Cost-Effective and Stable Policy Optimization Algorithm for Uplift Modeling with Multiple Treatments

Yuta Saito *† Hayato Sakata * Kazuhide Nakata †

Abstract

Uplift modeling aims to optimize treatment policies and is a promising method for causal-based personalization in various domains such as medicine and marketing. However, applying this method to real-world problems faces challenges such as the impossibility of validation and binary treatment limitation. The Contextual Treatment Selection (CTS) algorithm was proposed to overcome the binary treatment limitation and demonstrated state-of-the-art results. However, previous experiments have implied that CTS is cost-ineffective because it requires a large amount of training data. In this paper, we demonstrate that the estimator maximized in CTS is biased against the true metric. We then propose a variance reduced estimator based on the doubly robust estimation technique that provides unbiasedness and desirable variance. We further propose a treatment policy optimization algorithm called *VAriance Reduced Treatment Selection* (VARTS), which maximizes our estimator. Empirical experiments on synthetic and real-world datasets demonstrated that our method outperforms other existing methods, particularly under realistic conditions such as small sample sizes and high noise levels. These theoretical and empirical results imply that our method can overcome the critical challenges of uplift modeling and should be the first choice for optimizing personalization in various fields.

1 Introduction

In various real-world problems, choosing the optimal treatment to maximize the profit of interest is crucial [1]. For example, online advertising companies need to deliver the best advertisements to each user to achieve the highest conversion rate [2]. In medicine, the most effective medical treatment should be selected for each patient from numerous medical options [3, 4]. Accurately estimating the causal effects of the treatments is essential to deriving an optimal policy. Conventionally, the average treatment effect (ATE), which is the net effect on the whole population, is used to choose the single best treatment for everyone. ATE is estimated by randomly assigning treatments to subjects through randomized controlled trials (RCTs) and averaging the outcomes within each treatment. However, this sort of treatment policy is not always optimal. For example, the best medical treatment for the entire patient population may have negative side effects on some patients. In other words, finding an optimal personalized treatment for each patient is essential [1, 5].

Uplift modeling is a promising field for optimizing our metric of interest: the expected response of a treatment policy [5, 6]. The aim is to maximize the expected response by using specialized methods that deal with the unobservability of counterfactual outcomes (i.e., outcomes for unobserved treatment assignments) to develop an optimal policy. Several uplift modeling methods have been used to improve the survival rates of breast cancer patients by personalizing their radiotherapy treatment allocations [7] and to raise the revenue of an airline company by optimizing its flight reservation pricing strategy [5].

Despite expectations, most uplift modeling methods [8, 9, 10, 11] are only applicable to binary treatment problems (binary treatment limitation). This limitation is critical because the multiple treatment optimization problem is ubiquitous, such as in the choice of medication. Moreover, accurately evaluating the policy performance is impossible because the counterfactual outcomes are unobservable (impossibility of validation). Thus, a stable model performance across a range of settings is required for real-world applications. However, the variance in the metric estimation or the worst-case performance of policy optimization algorithms have not yet been fully investigated [11]. In summary, a treatment policy optimization algorithm that can directly and stably maximize the expected response in multiple treatment settings is highly desired.

Among the multiple treatment optimization methods, the Separate Model Approach (SMA) is the simplest and commonly used [12, 13]. Predictive models are trained to predict the outcomes under each treatment, and the treatment with the best predictive value of new

*Tokyo Institute of Technology
{saito.y.bj, nakata.k.ac}@m.titech.ac.jp
†SMN Corporation.
{yuta_saito, hayato_sakata}@so-netmedia.jp

data is chosen. SMA is easily implemented because it does not require a specialized algorithm. However, SMA cannot capture causal signals because it does not directly predict causal effects and often overestimates features related to the outcomes but not to the causal effects [14, 15].

To overcome the shortcoming of SMA, Contextual Treatment Selection (CTS) was proposed [5, 6]. CTS directly maximizes the local expected response estimator during the tree construction process and is currently the only algorithm that can handle multiple treatments and continuous outcomes at the same time [5, 6]. However, CTS suffers from high variance of its estimator and a costly data gathering process to ensure adequate performance. This is because it estimates the local expected response by using only factual (i.e., observed) outcomes and ignores counterfactual outcomes. In fact, the experimental results in [5, 6] showed that CTS needs a large amount of training data to be effective. This limits the real-world applicability of CTS.

