and qualitative experiments, demonstrating that *CURLS* delivers not only extra rule-based interpretative capabilities for subgroups, but also enhances the precision of effect estimation with a smaller variance. Our method outperforms state-of-the-art algorithms in estimated and true effect strength (CATE) by 16.1% and 13.8%, respectively, and reduces variance by 12.0%.

2 RELATED WORK

In this section, we first review the algorithms related to HTE, then discuss the progress of rule learning, and finally summarize the work of subgroup discovery.

2.1 Heterogeneous Treatment Effect Estimation

Causal inference identifies the effect of treatment on outcome. However, treatment effects are often not "one-size-fits-all"—they may vary across the population. Current HTE research is divided into conditional average treatment effect (CATE) and individual average treatment effect (ITE) by population level. Reviews [24, 62] provide detailed analyses on treatment effect estimation.

CATE examines treatment effects on specific subgroups of the population, conditional on similar covariates, such as certain demographic characteristics. Tree-based methods [1, 2, 55] are widely used by dividing the covariate space into subspaces to maximize the treatment effects heterogeneity. For example, Causal Tree [1] uses part of the data to construct the tree and another part to estimate the treatment effect in each subspace, avoiding overfitting by cross-validation. To make the estimation more robust and smooth, Wager *et al.* [55] proposed Causal Forest, which aggregates the results of causal tree ensembles. The advantage of the tree model is its interpretability, which naturally provides subgroups of heterogeneous CATEs defined by root-to-leaf node paths.

ITE measures the difference in outcomes for individuals with or without receiving the treatment. Since only one outcome can be observed in the actual scenario, another potential outcome needs to be estimated. Depending on whether the treatment and control groups are estimated separately, existing methods can be categorized as single-model-based and multi-model-based. The former fits treatment effects with regression models. For example, Hill *et al.* [27] use Bayesian additive regression trees to fit the outcome surface. The latter fits the treated and control groups separately and can achieve better performance when the difference between the outcomes of the two groups is significant. The base model can use off-the-shelf estimators, such as linear regression [11] or neural networks [30]. Although these models can be accurate in estimating effects with carefully tuned parameters, they are generally uninterpretable.

Previous work has focused on how to estimate effects more accurately, *i.e.*, to exclude confounding bias in the observational data. Instead, we aim to mine subgroups that have stronger effects with small variances. To this end, we utilize a propensity-score-based effect estimation method in our implementation and incorporate the ability of rules to characterize subgroups.

2.2 Rule Learning

Rules are simple logical structures of the form "IF P THEN Q". Since general rules are similar to the way humans think, rule learning is employed in prediction or classification scenarios that require high interpretability. The existing work can be broadly classified into pre-mining and uniform optimization approaches.

Most studies [17, 34, 44, 57, 63] adopted the two-stage paradigm consisting of rule generation (or rule pre-mining) and rule selection. First, rules are pre-mined through efficient algorithms such as decision trees and association rule mining to reduce the search space of rules significantly. In the second stage, appropriate rules are selected from the candidate rules to form an unordered rule set or an ordered rule list based on specific metrics (*e.g.*, classification accuracy). However, this separation paradigm can lead to sub-optimal results as important rules may be missed in the rule pre-mining stage, resulting in a loss of accuracy.

Recently, researchers have linked rule generation and selection in a single optimization framework for rule learning. For example, Dash *et al.* [14] formalized the rule set learning problem as an integer programming problem to balance classification accuracy and rule simplicity. Column generation (CG) algorithms were used to efficiently search for an exponential number of candidate clauses (conjunctive or disjunctive terms) without the need for heuristic rule mining. Yang *et al.* [61] approached rule set learning from the perspective of submodular optimization. They formulate the main problem as a task of selecting a subset from all possible rules, while the subproblem of searching for rules is mapped as another feature subset selection task. Since the objective function can be formulated as the difference between two submodular functions, it can be approximately solved by the optimization algorithm.

Most rule learning methods are used for classification tasks that solely examine correlations. However, correlation does not imply causation. A few researchers [6, 42, 56, 60] have attempted to mine causal rules from data. For example, CRE [6] and CRS [56] are both two-stage methods, which first generates a pool of rules using random forest, FP-Growth, *etc.*, and then select a subset among them based on some criteria, such as stability selection regularization. Li *et al.* [42] first mined the association rules from the data and then used a cohort study to test whether the association rules are causal or not. However, these methods lack a global optimization objective; therefore their results depend on the quality of the candidate rules developed in the first stage.

Unlike earlier methods, we mine causal rules from observational data from a unified optimization perspective. These rules represent those subgroups with large treatment effects and low variance.

2.3 Subgroup Discovery

Subgroup discovery (SD) is a descriptive data mining technique that identifies data subgroups with interesting patterns on specific targets. It differs slightly from rule learning that focuses on prediction/classification performance on upcoming data. A comprehensive study of SD is available in reviews [3, 26].

Data subgroups can be represented using description languages such as attribute-value pairs and different logical forms (*e.g.*, conjunctions, disjunctions, inequalities, fuzzy logic, *etc.*). Subgroup interestingness can be measured using binary, nominal, or numerical targets. Certain post-processing methods have been applied to select diverse and less redundant subgroups. Due to the enormous number of potential subgroups, different search strategies, such as exhaustive and heuristic search, have been applied.

The exhaustive method [5, 21, 22, 58] searches all feasible subgroups. The naive exhaustive search may be time-consuming because the viable subgroup is exponential. Examples of strategies for reducing hypothesis space include optimistic estimate pruning, generalization-aware pruning, minimum support. SD-Map [5] is a typical exhaustive SD method that extends the popular Frequent Pattern (FP) Growth-based association rule mining method, utilizing depth-first search to generate candidates. Piatetsky-Shapiro, unusualness, and binomial tests are utilized to determine precise and significant subgroups. The SD-Map* [4] is extended for use with binary, categorical, and continuous target variables.

Further studies [16, 19, 38, 53, 64] employed efficient heuristic methods. For example, DSSD [53] is an SD algorithm based on beam search. The search usually starts with an initial solution and is then expanded to a certain number of candidate solutions. The best ones are retained for the next iteration until a stopping condition is reached. SDIGA [16] is an evolutionary fuzzy rule induction algorithm. It facilitates the discovery of more general rules by allowing variables to take multiple values. Subgroups can be evaluated in terms of confidence, support, and unusualness.

SD is useful in many fields. For example, in medicine, it helps to discover high-risk groups for a certain disease [36]. During operation and maintenance, it helps troubleshoot and attribute anomalies in total KPI metrics to specific subgroups [9, 23]. In marketing, it helps to identify target customers of different brands [37].

However, conventional SD methods usually overlook the treatment effect. Thus, this paper seeks to uncover subgroups with significant treatment effects, which requires different optimization objectives and evaluation criteria from prior SD methods.

3 PRELIMINARIES

In this section, we present some preliminaries about causal inference, submodular and supermodular functions.

Causal Inference. We introduce causal inference under the potential outcome (PO) framework [49].

Unit: A unit is an individual or object under study. A medical study unit may be a patient. The subscript $_i$ denotes the i-th unit.

Treatment: A treatment is an intervention or exposure that subjects to a unit. A new medicine or therapy could be used as a treatment in a medical study. Let T indicate the treatment. $T = 1$ units are the treatment group, while $T = 0$ units are the control group. We assume one binary treatment for simplicity.

Outcomes: Outcomes are what would have happened under different treatments. Each unit has two potential outcomes: factual outcome and counterfactual outcome. Potential outcome with treatment value t is $Y(T = t)$, also abbreviated as $Y(t)$.

Covariates: Covariates are background variables that affect treatment assignment and outcome. Observational studies often control for covariates to mitigate confounding. Let $\mathbf{X}_i = (x_{i,1}, \cdots, x_{i,d})$ represent covariates.

Observational data: Observational data refers to data collected without the researcher manipulating the environment or the subjects being studied. It differs from RCTs, which randomly assign treatment to each unit. The observational data containing n units is denoted by $\mathcal{D} = \{(T_i, \mathbf{X}_i, Y_i)\}_{i=1}^{n}$.

Treatment effect refers to the impact of a treatment on an outcome. For observational data, Inverse Probability Weighting (IPW) [28] is proposed, which assigns appropriate weights $w_i = \frac{T_i}{e_i} + \frac{1-T_i}{1-e_i}$ to each unit based on propensity score e_i to balance covariates distribution in the treatment and control groups, thereby simulating RCTs. Then, the normalized weighted average of the factual outcomes for the treatment and control groups can be calculated to estimate treatment effects [29]:

$$\tau = \frac{\sum_{i;T_i=1} w_i Y_i}{\sum_{i;T_i=1} w_i} - \frac{\sum_{i;T_i=0} w_i Y_i}{\sum_{i;T_i=0} w_i}. \tag{1}$$

When the data satisfy the three causal assumptions (unconfoundedness, positivity and stable unit treatment value), it is shown that the adjustment of the scalar propensity score removes the bias due to the observed covariates [47].