In this paper, we prove that the local expected response estimator optimized by CTS is actually biased because of the regularization for variance reduction. We then propose a variance reduced local expected response estimator that is based on the doubly robust estimator, which is well-established in the literature on causal inference [16, 17]. The technique is useful for estimating the expected response efficiently but has not been applied to uplift tree methods. Our theoretical analysis showed that our estimator is unbiased and has a smaller variance and tighter estimation error tail bound than the naive one. We further propose a treatment policy optimization algorithm called *VAriance Reduced Treatment Selection* (VARTS) that maximizes our estimator during its learning process. Finally, we conducted extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of the proposed estimator and algorithm.

The contributions of this paper are summarized as follows:

- We investigated the theoretical and empirical properties of CTS and showed that the estimator maximized by the algorithm is actually biased.
- We propose a variance reduced estimator and prove that it is unbiased and achieves a smaller variance and tighter estimation error tail bound.
- We propose the VARTS algorithm maximizing the variance reduced estimator.
- We empirically demonstrated the effectiveness of the proposed estimator and algorithm using both synthetic and real-world datasets.

2 Problem Setting

Here, we formulate the uplift modeling with multiple treatments.

2.1 Notation Given a set of N individuals indexed by i, we define $\boldsymbol{X}_i \in \mathcal{X}$ as the feature vector for each unit. We consider settings where T treatments exist and let $W_i \in \{0, 1, 2, \ldots, T-1\} \in \mathcal{T}$ be a categorical random variable that represents i's treatment assignment. When i receives treatment t, $W_i = t$. We assume that data are gathered through RCTs and that the feature vectors and treatment assignments are statistically independent (i.e., $\boldsymbol{X}_i \perp W_i$, $\forall i \in \{1, 2, \ldots, N\}$). We also use $p^{(t)}$ to represent the treatment assignment probability (i.e., $p^{(t)} = \mathbb{P}(W_i = t)$).

Here, we follow the Rubin causal model [18] and assume that there exist T potential outcomes corresponding to T treatments for each data: $\boldsymbol{Y}_i = (Y_i^{(0)}, Y_i^{(1)}, \ldots, Y_i^{(T-1)}) \in \mathcal{Y}^T$. The fundamental problem of uplift modeling is that only the potential outcome corresponding to the realized treatment can be observed. Let Y_i^{obs} be the observed outcome; then, $Y_i^{\text{obs}} = Y_i^{(t)}$ when $W_i = t$, and the other potential outcomes remain counterfactual.

In addition, we use $\mu_i^{(t)}$ to denote expected potential outcomes for each unit, which means that $\mu_i^{(t)} = \mathbb{E}[Y_i^{(t)} | \boldsymbol{X}_i = \boldsymbol{x}_i]$. We use $\widehat{\mathcal{D}} = \left\{\left(\boldsymbol{x}_i, w_i, y_i^{\text{obs}}\right)\right\}_{i=1}^{N} \overset{iid}{\sim} \mathcal{P}(\boldsymbol{X}, W, \boldsymbol{Y})$ as the empirical distribution of N independent and identically distributed data.

In our theoretical analysis, n_ϕ and $\mathcal{D}_\phi$ represent $\sum_{i=1}^{N} \mathbb{I}\{\boldsymbol{x}_i \in \phi\}$ and $\mathcal{P}(\boldsymbol{X}, W, \boldsymbol{Y} \mid \boldsymbol{X} \in \phi)$, where $\phi \subset \mathcal{X}$ is an arbitrary subset of the feature space.

2.2 Treatment Policy and Expected Response The treatment policy $h(\cdot)$ is a mapping from the feature space to the treatment space. We consider the following expected response as the performance metric of a treatment policy; the main focus of this paper is to propose an algorithm to optimize this metric.

Definition 2.1. (Expected Response) *Given a treatment policy h, the expected response is*

$$V(h) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{Y}}\left[Y^{(h(\boldsymbol{X}))} \mid \boldsymbol{X}\right]\right]$$

The optimal treatment policy h^* is one that outputs the treatment corresponding to the potential outcome with the highest expected value [5]:

$$h^*(\boldsymbol{x}_i) \in \arg\max_{t \in \mathcal{T}} \mathbb{E}\left[Y_i^{(t)} \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right]$$

3 Existing CTS Algorithm

In this section, we discuss the existing state-of-the-art treatment policy optimization algorithm called CTS. The main idea of CTS is to gradually personalize the treatment assignments to maximize the estimated expected response while splitting the feature space.

3.1 Split Criterion of CTS Let $\phi \subset \mathcal{X}$ be a node of a tree; then, we use $\mathcal{S}$ to denote the candidate set of binary splits s that split ϕ into two child nodes: $\phi_l(s) \subset \phi, \phi_r(s) \subset \phi$. Let $V(\phi, t) = \mathbb{E}[Y_i^{(t)} \mid \boldsymbol{X} \in \phi]$ be the local expected response (i.e., expected response in ϕ given t). At each depth during the tree growing process, CTS attempts to find a split $s \in \mathcal{S}$ that leads to the maximum expected response. The optimal binary split is defined as:

$$
\begin{aligned}
s^* = \arg\max_{s \in \mathcal{S}} \; & \mathbb{P}(\boldsymbol{X} \in \phi_l(s) \mid \boldsymbol{X} \in \phi) \times \max_{t_l \in \mathcal{T}} V(\phi_l(s), t_l) \\
& + \mathbb{P}(\boldsymbol{X} \in \phi_r(s) \mid \boldsymbol{X} \in \phi) \times \max_{t_r \in \mathcal{T}} V(\phi_r(s), t_r)
\end{aligned} \tag{3.1}
$$

The main focus of CTS is finding the best split from factual training data. First, the conditional probabilities can be straightforwardly estimated from the given samples:

$$
\hat{p}(\phi' \mid \phi) = \frac{\sum_{i=1}^{N} \mathbb{I}\{\mathbf{x}_i \in \phi'\}}{\sum_{i=1}^{N} \mathbb{I}\{\mathbf{x}_i \in \phi\}}
$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and ϕ' is a child node of ϕ.

On the other hand, the local expected response is not straightforward to estimate because the unobservability of the counterfactuals makes directly observing the realizations of the local expected response impossible. To deal with this problem, CTS utilizes a random variable composed of observable variables:

$$
Z_{\text{naive}}(i, t) = \frac{Y_i^{\text{obs}} \mathbb{I}\{W_i = t\}}{p^{(t)}} \tag{3.2}
$$

Proposition 3.1 shows that the expectation of the random variable in (3.2) is equal to the true local expected response:

PROPOSITION 3.1. *The following holds for any $\phi \subset \mathcal{X}$ and $t \in \mathcal{T}$.*

$$
V(\phi, t) = \mathbb{E}_{\mathcal{D}_\phi}[Z_{naive}(i, t)]
$$

See Appendix A for the proof.

Proposition 3.1 implies that the local expected response can be estimated without bias by averaging the realizations of the random variable in (3.2):

$$
\widehat{V}_{\text{naive}}(\phi, t) = \frac{\sum_{i:\boldsymbol{x}_i \in \phi} y_i^{\text{obs}} \mathbb{I}\{w_i = t\}}{\sum_{i:\boldsymbol{x}_i \in \phi} \mathbb{I}\{w_i = t\}} \tag{3.3}
$$

However, this naive estimator is calculated from only the factual outcome of data assigned to the treatment t. This property implies that this naive estimator may suffer from high variance, especially when a small amount of data is available in the node ϕ.

CTS deals with the variance problem by introducing regularization, which exploits the expected response estimation of the parent node to some extent. The expected response estimator with regularization that CTS actually relies on is

$$
\widehat{V}_{\text{cts}}(\phi', t) = \frac{\sum_{i:\boldsymbol{x}_i \in \phi} y_i^{\text{obs}} \mathbb{I}\{w_i = t\} + n\_reg \times \widehat{V}_{\text{cts}}(\phi, t)}{\sum_{i:\boldsymbol{x}_i \in \phi} \mathbb{I}\{w_i = t\} + n\_reg} \tag{3.4}
$$

where n_reg is a predetermined hyper parameter and ϕ is the parent node of ϕ'.

The split criterion of CTS is defined on the basis of (3.1) and (3.4):

$$
\begin{aligned}
\hat{s} = \arg\max_{s \in \mathcal{S}} \; & \hat{p}(\phi_l(s) \mid \phi) \times \max_{t_l \in \mathcal{T}} \widehat{V}_{\text{cts}}(\phi_l(s), t_l) \\
& + \hat{p}(\phi_r(s) \mid \phi) \times \max_{t_r \in \mathcal{T}} \widehat{V}_{\text{cts}}(\phi_r(s), t_r)
\end{aligned} \tag{3.5}
$$

3.2 Theoretical Analysis of CTS Here, we theoretically analyze the estimator (3.4) that is maximized during the tree growing process of CTS. First, we show that, because of the additional regularization, the estimator (3.4) does not always satisfy unbiasedness, while the naive estimator (3.3) always does.