Submodular and supermodular functions. A submodular function is a set function with special properties. Its domain is a family of subsets of a given set. The output value is some measure of the subset. The inputs and outputs satisfy the relationship of diminishing returns, *i.e.*, the additional benefit of adding the set to the inputs declines. The supermodular function is the opposite of the submodular function, which satisfies the increasing returns. Formally, for a set function $f : 2^{\Omega} \to \mathbb{R}$, it is submodular if:

$$\forall A \subseteq B \subseteq \Omega \text{ and } v \in \Omega \setminus B, f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B).$$

Like the convexity in continuous optimization, submodularity is a good property in discrete optimization, making it suitable for many applications, such as approximation algorithms, game theory, automatic summarization, and feature selection [32, 43].

4 PROBLEM FORMULATION

This section introduces the formalization of the causal rule learning problem. The final optimization problem is presented in Eq. (4).

Without loss of generality, we consider observational data whose covariates are binary and outcome is a positive number, *i.e.*, $\mathbf{X}_i = (x_{i,1}, \cdots, x_{i,d}) \in \{0,1\}^d, Y \in \mathbb{R}_{>0}$. Categorical variables can be binarized by one-hot encoding, and numerical variables can be converted to binary by bucketing strategy. We determine the optimal number of bins (4-20) using 5-fold cross-validation and select the best-performing parameter for the final model. Negative outcomes can be made positive by adding an offset.

Formally, given the observational data $\mathcal{D}$, we aim to learn interpretable causal rules from it. A **causal rule** $\mathcal{R} : \boldsymbol{\alpha} \Rightarrow \tau$ contains the antecedent $\boldsymbol{\alpha}$ and the consequent τ.

A **antecedent** $\boldsymbol{\alpha}$ is the condition of the rule, expressed as the conjunctive normal form (CNF) of a series of atoms $\bigwedge_{j\in\Gamma} x_j$, *e.g.*, "age > 25 AND job != teacher". Γ is the covariate indices used in the antecedent, which is a subset of the indices of all binary covariates, *i.e.*, $\Gamma \in 2^{[d]}$, where $[d] = \{1, \cdots, d\}$ and $2^{[d]}$ means the power set of $[d]$. The atom x_j is the smallest interpretable element. We create the negation of covariate $\neg x_j$ to increase the expressiveness. The mapping from a $\mathcal{R}$ to a CNF is given by $\boldsymbol{\alpha}_{\mathcal{R}}(\mathbf{X}_i) = \bigwedge_{j\in\Gamma_{\mathcal{R}}} x_{i,j}$. For brevity, we also refer to it as $\bigwedge_{j\in\mathcal{R}} x_{i,j}$. When $\boldsymbol{\alpha}_{\mathcal{R}}(\mathbf{X}_i)$ is true, the i-th unit is **covered** by the rule $\mathcal{R}$.

The **consequent** τ is the prediction result of the rule, indicating the estimated treatment effect for the data covered by the rule. Define $\mathcal{D}_{\mathcal{R}}$ to denote the covered data, $\mathcal{D}^+$ to denote the data that received the treatment ($T = 1$), and $\mathcal{D}^-$ to denote the data that did not receive the treatment ($T = 0$). Define $\mathcal{D}_{\mathcal{R}}^+ = \{i | i \in \mathcal{D}^+ \wedge \boldsymbol{\alpha}_{\mathcal{R}}(\mathbf{X}_i) = 1\}$ denotes the units in the treatment group that are covered by the rule $\mathcal{R}$, $\mathcal{D}_{\mathcal{R}}^- = \{i | i \in \mathcal{D}^- \wedge \boldsymbol{\alpha}_{\mathcal{R}}(\mathbf{X}_i) = 1\}$ denotes the units in the control group that are covered by the rule $\mathcal{R}$. Therefore, the treatment effect of rule $\mathcal{R}$ can be represented as:

$$\tau_{\mathcal{R}} = \frac{\sum_{i\in\mathcal{D}_{\mathcal{R}}^+} w_i Y_i}{\sum_{i\in\mathcal{D}_{\mathcal{R}}^+} w_i} - \frac{\sum_{i\in\mathcal{D}_{\mathcal{R}}^-} w_i Y_i}{\sum_{i\in\mathcal{D}_{\mathcal{R}}^-} w_i} = \frac{Q_1}{Q_2} - \frac{Q_3}{Q_4}, \tag{2}$$

where $Q_1 = \sum_{i\in\mathcal{D}_{\mathcal{R}}^+} w_i Y_i$, $Q_2 = \sum_{i\in\mathcal{D}_{\mathcal{R}}^+} w_i$, $Q_3 = \sum_{i\in\mathcal{D}_{\mathcal{R}}^-} w_i Y_i$, $Q_4 = \sum_{i\in\mathcal{D}_{\mathcal{R}}^-} w_i$.

The **causal rule set** $\mathcal{S} = \{\mathcal{R}_1, \cdots, \mathcal{R}_k\}$ contains multiple rules. A rule set covers a unit if $\boldsymbol{\alpha}_{\mathcal{S}}(\mathbf{X}_i) = \bigvee_{\mathcal{R}\in\mathcal{S}} \bigwedge_{j\in\mathcal{R}} x_{i,j}$ is true. If a unit is covered by more than one rule, we take the average of the effects as the estimated treatment effect for that unit. However, an interpretable ruleset should minimize rule overlap.

Since obtaining treatment effects is a statistical estimation problem, it is important to consider the uncertainty of the treatment effect, which can be measured by the outcome variance of the treatment group, defined as:

$$\sigma_{\mathcal{R}}^2 = \frac{\sum_{i\in\mathcal{D}_{\mathcal{R}}^+} w_i (Y_i - \overline{Y}_w)^2}{\sum_{i\in\mathcal{D}_{\mathcal{R}}^+} w_i}, \tag{3}$$

For the effect, we want $\frac{Q_1}{Q_2}$ to be large, $\frac{Q_3}{Q_4}$ to be small, and the variance σ is also small. To facilitate the optimization, we take a log for the ratio. Then we get the **objective function** $f(\mathcal{R}) = \log Q_1 + \log Q_4 - \log Q_2 - \log Q_3 - \lambda \log \sigma_{\mathcal{R}}^2$, which is the profit of a rule, and λ is a coefficient that adjusts for the trade-off between treatment effect and variance. Therefore, to learn the causal rule set from observational data, we consider solving the following optimization problem:

$$\begin{aligned} \max_{\mathcal{S}} \sum_{\mathcal{R}\in\mathcal{S}} f(\mathcal{R}) \\ \text{s.t.} |\mathcal{S}| \leq K \\ |\mathcal{R}| \leq L. \end{aligned} \tag{4}$$

To ensure interpretability, $|\mathcal{S}| \leq K$ restricts the number of rules in the rule set to be no more than K, and $|\mathcal{R}| \leq L$ limits the antecedent length of the rules to be no more than L.

5 ALGORITHM

In this section, we introduce the proposed algorithm *CURLS* for solving the optimization problem in Eq. (4).

As shown in Fig. 2, to construct a precise causal rule set, we employ an iterative framework (Sec. 5.1), balancing constraints such as set size and antecedent length. This methodical approach allows us to sequentially select the most fitting rule, ensuring a coherent and

optimized causal rule set. Central to our strategy is the optimization of a set function, which maps covariates to treatment effects, requiring a nuanced balance between computational feasibility and accuracy. We address this challenge by crafting an approximate submodular lower bound for the objective function, a strategic choice that simplifies the optimization process while maintaining solution quality. Leveraging the minorize-maximization (MM) procedure (Sec. 5.2) and submodular optimization(Sec. 5.3), we efficiently derive each rule's antecedents and consequents, resulting in significant and interpretable causal rules.

5.1 Causal Rule Set Learning

Directly optimizing the rule set is not a trivial problem. Typical correlation rule set learning algorithms usually adopt sequential covering paradigms [13], that is, removing data covered by previous rules and learning a new rule. However, this can easily result in overlapping rules, thus affecting the interpretability of the rule set. To solve this problem, instead of removing the covered data, we explicitly introduce a penalty for overlapping data in the iterative process, thus increasing the diversity of the rule set.

Overlap Penalty. Since the purpose of the causal rule is to cover those units that have a strongly positive outcome after receiving treatment, we can set the weighted outcome of the covered units belonging to the treatment group to a smaller value ϵ, *i.e.*,

$$Q_1(\mathcal{R}) = \sum\nolimits_{i \in \mathcal{D}_{\mathcal{R}}^+ \setminus \mathcal{D}_{\mathcal{S}}} w_i Y_i + \sum\nolimits_{i \in \mathcal{D}_{\mathcal{R}}^+ \cap \mathcal{D}_{\mathcal{S}}} \epsilon. \tag{5}$$

Therefore, if the new rule $\mathcal{R}$ searches for units that have been covered by the current rule set $\mathcal{S}$, its estimated effect will be low, and thus its probability of being selected during the optimization process will decrease. The overall causal rule set learning process is shown in Alg. 1.