Suppose that the true local expected response of the parent node is given (i.e., $\widehat{V}_{\text{cts}}(\phi, t) = V(\phi, t)$). The local expected response estimator (3.4) is interpreted as the sample average of the following random variable[1]:

$$
\begin{aligned}
Z_{\text{cts}}(i, t) = & \frac{p^{(t)}}{p^{(t)} + \overline{n\_reg}} \frac{Y_i^{\text{obs}} \mathbb{I}\{W_i = t\}}{p^{(t)}} \\
& + \frac{\overline{n\_reg}}{p^{(t)} + \overline{n\_reg}} V(\phi, t)
\end{aligned} \tag{3.6}
$$

where $\overline{n\_reg}$[2] is a hyper parameter that depends on the child node. This parameter determines how much information of the parent node $Z_{\text{cts}}(i, t)$ is considered when the local expected response of the child node is estimated.

[1] Here, we use the theoretical treatment assignment probability $p^{(t)}$ instead of its empirical estimate $\frac{1}{n_\phi} \sum_{i:\boldsymbol{x}_i \in \phi} \mathbb{I}\{w_i = t\}$.

[2] $\overline{n\_reg} = n\_reg / n_\phi$

The following proposition states that the expectation of (3.6) is not always equal to the true local expected response.

PROPOSITION 3.2. *Suppose* $V(\phi', t) \neq V(\phi, t)$*, then,*

$$V(\phi', t) \neq \mathbb{E}_{\mathcal{D}_{\phi'}}[Z_{cts}(i,t)]$$

See Appendix A for the Proof.

The tree construction of CTS depends on the sample average of (3.6). However, Proposition 3.2 indicates that the local expected response estimator of CTS is biased. Thus, CTS can lead to a suboptimal treatment policy. In addition, the variance of the estimator has not been analyzed; whether the regularization actually reduces the variance of the naive estimator remains unknown.

4 VARTS Algorithm

In this section, we propose a variance reduced local expected response estimator inspired by the doubly robust estimation technique [16, 17]. This estimator exploits missing counterfactuals by predicting them in order to overcome the drawbacks of existing estimators. After a theoretical analysis of our estimator, we present our treatment policy optimization algorithm called VARTS.

4.1 Variance Reduced Expected Response Estimator Our estimator utilizes the counterfactual outcomes predicted by an arbitrary machine learning algorithm before the estimation. Let $\hat{\mu}_i^{(t)}$ be the predicted value of $\mu_i^{(t)}$. We introduce the following random variable:

$$Z_{\text{varts}}(i,t) = \frac{\left(Y_i^{\text{obs}} - \hat{\mu}_i^{(t)}\right)\mathbb{I}\{W_i = t\}}{p^{(t)}} + \hat{\mu}_i^{(t)} \tag{4.7}$$

Theorem 4.1 shows that the expectation of the random variable in (4.7) is equal to the true local expected response.

THEOREM 4.1. *The following holds for any* $\phi \subset \mathcal{X}$ *and* $t \in \mathcal{T}$*.*

$$V(\phi, t) = \mathbb{E}_{\mathcal{D}_\phi}[Z_{varts}(i,t)]$$

See Appendix A for the Proof.

Next, Theorem 4.2 shows that (3.2) and (4.7) have the following variances.

THEOREM 4.2. *The following equations hold:*

$$\mathbb{V}_{\mathcal{D}_\phi}(Z_{naive}(i,t)) = \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}}\left(\mu_i^{(t)} \mid \boldsymbol{X} \in \phi\right) + \frac{1-p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right]$$

$$\mathbb{V}_{\mathcal{D}_\phi}(Z_{varts}(i,t)) = \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}}\left(\mu_i^{(t)} \mid \boldsymbol{X} \in \phi\right) + \frac{1-p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)} - \hat{\mu}_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right]$$

where $\epsilon = \frac{\left(Y_i^{obs} - \mu_i^{(t)}\right)\mathbb{I}\{W_i = t\}}{p^{(t)}}$.

See Appendix A for the proof.

Theorem 4.2 indicates that (4.7) has a smaller variance than (3.2) when the following inequality holds:

$$\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)} - \hat{\mu}_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] < \mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] \tag{4.8}$$

This condition requires that the counterfactual outcome predictions (i.e., $\hat{\mu}_i^{(t)}$) introduced in (4.7) achieve a smaller MSE than the all zero predicted values. We assume that the training data are gathered through RCT, and thus, the distribution of the feature vector conditional on any of the treatments is equal to the marginalized distribution of the feature vector. This means that the predictive models trained with each treatment group can be used to predict the counterfactual outcomes. Simple machine learning algorithms such as elastic net [19] and random forest [20] have enough power to accurately predict counterfactuals and satisfy (4.8). Thus, the inequality is a relatively mild condition.

Our local expected response estimator is calculated with the random variable in (4.7) and is defined as follows:

$$\widehat{V}_{\text{varts}}(\phi, t) = \frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i \in \phi}\left(\frac{\left(Y_i^{\text{obs}} - \hat{\mu}_i^{(t)}\right)\mathbb{I}\{W_i = t\}}{p^{(t)}} + \hat{\mu}_i^{(t)}\right) \tag{4.9}$$

The following corollaries prove that the estimator (4.9) is unbiased for the true expected response and achieves a smaller variance than the naive estimator (3.3) under the mild condition (4.8).

COROLLARY 4.1. (UNBIASEDNESS) *Our local expected response estimator (4.9) is unbiased for the true local expected response:*

$$V(\phi, t) = \mathbb{E}_{\mathcal{D}_\phi}\left[\widehat{V}_{varts}(\phi, t)\right]$$

See Appendix A for the proof.

COROLLARY 4.2. (SMALLER VARIANCE) *Our local expected response estimator (4.9) achieves a smaller variance than the naive estimator (3.3) with (4.8).*

$$\mathbb{V}_{\mathcal{D}_\phi}\left(\widehat{V}_{varts}(\phi, t)\right) < \mathbb{V}_{\mathcal{D}_\phi}\left(\widehat{V}_{naive}(\phi, t)\right)$$

See Appendix A for the Proof.

Finally, we show that our estimator (4.9) achieves a better estimation error tail bound than the naive estimator (3.3).

THEOREM 4.3. (ESTIMATION ERROR TAIL BOUND) *Let* $\{(\boldsymbol{x}_i, (y_i^{(0)}, y_i^{(1)}, \ldots, y_i^{(T-1)}))\}_{i=1}^N$ *be an arbitrary set of* N *independently sampled realizations. Note that* $p^{(t)} = \mathbb{P}(W_i = t)$ *is the probability of* $y_i^{(t)}$ *being observed. Then, for any* $\delta \in (0, 1)$*, the following inequalities hold with the probability of at least* $1 - \delta$*:*

$$\left|\widehat{V}_{naive}(\phi, t) - V(\phi, t)\right| \leq \frac{1}{n_\phi}\sqrt{\frac{\log\frac{2}{\delta}}{2}\sum_{i:\boldsymbol{x}_i\in\phi}\rho_i^2}$$

$$\left|\widehat{V}_{varts}(\phi, t) - V(\phi, t)\right| \leq \frac{1}{n_\phi}\sqrt{\frac{\log\frac{2}{\delta}}{2}\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2}$$

where $\rho_i = \frac{y_i^{obs}}{p^{(t)}}$, $\xi_i = \frac{y_i^{obs} - \hat{\mu}_i^{(t)}}{p^{(t)}}$.

See Appendix A for the Proof.

As stated in Theorem 4.3, if the counterfactual predictions satisfy the condition (4.10), our estimator (4.9) has a tighter estimation error tail bound than the naive estimator (3.3). Condition (4.10) requires that the counterfactual outcome predictions (i.e., $\hat{\mu}_i^{(t)}$) introduced by the random variable in (4.7) achieves a smaller empirical MSE than the all zero predicted values. Thus, it is easy for the condition to hold like (4.8), and our estimator is stable in most cases.

$$\begin{aligned}&\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2 < \sum_{i:\boldsymbol{x}_i\in\phi}\rho_i^2\\ \Leftrightarrow\ &\sum_{i:\boldsymbol{x}_i\in\phi}\left(y_i^{\mathrm{obs}} - \hat{\mu}_i^{(t)}\right)^2 < \sum_{i:\boldsymbol{x}_i\in\phi}\left(y_i^{\mathrm{obs}}\right)^2\end{aligned} \tag{4.10}$$