Algorithm 1 Causal rule set learning

```
Input: Training data D = {(T_i, X_i, Y_i)}_{i=1}^n, hyperparameters λ, K, and L
Initialize S ← ∅
for k = 1 to K do
    Solve R* ← arg max_R f(R)                    /* See Sec. 5.2 */
    if f(R*) > 0 then
        S ← S ∪ {R*}
        Change weighted outcome to ε for covered units
    end if
end for
Output: S
```

5.2 MM Procedure

For a single rule, its maximization objective function is $f(\mathcal{R})$. However, the complexity of $f(\mathcal{R})$ makes it difficult for traditional optimization algorithms to handle it directly. To this end, we propose to use the MM procedure. It is an iterative optimization method that, instead of finding the optimal solution to the original objective function $f(\mathcal{R})$, first finds an easy-to-optimize surrogate function $g(\mathcal{R})$ that approximates the original one (see Sec. 5.3 for detail). The solution of the surrogate function makes the optimal solution of $g(\mathcal{R})$ close to the optimal solution of $f(\mathcal{R})$. In each iteration, a new surrogate function for the next iteration is constructed based on the current solution. Mathematically, the solution can converge to the optimal solution to the original optimization problem [51].

Formally, taking the minorize-maximization version, $f(\mathcal{R})$ is the original objective function to be maximized. At the $m-$th ($m = 0, 1, \dots$) step of MM, the objective function $f(\mathcal{R})$ can be replaced by a surrogate function $g_m(\mathcal{R})$ if the following conditions are satisfied:

$$\begin{aligned} &g_m(\mathcal{R}) \le f(\mathcal{R}) \quad \forall \mathcal{R} \\ &g_m(\mathcal{R}_m) = f(\mathcal{R}_m). \end{aligned} \tag{6}$$

Formmally, we summarize the steps of the MM procedure in Alg. 2.

Algorithm 2 Single causal rule optimization

```
m = 0
Initialize R_m
while true do
    Construct g_m(R)                             /* See Sec. 5.3 */
    R_{m+1} = arg max_R g_m(R)
    if R_{m+1} = R_m then R* = R_m break end if
    m = m + 1
end while
Output: R*
```

5.3 Submodular Lower Bound Optimization

Here, we introduce how to construct a submodular approximation lower bound of the original objective function and the corresponding optimization method. Specifically, we aim to develop a rule $\mathcal{R}^\star$ that maximizing $f(\mathcal{R})$. First, we introduce the following inequality.

Proposition 1. $\sigma^2_{\mathcal{R}} \le \frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i (Y_i - \mu_{(m)})^2}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i}$, where $\mu_{(m)}$ is the weighted mean of outcome of the previous step in the MM procedure.

Proof. $\sigma^2_{\mathcal{R}}$ is the weighted variance. According to the definition, it can also be written as: $\sigma^2_{\mathcal{R}} = \mathbb{E}_w[Y_i^2] - (\mathbb{E}_w[Y_i])^2 = \frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i Y_i^2}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i} - (\frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i Y_i}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i})^2$. We perform Taylor expansion of the latter term, so that $\sigma^2_R \le \frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i Y_i^2}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i} - (2\mu_{(m)} \frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i Y_i}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i} - \mu^2_{(m)}) =$
$\frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i (Y_i^2 - 2\mu_{(t)} Y_i + \mu^2_{(m)})}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i} = \frac{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i (Y_i - \mu_{(m)})^2}{\sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i}$. □

Define $Q_5 = \sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i (Y_i - \mu_{(m)})^2$, $Q_6 = \sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i = Q_3$, then $f(\mathcal{R}) \ge \log Q_1 + \log Q_4 - \log Q_3 - \log Q_4 - \lambda(\log \frac{Q_5}{Q_6}) = \log Q_1 + \log Q_4 + \lambda \log Q_6 - \log Q_2 - \log Q_3 - \lambda \log Q_5$.

Each atom (covariate) of the antecedent in the rule corresponds to a part of the unit, and the unit corresponding to the entire rule is the intersection of the units corresponding to these atoms. Based on the formula for Q, Q can be viewed as the set function. Taking $Q_1 = \sum_{i \in \mathcal{D}_{\mathcal{R}}^+} w_i Y_i$ as an example, its corresponding units is $\mathcal{D}_{\mathcal{R}}^+$, and the corresponding value is the sum of the weighted outcome of these units. Then, we have the following property for Q functions.

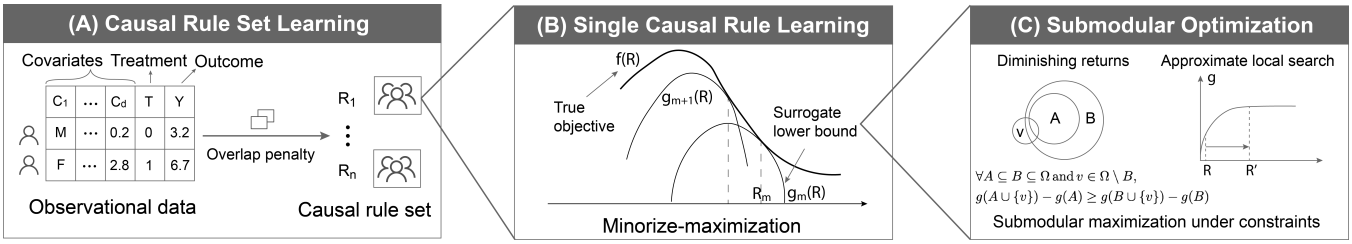

Figure 2: Illustration of the proposed algorithm. (A) A causal rule set is learned from the observational data, and overlap penalities are applied to minimize the case where a unit is covered by multiple rules. (B) A single causal rule is solved by the MM framework, and the rule is improved by iteratively optimizing the surrogate lower bound of the original objective. (C) We prove an surrogate lower bound with submodular properties, allowing us to efficiently solve the surrogate optimization problem using efficient submodular optimization.

Proposition 2. Q functions are supermodular.

Proof. Taking Q_1 as an example, $\mathcal{D}^+_{\mathcal{R}} = \{i | i \in \mathcal{D}^+ \wedge \boldsymbol{\alpha}_{\mathcal{R}}(\mathbf{X}_i) = 1\} = \{i | i \in \mathcal{D}^+ \wedge (\wedge_{j\in\mathcal{R}} x_{i,j}) = 1\}$. Thus Q_1 can be regarded as the weighted outcome sum of set $|\mathcal{D}^+ \cap (\cap_{j\in\mathcal{R}} x_j)|$. We can rewrite it as $|\mathcal{D}^+| - |\mathcal{D}^+ \cap (\overline{\cap_{j\in\mathcal{R}} x_j})| = |\mathcal{D}^+| - |\mathcal{D}^+ \cap (\cup_{j\in\mathcal{R}} \overline{x_j})| = |\mathcal{D}^+| - |\cup_{j\in\mathcal{R}} (\mathcal{D}^+ \cap \overline{x_j})|$. The latter term is the union of sets (coverage functions), which is a well-known submodular function, so Q_1 is a supermodular function. Similarly, it can be shown that other Q functions are also supermodular. □

For the supermodular function $Q : 2^V \to \mathbb{R}_{\geq 0}$, where $V = [d]$ is the universal set. The following modular functions gives two tight lower bounds approximating Q at $\mathcal{R}_m$ [45]:

$$\begin{aligned} b^1_{Q,\mathcal{R}_m}(\mathcal{R}) &= Q(\mathcal{R}_m) - \sum\nolimits_{j\in\mathcal{R}_m\setminus\mathcal{R}} Q(j|\mathcal{R}_m \setminus \{j\}) \\ &\quad + \sum\nolimits_{j\in\mathcal{R}\setminus\mathcal{R}_m} Q(j|\emptyset) \leq Q(\mathcal{R}), \forall \mathcal{R} \subseteq V \\ b^2_{Q,\mathcal{R}_m}(\mathcal{R}) &= Q(\mathcal{R}_m) - \sum\nolimits_{j\in\mathcal{R}_m\setminus\mathcal{R}} Q(j|V \setminus \{j\}) \\ &\quad + \sum\nolimits_{j\in\mathcal{R}\setminus\mathcal{R}_m} Q(j|\mathcal{R}_m) \leq Q(\mathcal{R}), \forall \mathcal{R} \subseteq V, \end{aligned} \tag{7}$$

where $Q(A|B) = Q(A \cup B) - Q(B)$ denotes the marginal gain from adding A to B. For ease of expression, we define

$$b_{Q,\mathcal{R}_m}(\mathcal{R}) = \max(b^1_{Q,\mathcal{R}_m}(\mathcal{R}), b^2_{Q,\mathcal{R}_m}(\mathcal{R})). \tag{8}$$

Then we have the following result:

Proposition 3. The function $g(\mathcal{R})$ defined below is a submodular lower bound of $f(\mathcal{R})$:

$$\begin{aligned} g(\mathcal{R}) = &\log b_{Q_1,\mathcal{R}_m}(\mathcal{R}) + \log b_{Q_4,\mathcal{R}_m}(\mathcal{R}) + \lambda \log b_{Q_6,\mathcal{R}_m}(\mathcal{R}) \\ &- (\log Q_{2,\mathcal{R}_m} + \frac{Q_{2,\mathcal{R}} - Q_{2,\mathcal{R}_m}}{Q_{2,\mathcal{R}_m}}) - (\log Q_{3,\mathcal{R}_m} + \frac{Q_{3,\mathcal{R}} - Q_{3,\mathcal{R}_m}}{Q_{3,\mathcal{R}_m}}) \\ &- \lambda(\log Q_{5,\mathcal{R}_m} + \frac{Q_{5,\mathcal{R}} - Q_{5,\mathcal{R}_m}}{Q_{5,\mathcal{R}_m}}). \end{aligned}$$

Proof. Since $b_{Q,\mathcal{R}_m}(\mathcal{R})$ is a modular function, then $\log b_{Q,\mathcal{R}_m}(\mathcal{R})$ is a submodular function. The first-order Taylor expansion of $\log Q$ is $\log Q_{\mathcal{R}_m} + \frac{Q_{\mathcal{R}} - Q_{\mathcal{R}_m}}{Q_{\mathcal{R}_m}}$, where $Q_{\mathcal{R}_m}$ are constants, so $-Q_{\mathcal{R}}$ is a submodular function. Thus, the $g(\mathcal{R})$ is a submodular function. □

For the submodular lower bound $g(\mathcal{R})$, there exists an approximate local search algorithm [40] that approaches the optimum by continuously performing local improvements. Specifically, starting from the initial rule, $g(\mathcal{R})$ is gradually maximized by local operations, including adding, removing, or replacing covariates. Formally, the detailed algorithm procedures are summarized in Alg. 3.

Algorithm 3 Submodular lower bound optimization

```
Input: Current rule R
while true do
    R' ← R
    while ∃j ∈ [d] \ R s.t. g(j|R) > 0 do R ← R ∪ {j} end while
    while ∃j ∈ R s.t. g(j|R \ {j}) ≤ 0 do R ← R \ {j} end while
    while ∃i ∈ R, j ∈ [d] \ R s.t. g(j|R \ {i}) > 0 do R ← (R \ {i}) ∪ {j} end while
    if R = R' then break end if
end while
Output: R
```

6 EVALUATION

We present detailed experimental evaluation of *CURLS*, including quantitative experiments and qualitative case studies.

6.1 Quantitative Experiments

The quantitative experiments aim to evaluate the efficacy of *CURLS* in identifying significant treatment effects in subgroups. Since real-world datasets lack the groundtruth of CATE, which affects the calculation of evaluation metrics, we compare *CURLS* with various baselines on synthetic and semi-synthetic datasets.

Datasets. For synthetic data, following the settings in [1, 59], we sampled units under the assumption of unconfoundedness, where the covariates are generated from the following distribution:

$$\begin{aligned} &X_1, \cdots, X_i \sim \text{Categorical}(\{A, B, C, D, E\}), \\ &X_{i+1}, \cdots, X_d \sim \text{Normal}(0, 1). \end{aligned} \tag{9}$$

The treatment T is generated according to a Bernoulli distribution, where the probability of $T = 1$ is given by the sigmod function

with respect to X. This simulates the non-randomness of treatment assignment in the observational data. Categorical variables are converted to one-hot encoding for calculation. Formally, we have

$$\begin{aligned} f(X) &= \sigma(\langle X, \beta\rangle + \eta), \\ \eta &\sim \text{Uniform}(-1, 1), \\ \beta &\sim \text{Uniform}(0, b)^{|X|}, \\ T &\sim \text{Bernoulli}(f(X)). \end{aligned} \tag{10}$$

The treatment effect TE and the outcome Y is generated by the following formula. An offset is added to Y to ensure that Y is positive. That is,

$$\begin{aligned} TE = \langle X, \alpha\rangle, \alpha &\sim \text{Uniform}(0, 2)^{|X|}, \\ Y = T \cdot \text{TE} + \langle X, \gamma\rangle + Y_{\text{offset}} &+ \epsilon, \\ Y_{\text{offset}} &= \max(0, -Y_{\min}), \\ \epsilon \sim \text{Uniform}(-1, 1), \gamma &\sim \text{Uniform}(0, 1)^{|X|}. \end{aligned} \tag{11}$$

We also collected the famous semi-synthetic dataset IHDP [1], which is constructed from the infant health and development program. The detail information of the datasets is shown in Table 1.

Table 1: Dataset statistics.

Dataset	#Units	#Categorical_cov	#Numerical_cov	b
Syn-data1	3000	5	5	0.6
Syn-data2	3000	5	10	0.5
Syn-data3	4000	5	15	0.3
IHDP	7470	19	6	/

Baselines. We compare the proposed algorithm *CURLS* with two groups of algorithms. The first group is the popular heterogeneous treatment effect estimation algorithms: (1) Causal Tree (CT) [1]; (2) Causal Forest (CF) [55]; and (3) Causal Rule Ensemble (CRE) [6]. The second group is the correlation rule learning and subgroup discovery algorithms: (1) BRCG [14]; (2) Decision Tree (DT) [10]; (3) Pysubgroup (PYS) [41]. In the first group, CRE can explicitly obtain the antecedent and treatment effect of the rule. For CT and CF, it can be considered that the path from the root to the leaf nodes in the tree structure is the antecedent of the causal rule, and the CATE value of the leaf node is the effect corresponding to the rule. The second group of methods can only get the correlation rules. In order to adapt to the setting of causal rule learning, we add a post-processing step. CATE is calculated on the data covered by each rule via the normalized IPW method [29]. For the consistency of comparison, the rules with the top K effect values in the baselines are taken out and compared with *CURLS*.

Metrics. In order to evaluate the effectiveness of the causal rules. On the one hand, we assess the subgroup treatment effect from the perspectives of effect strength, uncertainty, and accuracy. The specific metrics are described as follows: (1) Estimated CATE; (2) True CATE (mean value of ITE within subgroup); (3) The variance of the outcome of the treated units in the subgroup. (4) The precision in the estimation of heterogeneous effects PEHE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}(\mathbf{x}_i) - \tau(\mathbf{x}_i))^2}$; (5) The mean absolute percentage error MAPE = $\frac{1}{n}\sum_{i=1}^{n}|\frac{\hat{\tau}(\mathbf{x}_i)-\tau(\mathbf{x}_i)}{\tau(\mathbf{x}_i)}|$. On the other hand, we have also measured the interpretability of the rule set, including the following metrics: (1) Average length of rule antecedent; (2) Average overlap between pairs of rules; (3) Rule set coverage.

Implement detail. We used 5-fold cross-validation and Bayesian optimization to tune parameters. Specifically, we optimize the parameters of *CURLS* with max rule length $L \in \{3, 4, 5, 6\}$, variance weight $\lambda \in \{0.1, \cdots, 1.5\}$. CT's hyperparameters include cross-validation method cv.option="matching" and pruning factor pru_coef $\in \{0.4, 0.9, 1.5\}$. For CF, its hyperparameter takes the values num.trees $\in \{5, 8, 10\}$, honest version of the CT split.Honest=TRUE and tradeoff between effect and variance split.alpha $\in \{0.2, 0.5, 0.8\}$. The CRE parameters include ntrees $\in \{20, 25\}$, max_depth $\in \{3, 4\}$ and the decay threshold for rules pruning t_decay $\in \{0.025, 0.01, 0.04\}$. For DT, the depth of tree max_depth is fixed as 4. In PYS, the result set has 10 rules, with a maximum rule depth of $\{2, 5\}$, using the subgroup scoring method qf=ps.WRAccQF(). For BRCG, we tune the maximum number of columns generated per iteration K from 8 to 12, and the max rule length is chosen from $\{5, 10\}$.

Results. The evaluation results are reported in Table 2. We recorded the estimated CATE and the ground truth CATE (Avg_ITE) to assess the strength of the treatment effect. The results show that the estimated CATE and true CATE of both rules of *CURLS* is about 16.1% and 13.8% higher than the effect values of other baselines. We also observe that for correlation learning methods, such as DT and PYS, the highest effect is not the rule with a high predicted value of Y (in our experiment, it is the rules with the probability of Y around 0.6-0.8), which reflects the difference between correlation and causation. When it comes to variance, *CURLS* reduces variance by about 12.0% compared to other baselines. In addition, *CURLS* sacrifices certain effects to reduce variance when necessary. For example, in rule2 of data3, the effect and variance of *CURLS* are 12.47 and 16.36, while the effect of PYS and BRCG are 12.91 and 13.90, respectively, which are larger than *CURLS*, but their variance is also large at 21.92 and 23.58. We also compared PEHE and MAPE, which measure the accuracy of ITE estimates. The results show that the estimation accuracy of *CURLS* is, on average, 0.05% higher on PEHE and 1.6% smaller on MAPE compared to the other methods. This suggests that *CURLS* is able to find subgroups with more significant treatment effects with similar or better estimation accuracy. It is worth noting that while CATE is the best estimate of ITE in terms of the mean squared error [33], it can also lead to inaccurate estimates because the estimation method we used, IPW, inherently has errors in estimating when the propensity score approaches 0 or 1. A potential solution is to introduce more robust estimators, such as doubly robust estimation [18].