4.2 Algorithm VARTS decides the split according to the following criterion:

$$\begin{aligned}\hat{s} = \arg\max_{s\in\mathcal{S}}\ &\hat{p}(\phi_l(s)\,|\,\phi)\times\max_{t_l\in\mathcal{T}}\widehat{V}_{\mathrm{varts}}(\phi_l(s), t_l)\\ &+\hat{p}(\phi_r(s)\,|\,\phi)\times\max_{t_r\in\mathcal{T}}\widehat{V}_{\mathrm{varts}}(\phi_r(s), t_r)\end{aligned} \tag{4.11}$$

Similar to CTS [5] and random forest [20], VARTS utilizes a bagging ensemble of trees to mitigate the overfitting. The whole algorithm is shown in Algorithm 1.

5 Synthetic Data Experiment

We experimentally compared VARTS with existing methods using synthetic datasets. Note that the detailed experimental conditions and hyperparameter tuning procedure are presented in Appendix E.

Algorithm 1 VAriance Reduced Treatment Selection

Input: training data: $\widehat{\mathcal{D}}$, number of samples used in each tree: $\mathbf{B}\,(\leqq N)$, number of trees n_{trees}, number of variables to be considered when looking for the best split: **mtry**, maximum depth of a tree: Δ_{depth}, minimum number of samples required to be at a leaf node: $n_{\text{min-leaf}}$, ML algorithm to be used to predict counterfactual outcomes: **base_learner**.

Training:

for $n = 1, \ldots, n_{\mathrm{trees}}$ **do**

- Train **base_learner** to predict y_i^{obs} from the feature vectors with each treatment.
- Draw B samples from $\widehat{\mathcal{D}}$ to create $\widehat{\mathcal{D}}_B$ at random with replacement.
- Build a tree from $\widehat{\mathcal{D}}_B$. At each node, draw **mtry** coordinates at random; then, find the split with the largest local expected response as measured by the splitting criterion (4.11).
- Output a partition of the feature space as represented by the leaf nodes; for each leaf node, estimate the local expected response to each treatment with (4.9).

Prediction:

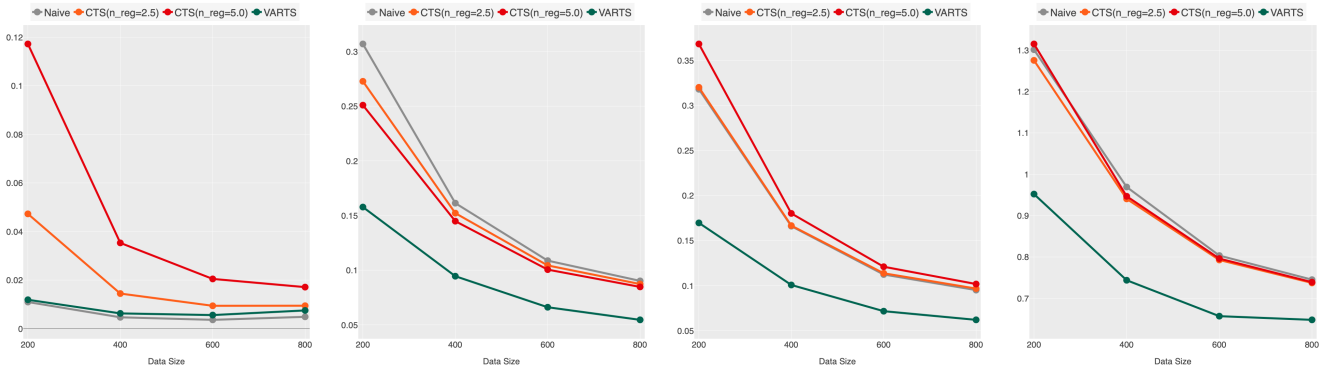
Given a new point in the feature space, the predicted expected response to a treatment is the average of the estimations based on (4.9) from all trees. The optimal treatment is the one with the largest estimated expected response.

5.1 Dataset Generation The synthetic datasets comprised seven scenarios based on the simulation study in [21] with the following elements:

1. The number N of samples in the training set, number p of features, treatment assignment probabilities $p^{(t)}$ for each treatment, and conditional variance σ^2 of Y_i^{obs}. The definitions of these variables depend on the experimental condition in Table 1.
2. The distribution $\mathcal{D}_{\boldsymbol{X}}$ of the feature vectors. Odd-numbered features were drawn independently from a standard Gaussian distribution, and even-numbered features were drawn independently from a Bernoulli distribution with the probability 1/2.