Table 3 shows the relevant metrics on the rule set readability. We found that the average rule length of *CURLS* is around 3, which is mostly smaller than tree-based methods such as CT and CF. In addition, the overlap between rules in *CURLS* is also at a low level of 0.1%, which is favorable for user understanding. The coverage metric shows that *CURLS* focuses on a small number of groups with strong effects, while other methods, like CT and CRE, have coverage rates as high as more than 50%. However, the variance of their coverage is also very large, indicating that there are significant differences between their rules. Also, the high coverage may be responsible for their more average treatment effects.

[1] https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP

Table 2: 5-fold average performance metrics for different rules. Numbers in parentheses represent standard deviations. (The CRE results on IHDP are missing because it cannot find any rules, and some standard deviations of BRCG are nan since sometimes only one fold can obtain a rule that meet the requirements.)

Dataset	Metrics	CURLS		CT		CF		CRE		DT		PYS		BRCG	
		Rule1	Rule2	Rule1	Rule2	Rule1	Rule2	Rule1	Rule2	Rule1	Rule2	Rule1	Rule2	Rule1	Rule2
Syn-data1	CATE↑	10.00(0.65)	8.39(0.25)	8.97(0.61)	7.36(0.88)	7.42(0.63)	6.33(1.45)	7.27(0.31)	6.97(0.37)	6.93(0.24)	6.72(0.25)	7.53(0.55)	7.13(0.09)	8.16(0.25)	7.60(0.29)
	Avg_ITE↑	9.17(0.38)	7.98(0.69)	7.79(0.52)	5.89(0.72)	7.69(0.68)	6.15(1.08)	6.81(0.50)	6.95(0.96)	6.75(0.44)	7.22(0.52)	7.57(0.45)	7.17(0.27)	7.99(0.13)	7.82(0.62)
	Variance↓	9.20(2.24)	9.69(3.08)	10.32(2.60)	10.37(1.55)	8.94(1.36)	7.92(2.18)	9.79(2.90)	9.85(1.73)	8.32(3.14)	9.23(0.94)	11.59(2.30)	12.07(1.70)	11.60(2.21)	9.91(2.15)
	PEHE↓	2.25(0.39)	2.38(0.33)	2.46(0.14)	2.62(0.05)	2.17(0.15)	1.84(0.20)	2.14(0.28)	2.23(0.16)	1.94(0.24)	2.13(0.17)	2.22(0.06)	2.36(0.10)	2.28(0.20)	2.23(0.11)
	MAPE↓	0.23(0.05)	0.28(0.06)	0.33(0.03)	0.56(0.10)	0.23(0.06)	0.28(0.05)	0.32(0.08)	0.30(0.07)	0.27(0.03)	0.24(0.04)	0.27(0.02)	0.31(0.03)	0.26(0.03)	0.25(0.05)
Syn-data2	CATE↑	12.72(0.55)	11.44(0.69)	10.95(0.50)	9.93(0.48)	11.30(0.58)	10.89(0.62)	10.23(0.44)	9.89(0.29)	9.40(0.46)	8.68(0.37)	10.24(0.07)	10.05(0.06)	11.55(0.25)	10.99(0.50)
	Avg_ITE↑	11.23(0.64)	10.53(0.75)	8.99(0.49)	8.55(0.52)	10.29(0.42)	9.70(0.69)	9.89(0.52)	10.11(1.27)	9.39(0.61)	8.87(0.69)	9.60(0.30)	9.50(0.27)	10.42(0.62)	10.06(0.61)
	Variance↓	11.15(2.91)	10.65(3.61)	12.79(2.43)	14.37(1.08)	11.78(3.76)	15.28(2.08)	13.15(2.55)	12.51(1.27)	8.71(4.18)	11.79(3.12)	13.78(1.85)	14.50(2.56)	11.82(2.92)	14.53(2.39)
	PEHE↓	2.64(0.49)	2.13(0.31)	2.93(0.24)	2.62(0.29)	2.32(0.35)	2.55(0.19)	2.19(0.24)	2.47(0.33)	2.10(0.30)	1.95(0.13)	2.29(0.19)	2.30(0.20)	2.46(0.39)	2.51(0.09)
	MAPE↓	0.23(0.06)	0.19(0.03)	0.34(0.05)	0.32(0.06)	0.21(0.05)	0.25(0.04)	0.20(0.04)	0.22(0.11)	0.19(0.04)	0.19(0.04)	0.22(0.03)	0.23(0.03)	0.22(0.06)	0.19(0.04)
Syn-data3	CATE↑	14.06(0.25)	12.70(1.55)	14.56(0.54)	13.65(0.70)	12.19(0.85)	11.15(1.06)	13.37(0.46)	12.73(0.60)	11.90(0.73)	11.03(0.67)	12.73(0.11)	12.56(0.03)	13.53(0.26)	13.22(0.15)
	Avg_ITE↑	13.80(0.61)	12.47(1.91)	12.87(0.16)	12.08(0.97)	12.74(1.86)	11.08(1.34)	12.73(1.72)	11.89(1.47)	12.53(0.68)	11.75(1.15)	12.74(0.22)	12.91(0.33)	13.31(0.32)	13.90(0.47)
	Variance↓	16.18(4.44)	16.36(5.34)	19.31(4.81)	19.10(3.38)	23.45(9.63)	16.42(4.48)	22.36(7.74)	19.99(3.78)	18.06(2.88)	17.48(1.87)	20.83(1.43)	21.92(3.52)	20.04(2.33)	23.58(7.62)
	PEHE↓	2.99(0.19)	3.08(0.38)	3.38(0.27)	3.37(0.12)	3.76(0.70)	2.75(0.23)	3.52(0.84)	3.30(0.32)	3.17(0.22)	2.92(0.40)	3.10(0.21)	3.24(0.22)	3.09(0.34)	3.39(0.50)
	MAPE↓	0.19(0.02)	0.23(0.07)	0.26(0.03)	0.28(0.04)	0.25(0.03)	0.23(0.03)	0.26(0.12)	0.28(0.08)	0.21(0.04)	0.21(0.03)	0.21(0.02)	0.21(0.02)	0.20(0.01)	0.20(0.01)
IHDP	CATE↑	10.98(1.72)	9.98(0.45)	6.52(0.37)	3.19(0.87)	8.11(0.47)	7.75(0.61)	/	/	8.79(1.33)	6.90(1.14)	4.24(0.39)	4.05(0.26)	3.36(0.80)	2.06(nan)
	Avg_ITE↑	8.44(0.70)	8.51(0.71)	6.08(0.36)	3.15(1.26)	6.39(1.10)	7.01(0.57)	/	/	7.84(0.65)	6.22(0.55)	4.39(0.40)	4.19(0.44)	3.47(0.60)	−0.93(nan)
	Variance↓	4.17(5.03)	24.06(16.35)	139.72(31.36)	210.22(33.57)	19.36(29.37)	148.97(130.44)	/	/	28.80(59.60)	126.27(104.86)	172.14(17.54)	166.78(28.04)	173.12(28.99)	198.46(nan)
	PEHE↓	10.78(1.22)	10.42(1.57)	8.58(0.39)	10.12(2.78)	8.95(1.63)	8.59(1.15)	/	/	9.96(1.75)	7.36(1.64)	9.98(1.51)	9.75(1.66)	9.34(0.88)	23.10(nan)
	MAPE↓	2.03(0.65)	2.08(0.94)	1.86(0.34)	1.25(0.29)	1.58(0.21)	1.54(0.34)	/	/	2.01(0.90)	2.21(1.74)	1.62(0.41)	1.51(0.37)	1.32(0.26)	2.29(nan)

Table 3: Interpretability metrics for the rule set, reported as mean and standard deviation.

Dataset	Metrics	CURLS	CT	CF	CRE	DT	PYS	BRCG
Syn-data1	Avg_len	2.9 (0.8)	2.3 (0.8)	5.1 (1.5)	2.1 (0.7)	4.0 (0.0)	1.1 (0.2)	2.3 (0.3)
	Overlap(%)	0.1 (0.1)	0.0 (0.0)	0.0 (0.1)	1.4 (0.7)	0.0 (0.0)	3.7 (0.7)	1.0 (0.9)
	Coverage(%)	6.0 (0.6)	46.1 (41.5)	10.6 (2.8)	24.1 (6.3)	14.5 (1.4)	31.3 (6.5)	16.5 (6.7)
Syn-data2	Avg_len	3.0 (0.0)	5.9 (2.1)	2.7 (0.8)	1.9 (0.5)	4.0 (0.0)	1.0 (0.0)	2.5 (0.4)
	Overlap(%)	0.2 (0.2)	0.0 (0.0)	1.2 (0.8)	1.1 (2.2)	0.0 (0.0)	4.1 (1.0)	2.1 (1.8)
	Coverage(%)	7.4 (1.0)	36.9 (22.2)	11.7 (1.6)	38.9 (34.3)	14.7 (1.9)	36.4 (2.2)	13.9 (1.4)
Syn-data3	Avg_len	2.8 (0.5)	4.6 (2.0)	5.3 (1.2)	1.4 (0.6)	4.0 (0.0)	1.0 (0.0)	2.3 (0.7)
	Overlap(%)	0.1 (0.2)	0.0 (0.0)	0.5 (0.6)	0.5 (1.2)	0.0 (0.0)	3.7 (0.6)	0.6 (0.1)
	Coverage(%)	12.1 (1.4)	27.5 (21.4)	8.4 (1.6)	68.1 (36.5)	16.4 (3.8)	35.8 (2.7)	14.7 (1.1)
IHDP	Avg_len	3.0 (0.8)	4.7 (2.3)	3.9 (1.6)	/	4.0 (0.0)	4.8 (0.3)	4.3 (0.8)
	Overlap(%)	0.7 (0.8)	0.0 (0.0)	5.0 (6.8)	/	0.0 (0.0)	41.6 (9.5)	0.0 (0.0)
	Coverage(%)	8.3 (1.8)	25.8 (24.6)	11.5 (3.4)	/	21.3 (10.9)	55.8 (5.0)	53.5 (8.8)

6.2 Case Studies

We qualitatively evaluate the performance of *CURLS* on two real and easily understandable accident analysis and policy making datasets. The Titanic dataset[2] provides passenger data on survival, sex, age, fares, number of siblings/spouses on board (sibsp) and number of parents/children (parch) *etc.* We want to determine how premium class (treatment) affects passenger survival (outcome). For the treatment, $T = 1$ means that the passenger is in premium class 1 and 2 cabins, and $T = 0$ means the lowest class 3 cabin. We extracted three causal rules from the dataset; the first shows that upgrading to a higher class improves survivability for passengers with more family members, who may be more willing to aid each other. The second and third rules correspond to the fact that women and children who pay higher ticket fees are more likely to be rescued due to the abundant rescue resources in the higher class and the "women and children first" policy.

The Lalonde dataset (part of the famous Jobs dataset) [35] includes data from participants and non-participants in a job training program (National Supported Work Demonstration, NSW). NSW is an experimental program that aims to help economically disadvantaged people (*e.g.*, long-term unemployed, high school dropouts) return to work. It would train underprivileged workers in work skills for 9-18 months. We evaluated how the training program (treatment) affected income (outcome). The covariates include individual background information (age, race, academic background, and past income). Table 4 reveals two subgroups with high treatment effects yielded by *CURLS*. The first subgroup is married people over 29 who may be living a stressful life and will study hard to increase their income during training. The second subgroup is 18-19-years-olds. They have good learning capacity and ambition, so they can get high-paying employment through training despite their lack of experience. We used DoWhy[3], a famous causal inference package, to calculate the treatment effect for the entire population, which is 1639.8, less than half of the two mentioned subgroups. With reference to causal rules, policymakers can better choose target groups to improve program implementation outcomes.

7 DISCUSSION

In this section, we discuss the implications, scalability, limitations and future work of *CURLS*.

[2]https://www.kaggle.com/c/titanic/data

[3]https://github.com/py-why/dowhy

Table 4: Examples of learned causal rules.

Titanic
IF sibsp > 1.0 AND parch > 1.0 THEN τ = 0.88
IF sex == female AND fare > 39.7 THEN τ = 0.81
IF age <= 20.0 AND fare > 21.7 THEN τ = 0.75
Lalonde
IF age > 29.0 AND married == 1.0 THEN τ = 4722.3
IF age <= 19.0 AND age > 18.0 THEN τ = 4165.8

Implications. This research leads to two key implications. Firstly, causal rule learning is helpful, and exploration of its algorithmic design is encouraging. The concise rules in the form of "IF-THEN" are similar to the logic of human decision-making. They are easier to understand compared to complicated tree structures and black-box methods for treatment effect estimation, making it simpler to discover new causal knowledge. Secondly, *CURLS* can be adopted in practical applications, as our real-world data cases demonstrate. In addition to those examples, *CURLS* can also be adapted to scenarios requiring causal-assisted decision-making, such as education and industry. For example, *CURLS* may help teachers find the best way to teach different students based on their characteristics. It may also assist quality inspectors troubleshoot metric anomalies and attribute them to specific models.

Scalability. The computational complexity of *CURLS* is mainly dominated by local search in Alg. 3, which is linear in the number of covariates. To demonstrate this empirically, we conducted an experiment to measure the algorithm running time under different numbers of covariates and units. As shown in Fig. 3, the training time grows linearly with the number of covariates.

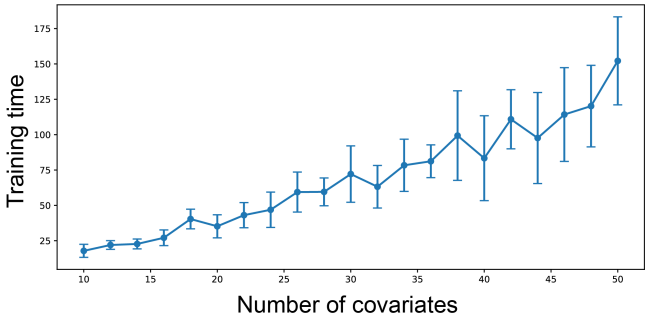

Figure 3: Scalability test. Training time scales linearly with the number of covariates.

Limitations and future work. We believe there are three potential directions that *CURLS* can further explore. First, *CURLS* may converge to suboptimal results due to its iterative optimization procedure, where poor initialization and iteration paths can lead to local optima. To address this, we use a greedy strategy for initialization for MM and incorporate local search techniques at the end. We are also exploring other optimization methods, such as neural combinatorial optimization [8] and multi-objective learning [50], to further improve the quality of the final solution. Second, the descriptive ability of antecedents is limited. While *CURLS* uses a form of CNF with AND logical connectives and binary covariates, its inability to handle OR connectives and limitations in discretization may restrict it from describing certain refined subgroups. Future work involves integrating logical connectives into the learning process and adaptively determining the discretization. Finally, the assumptions on single treatment and single outcome may not be compatible with real-world scenarios. In practice, there may be multiple treatments, unobserved variables, multiple outcomes, or more complex causal relationships. For non-binary treatments, we can extend our method by utilizing One-Versus-The-Rest (OvR), which is commonly used for multi-classification tasks, to handle each treatment value. Other methods such as robust HTE [31] and instrumental variables [52] can also be investigated to ensure the validity of causal inference.

8 CONCLUSION

In this paper, we propose a new method called *CURLS* for learning causal rules from observational data. To the best of our knowledge, this is the first method that employs an optimization problem to generate rules to explain subgroups with significant treatment effects. We formally define these causal rules composed of antecedents that form conditions to characterize the subgroup and the associated effects. We model the rule learning process as a discrete optimization problem. By constructing an approximate submodular lower bound for the original objective, the problem can be solved iteratively based on the minorize-maximization algorithm. Quantitative experiments and qualitative case studies demonstrate that our method is effective in identifying meaningful causal rules from observational data. Future works involve more effective optimization algorithms, refining rule formation, and addressing more complex scenarios.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (62132017), Zhejiang Provincial Natural Science Foundation of China (LD24F020011) and Alibaba Group through Alibaba Research Intern Program.

REFERENCES

[1] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360. https://doi.org/10.1073/pnas.1510489113

[2] Susan Athey and Stefan Wager. 2019. Estimating treatment effects with causal forests: An application. *Observational Studies* 5, 2 (2019), 37–51. https://doi.org/10.1353/obs.2019.0001

[3] Martin Atzmueller. 2015. Subgroup discovery. *WIREs Data Mining and Knowledge Discovery* 5, 1 (2015), 35–49. https://doi.org/10.1002/widm.1144

[4] Martin Atzmüller and Florian Lemmerich. 2009. Fast Subgroup Discovery for Continuous Target Concepts. In *Proceedings of International Syposium on Methodologies for Intelligent Systems*. 35–44. https://doi.org/10.1007/978-3-642-04125-9_7

[5] Martin Atzmüller and Frank Puppe. 2006. SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*. 6–17. https://doi.org/10.1007/11871637_6

[6] Falco J. Bargagli-Stoffi, Riccardo Cadei, Kwonsang Lee, and Francesca Dominici. 2020. Causal Rule Ensemble: Interpretable Discovery and Inference of Heterogeneous Causal Effects. *arXiv preprint arXiv:2009.09036* (2020). https://doi.org/10.48550/arXiv.2009.09036

[7] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML. Version 0.x.