Any policy's performance can be accurately evaluated with synthetic datasets because they give access to all potential outcomes.

5.2 Comparison of Expected Response Estimators We empirically evaluated the bias, variance, MSE,



(a) Bias (b) Variance (c) Mean-Squared Error (d) Estimation Error Bound

Figure 1: Metric estimation performance of expected response estimators.

and estimation error tail bound of the expected response estimators.

5.2.1 Experimental Setup We compared the following estimators: $\widehat{V}_{\text{naive}}$, $\widehat{V}_{\text{cts}}(n\_reg = 2.5, 5.0)$, and $\widehat{V}_{\text{varts}}$. We iterated the steps (1)–(3) below 50 times, and report their bias, variance, MSE, and 99% confidence upper bound of the absolute bias (estimation error tail bound).

1. Generate data ($N = 200$–$800, T = 4, \sigma^2 = 1.0$)
2. Estimate the expected response of all generated data (parent node) with $\widehat{V}_{\text{naive}}$.
3. Estimate the expected response of a subgroup of the generated data ($x_1 \geqq 0$, child node) with the four estimators. Note that the estimation of $\widehat{V}_{\text{cts}}$ depends on the estimation of $\widehat{V}_{\text{naive}}$ in the previous step.

5.2.2 Results The result of the experiment on expected response estimators are reported in Figure 1. We present the results with regard to each performance metric below.

Bias: $\widehat{V}_{\text{naive}}$ and $\widehat{V}_{\text{varts}}$ estimated the expected responses with almost no bias; these results are consistent with the theoretical analysis in Corollary 4.1. On the other hand, $\widehat{V}_{\text{cts}}$ suffered from huge bias, especially when N was small and n_reg was large. The results are also explained by our theoretical results in Proposition 3.2.

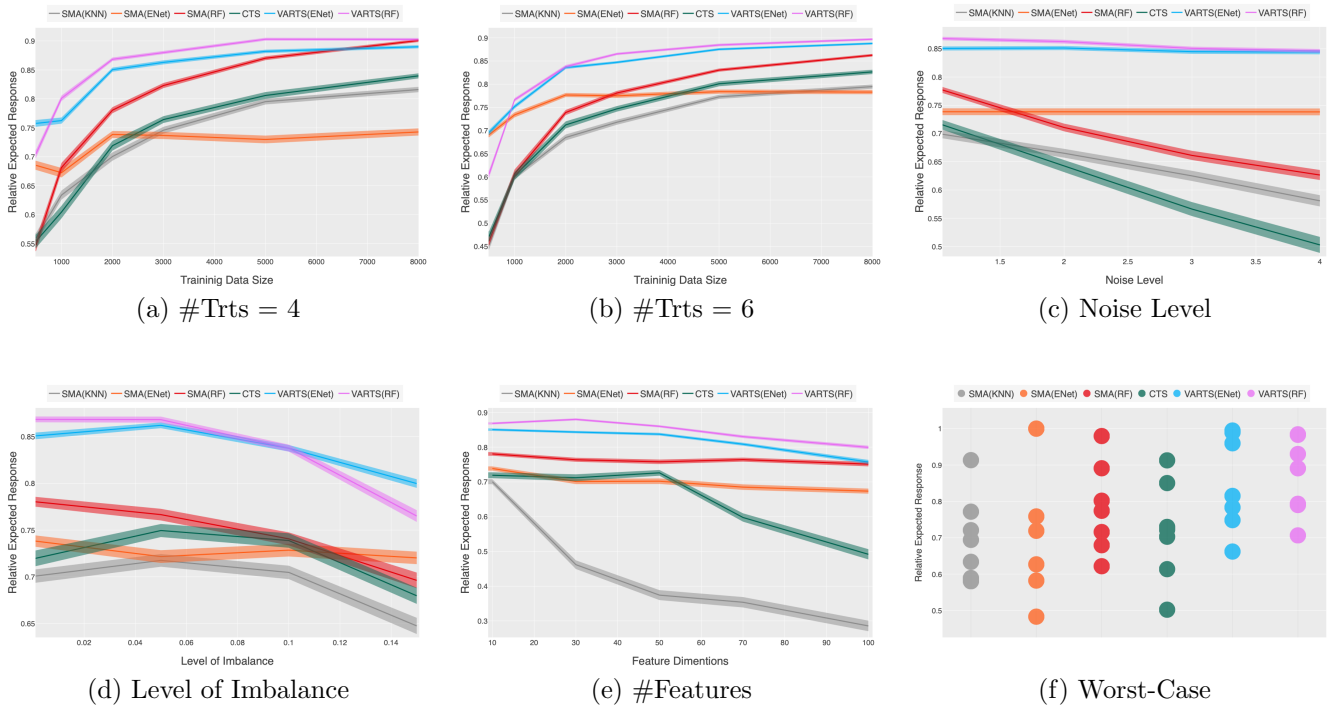
Variance: A larger n_reg led to a smaller variance of $\widehat{V}_{\text{cts}}$ as expected, but the effect of the variance reduction was slight. Our $\widehat{V}_{\text{varts}}$ demonstrated the smallest variance for all data sizes, as guaranteed in Corollary 4.2.