[8] Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. 2016. Neural Combinatorial Optimization with Reinforcement Learning. *ArXiv* abs/1611.09940 (2016). https://doi.org/10.48550/arXiv.1611.09940

[9] Ranjita Bhagwan, Rahul Kumar, Ramachandran Ramjee, George Varghese, Surjyakanta Mohapatra, Hemanth Manoharan, and Piyush Shah. 2014. Adtributor: Revenue Debugging in Advertising Systems. In *Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation*. 43–55. https://www.usenix.org/conference/nsdi14/technical-sessions/presentation/bhagwan

[10] L. Breiman and Richard A. Olshen. 2017. Points of Significance: Classification and regression trees. *Nature Methods* 14 (2017), 757–758. https://doi.org/10.1038/nmeth.4370

[11] Tianxi Cai, Lu Tian, Peggy H. Wong, and L. J. Wei. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12, 2 (2011), 270–282. https://doi.org/10.1093/biostatistics/kxq060

[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68. https://doi.org/10.1111/ectj.12097

[13] William W. Cohen. 1995. Fast Effective Rule Induction. In *Proceedings of International Conference on Machine Learning*. 115–123. https://doi.org/10.1016/B978-1-55860-377-6.50023-2

[14] Sanjeeb Dash, Oktay Günlük, and Dennis Wei. 2018. Boolean Decision Rules via Column Generation. In *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1805.09901

[15] Jonathan M.V. Davis and Sara B. Heller. 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review* 107, 5 (2017), 546–550. https://doi.org/10.1257/aer.p20171000

[16] María José del Jesús, Pedro González, Francisco Herrera, and Mikel Mesonero. 2007. Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A Case Study in Marketing. *IEEE Transactions on Fuzzy Systems* 15, 4 (2007), 578–592. https://doi.org/10.1109/TFUZZ.2006.890662

[17] Jerome H. Friedman and Bogdan E. Popescu. 2008. PREDICTIVE LEARNING VIA RULE ENSEMBLES. *The Annals of Applied Statistics* 2, 3 (2008), 916–954. https://doi.org/10.1214/07-AOAS148

[18] Michele Jonsson Funk, Daniel J. Westreich, Christopher A. Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American Journal of Epidemiology* 173, 7 (2011), 761–767. https://doi.org/10.1093/aje/kwq439

[19] Dragan Gamberger and Nada Lavrac. 2002. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research* 17 (2002), 501–527. https://doi.org/10.1613/jair.1089

[20] Markus Gangl. 2010. Causal inference in sociological research. *Annual review of sociology* 36, 1 (2010), 21–47. https://doi.org/10.1146/annurev.soc.012809.102702

[21] Henrik Grosskreutz and Stefan Rüping. 2009. On subgroup discovery in numerical domains. *Data Mining and Knowledge Discovery* 19 (2009), 210–226. https://doi.org/10.1007/s10618-009-0136-3

[22] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. 2008. Tight Optimistic Estimates for Fast Subgroup Discovery. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 440–456. https://doi.org/10.1007/978-3-540-87479-9_47

[23] Jiazhen Gu, Chuan Luo, Si Qin, Bo Qiao, Qingwei Lin, Hongyu Zhang, Ze Li, Yingnong Dang, Shaowei Cai, Wei Wu, Yangfan Zhou, Murali Chintalapati, and Dongmei Zhang. 2020. Efficient incident identification from multi-dimensional issue reports via meta-heuristic search. *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2020), 292–303. https://doi.org/10.1145/3368089.3409741

[24] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2020. A Survey of Learning Causality with Data: Problems and Methods. *Comput. Surveys* 53, 4, Article 75 (jul 2020), 37 pages. https://doi.org/10.1145/3397269

[25] Jens Hainmueller. 2012. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis* 20, 1 (2012), 25–46. https://doi.org/10.1093/pan/mpr025

[26] Francisco Herrera, Cristóbal José Carmona, Pedro González, and María José del Jesus. 2011. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems* 29 (2011), 495–525. https://doi.org/10.1007/s10115-010-0356-2

[27] Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240. https://doi.org/10.1198/jcgs.2010.08162

[28] Keisuke Hirano, Guido Imbens, and Geert Ridder. 2000. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71, 4 (2000), 1161–1189. https://doi.org/10.1111/1468-0262.00442

[29] Guido W. Imbens. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics* 86, 1 (2004), 4–29. https://doi.org/10.1162/003465304323023651

[30] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *Proceedings of International Conference on Machine Learning*, Vol. 48. 3020–3029. https://doi.org/10.48550/arXiv.1605.03661

[31] Edward H. Kennedy. 2023. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17, 2 (2023), 3008–3049. https://doi.org/10.1214/23-EJS2157

[32] Andreas Krause and Carlos Guestrin. 2008. Beyond convexity: Submodularity in machine learning. *International Conference on Machine Learning* (2008). https://las.inf.ethz.ch/submodularity/icml08/index.html

[33] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of National Academy of Sciences* 116, 10 (2019), 4156–4165. https://doi.org/10.1073/pnas.1804597116

[34] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1675––1684. https://doi.org/10.1145/2939672.2939874

[35] Robert J. LaLonde. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review* 76, 4 (1986), 604–620. https://www.jstor.org/stable/1806062

[36] Nada Lavrač. 2005. Subgroup Discovery Techniques and Applications. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2–14. https://doi.org/10.1007/11430919_2

[37] Nada Lavrač, Bojan Cestnik, Dragan Gamberger, and Peter A. Flach. 2004. Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned. *Machine Learning* 57 (2004), 115–143. https://doi.org/10.1023/B:MACH.0000035474.48771.cd

[38] Nada Lavrač, Branko Kavšek, Peter A. Flach, and Ljupčo Todorovski. 2004. Subgroup Discovery with CN2-SD. *The Journal of Machine Learning Research* 5 (2004), 153–188. https://api.semanticscholar.org/CorpusID:60988

[39] Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. 2009. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing* (Bethesda, MD, USA) *(STOC '09)*. Association for Computing Machinery, New York, NY, USA, 323–332. https://doi.org/10.1145/1536414.1536459

[40] Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. 2010. Maximizing Nonmonotone Submodular Functions under Matroid or Knapsack Constraints. *SIAM Journal on Discrete Mathematics* 23, 4 (2010), 2053–2078. https://doi.org/10.1137/090750020

[41] Florian Lemmerich and Martin Becker. 2018. pysubgroup: Easy-to-use subgroup discovery in python. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 658–662. https://doi.org/10.1007/978-3-030-10997-4_46

[42] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma. 2015. From Observational Studies to Causal Rule Mining. *ACM Transactions on Intelligent Systems and Technology* 7, 2 (2015), 1–27. https://doi.org/10.1145/2746410

[43] Hui Lin and Jeff Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 510–520. https://aclanthology.org/P11-1052

[44] Graziano Mita, Paolo Papotti, Maurizio Filippone, and Pietro Michiardi. 2020. LIBRE: Learning Interpretable Boolean Rule Ensembles. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Vol. 108. 245–255. https://doi.org/10.48550/arXiv.1911.06537

[45] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming* 14 (1978), 265–294. https://doi.org/10.1007/BF01588971

[46] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms.* https://doi.org/10.1080/00949655.2018.1505197
[47] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55. https://doi.org/10.1093/biomet/70.1.41
[48] Kenneth J. Rothman and Sander Greenland. 2005. Causation and causal inference in epidemiology. *American journal of public health* 95, S1 (2005), S144–S150. https://doi.org/10.2105/AJPH.2004.059204
[49] Jasjeet Sekhon. 2008. The Neyman—Rubin Model of Causal Inference and Estimation Via Matching Methods. In *The Oxford Handbook of Political Methodology*. 271–299. https://doi.org/10.1093/oxfordhb/9780199286546.003.0011
[50] Yinan Shao, Jerry Chun-Wei Lin, Gautam Srivastava, Dongdong Guo, Hongchun Zhang, Hu Yi, and Alireza Jolfaei. 2023. Multi-Objective Neural Evolutionary Algorithm for Combinatorial Optimization Problems. *IEEE Transactions on Neural Networks and Learning Systems* 34, 4 (2023), 2133–2143. https://doi.org/10.1109/TNNLS.2021.3105937
[51] Ying Sun, Prabhu Babu, and Daniel P. Palomar. 2017. Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning. *IEEE Transactions on Signal Processing* 65, 3 (2017), 794–816. https://doi.org/10.1109/TSP.2016.2601299
[52] Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. 2019. Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems* 32 (2019). https://doi.org/10.48550/arXiv.1905.10176
[53] Matthijs van Leeuwen and Arno J. Knobbe. 2012. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery* 25 (2012), 208–242. https://doi.org/10.1007/s10618-012-0273-y
[54] Hal R. Varian. 2016. Causal inference in economics and marketing. *Proceedings of National Academy of Sciences* 113, 27 (2016), 7310–7315. https://doi.org/10.1073/pnas.1510479113
[55] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839
[56] Tong Wang and Cynthia Rudin. 2022. Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS Journal on Computing* 34, 3 (2022), 1626–1643. https://doi.org/10.1287/ijoc.2021.1143
[57] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 2017. A Bayesian Framework for Learning Rule Sets for Interpretable Classification. *Journal of Machine Learning Research* 18, 70 (2017), 1–37. http://jmlr.org/papers/v18/16-003.html
[58] Stefan Wrobel. 1997. An Algorithm for Multi-relational Discovery of Subgroups. In *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery*. 78–87. https://doi.org/10.1007/3-540-63223-9_108
[59] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. 2023. Stable Estimation of Heterogeneous Treatment Effects. In *Proceedings of International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:260814131
[60] Ying Wu, Hanzhong Liu, Kai Ren, and Xiangyu Chang. 2023. Causal Rule Learning: Enhancing the Understanding of Heterogeneous Treatment Effect via Weighted Causal Rules. *arXiv preprint arXiv:2310.06746* (2023). https://doi.org/10.48550/arXiv.2310.06746
[61] Fan Yang, Kai He, Linxiao Yang, Hongxia Du, Jingbang Yang, Bo Yang, and Liang Sun. 2021. Learning Interpretable Decision Rule Sets: A Submodular Optimization Approach. In *Advances in Neural Information Processing Systems*, Vol. 34. 27890–27902. https://doi.org/10.48550/arXiv.2206.03718
[62] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data* 15, 5, Article 74 (may 2021), 46 pages. https://doi.org/10.1145/3444944
[63] Guangyi Zhang and Aristides Gionis. 2020. Diverse Rule Sets. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), 1532—-1541. https://doi.org/10.1145/3394486.3403204
[64] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. 2001. SPEA2: Improving the strength Pareto evolutionary algorithm. *TIK report* 103 (2001). https://doi.org/10.3929/ethz-a-004284029

A SUPPLEMENTAL EXPERIMENTS

The above experiments compares *CURLS* to popular tree-based CATE methods (CT, CF), causal rule learning method (CRE), subgroup discovery methods (DT, PYS), and rule learning method (BRCG). The path from the root to the leaf nodes can be viewed as a description of the subgroups in these tree- or rule-based methods, whereas other black-box causal heterogeneity or uplift modeling models usually lack interpretability and are not used for the comparison. As a workaround, we supplemented new baselines, including double machine learning, doubly robust, and orthogonal random forest. We train a decision tree on the effects of these estimators using Tree Interpreter from a popular causal inference package, EconML [7], to indirectly obtain subgroup descriptions. New baselines includes:

- LDML: The double machine learning estimator with a low-dimensional linear final stage implemented as a statsmodel regression.
- DRL: CATE estimator that uses doubly-robust correction techniques to account for covariate shift (selection bias) between the treatment arms.
- DROF: Orthogonal random forest for discrete treatments using the doubly robust moment function.

As shown in Table 5, *CURLS* still achieves competitive performance, being able to identify subgroups described by rules with stronger treatment effects and smaller outcome variance, while maintaining similar PEHE and MAPE accuracies.

The evaluation results of the interpretability metrics are shown in Table 6, in which the average length of the rules of *CURLS* is shorter and the overlap rate is small, which helps users to understand. In addition, the coverage rate shows that *CURLS* can find more fine-grained subgroups with significant treatment effects.

Table 6: Results of the interpretability metrics of new baselines, reported as mean and standard deviation.

Dataset	Metrics	CURLS	LDML	DRL	DROF
Syn-data1	Avg_len	2.9 (0.8)	3.7 (0.3)	2.6 (0.2)	3.4 (0.2)
	Overlap(%)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Coverage(%)	6.0 (0.6)	14.5 (2.7)	9.9 (1.5)	10.1 (1.8)
Syn-data2	Avg_len	3.0 (0.0)	3.7 (0.4)	2.5 (0.4)	4.0 (0.0)
	Overlap(%)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Coverage(%)	7.4 (1.0)	19.0 (4.2)	14.7 (3.8)	23.5 (8.3)
Syn-data3	Avg_len	2.8 (0.5)	3.6 (0.2)	2.9 (0.7)	3.4 (0.2)
	Overlap(%)	0.1 (0.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Coverage(%)	12.1 (1.4)	14.6 (3.7)	12.4 (1.8)	12.1 (2.1)
IHDP	Avg_len	3.0 (0.8)	3.8 (0.4)	4.0 (0.0)	4.0 (0.0)
	Overlap(%)	0.7 (0.8)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Coverage(%)	8.3 (1.8)	22.4 (20.8)	22.7 (12.1)	14.8 (4.1)

B THEORETICAL ANALYSIS OF APPROXIMATE BOUNDS

We analyse the approximation bounds for $g(\mathcal{R})$. In Proposition 3, we have shown that $g(\mathcal{R})$ is submodular. Since the parts of $g(\mathcal{R})$ are either increasing (*e.g.*, $Q_{3,\mathcal{R}}$) or decreasing (*e.g.*, $-Q_{2,\mathcal{R}}$) with

Table 5: 5-fold average performance metrics of the new baselines. Numbers in parentheses represent standard deviations.

Dataset	Metrics	CURLS		LDML		DRL		DROF	
		Rule1	Rule2	Rule1	Rule2	Rule1	Rule2	Rule1	Rule2
Syn-data1	CATE↑	10.00(0.65)	8.39(0.25)	8.06(0.27)	7.28(0.15)	8.17(0.20)	7.39(0.37)	9.51(0.47)	8.55(0.21)
	Avg_ITE↑	9.17(0.38)	7.98(0.69)	8.41(0.19)	7.08(0.48)	8.42(0.44)	7.74(0.44)	8.19(0.53)	7.58(0.57)
	Variance↓	9.20(2.24)	9.69(3.08)	9.63(1.50)	10.08(3.08)	9.10(3.33)	9.11(1.78)	9.48(4.16)	9.89(2.01)
	PEHE↓	2.25(0.39)	2.38(0.33)	2.14(0.21)	2.15(0.23)	2.24(0.31)	1.98(0.38)	2.69(0.15)	2.27(0.22)
	MAPE↓	0.23(0.05)	0.28(0.06)	0.21(0.02)	0.29(0.06)	0.23(0.02)	0.22(0.05)	0.35(0.06)	0.30(0.05)
Syn-data2	CATE↑	12.72(0.55)	11.44(0.69)	10.44(0.23)	9.54(0.20)	10.85(0.65)	10.00(0.40)	10.80(0.45)	9.48(0.40)
	Avg_ITE↑	11.23(0.64)	10.53(0.75)	10.19(0.56)	9.52(0.41)	10.66(0.54)	9.86(0.47)	9.47(0.45)	8.91(0.53)
	Variance↓	11.15(2.91)	10.65(3.61)	13.28(3.12)	13.03(1.97)	13.51(1.93)	10.07(1.27)	14.64(4.17)	13.73(1.38)
	PEHE↓	2.64(0.49)	2.13(0.31)	2.23(0.18)	2.20(0.25)	2.15(0.25)	1.97(0.32)	2.67(0.30)	2.26(0.13)
	MAPE↓	0.23(0.06)	0.19(0.03)	0.19(0.03)	0.21(0.03)	0.18(0.02)	0.17(0.04)	0.28(0.06)	0.25(0.04)
Syn-data3	CATE↑	14.06(0.25)	12.70(1.55)	14.39(0.45)	13.30(0.07)	13.08(0.99)	12.37(0.78)	16.02(0.69)	14.61(0.46)
	Avg_ITE↑	13.80(0.61)	12.47(1.91)	13.76(0.42)	12.66(0.27)	13.88(0.46)	13.21(1.45)	13.98(0.62)	12.96(0.45)
	Variance↓	16.18(4.44)	16.36(5.34)	24.67(15.93)	24.54(6.54)	23.22(3.99)	18.32(2.76)	16.45(5.79)	20.25(3.21)
	PEHE↓	2.99(0.19)	3.08(0.38)	3.27(0.48)	3.21(0.29)	3.13(0.34)	3.13(0.62)	3.57(0.50)	3.35(0.40)
	MAPE↓	0.19(0.02)	0.23(0.07)	0.22(0.03)	0.24(0.02)	0.18(0.03)	0.20(0.02)	0.25(0.05)	0.25(0.04)
IHDP	CATE↑	10.98(1.72)	9.98(0.45)	2.15(0.98)	1.29(0.82)	7.31(0.80)	6.52(0.74)	7.70(0.45)	6.61(0.40)
	Avg_ITE↑	8.44(0.70)	8.51(0.71)	5.36(0.69)	3.00(2.98)	7.74(0.82)	7.04(0.61)	8.32(1.21)	7.84(1.23)
	Variance↓	4.17(5.03)	24.06(16.35)	128.78(44.79)	153.16(51.99)	95.34(92.46)	148.74(89.66)	79.55(53.33)	172.09(85.54)
	PEHE↓	10.78(1.22)	10.42(1.57)	8.56(2.32)	12.49(6.05)	9.59(1.08)	9.37(0.87)	10.24(1.61)	9.51(1.89)
	MAPE↓	2.03(0.65)	2.08(0.94)	0.75(0.17)	1.07(0.34)	1.75(0.70)	2.18(1.27)	1.68(0.66)	1.19(0.37)

the addition of covariates, $g(\mathcal{R})$ is not necessarily monotone. There is a $\frac{1}{k+2+\frac{1}{k}+\epsilon}$-approximation bound for non-monotone submodular functions under k matroid constraints [39]. Our problem formulation uses a cardinality constraints $|\mathcal{R}| \leq L$, which can be viewed as a $k = 1$ matroid constraint; hence $g(\mathcal{R})$ has $\frac{1}{4}$-approximation bound.