MSE: The overall performance of the expected response estimators was evaluated according to the MSE. Our $\widehat{V}_{\text{varts}}$ performed the best and especially outperformed the other baselines by a large margin when N was small. This is because our estimator greatly reduces the variance of $\widehat{V}_{\text{naive}}$ while retaining its unbiasedness; the results empirically justified the benefits of our estimator. On the other hand, $\widehat{V}_{\text{cts}}$ did not improve the MSE of $\widehat{V}_{\text{naive}}$ because it had to deal with the bias–variance tradeoff depending on the value of n_reg.

Estimation Error Tail Bound: We reported the 99% upper confidence upper bound of the absolute bias for each estimator to empirically test the property discussed in Theorem 4.3. The results suggested that a positive n_reg does not improve the tail bound over the naive estimator. On the other hand, $\widehat{V}_{\text{varts}}$ significantly outperformed the others, which validated the theoretical findings in Theorem 4.3 and empirically emphasized the stability of our estimator.

5.3 Comparison of Treatment Policy Optimization Algorithms We compared the proposed policy optimization algorithms with the baseline algorithms. In particular, we investigated the effects of **#Train**, **noise level**, **level of imbalance** of treatment assignment probabilities, **#features**, and the **worst-case performance**.

5.3.1 Experimental Setup: We compared the following methods: SMA with KNN, SMA with elastic net, SMA with random forest, CTS, VARTS with elastic net, and VARTS with random forest. We iterated the steps (2)–(3) below 50 times, and evaluate the performance in terms of the means and standard deviations



(a) #Trts = 4 (b) #Trts = 6 (c) Noise Level

(d) Level of Imbalance (e) #Features (f) Worst-Case

Figure 2: Performance of each method under each experimental condition.

of the expected responses of methods relative to the best achievable values.

1. Generate training and test data for the corresponding experimental conditions, and train each method.

2. Sub-sample the test data with replacements.

3. Calculate the true expected response of the treatment policy relative to the optimal expected response as the performance metric.

Table 1 presents the experimental conditions. We investigated each method's performance by varying the levels of the factors described in the Table. Note that the level of each factor, except for the one being varied, was fixed at the base level. In addition, we examined the worst-case performance by comparing the worst performances across the seven scenarios.

[3]Noise level is the conditional variance on the observed outcome, i.e., σ^2.

[4]These values represent the additive derivation of the treatment assignment probabilities, i.e., if the level of imbalance is 0.10, the treatment assignment probabilities of the four treatments are $(0.1, 0.2, 0.3, 0.4)$.

Table 1: Experimental Conditions.

factor	base level	levels
#Train	2000	500, 2000, 1000, 3000, 5000, 8000
#Trts	4	4, 6
Noise level[3]	1.0	0.5, 1.0, 1.5, 2.0, 3.0, 4.0
Level of imbalance[4]	0.00	0.0, 0.05, 0.10, 0.15
#Features	10	10, 30, 50, 70, 100
Worst case	Average	Seven Scenarios

5.3.2 Results Here, we present the results of our experiments below.

#Train: Figure 2 (a) and (b) show that VARTS performed the best in almost all situations. In particular, VARTS outperformed SMA(RF) and CTS by a large margin, especially when #Train was small or #Trts was large. Therefore, the benefits of our method are emphasized when the training data size per treatment is quite small. This result is because of the variance reduction effect of $\widehat{V}_{\text{varts}}$. In contrast, the results empirically suggest that the additional regularization of CTS does not always improve the performance, and it requires a large amount of training data to outperform SMAs.

Noise level: Figure 2 (c) shows that VARTS(ENet) and VARTS(RF) consistently outperformed the other methods, especially when the

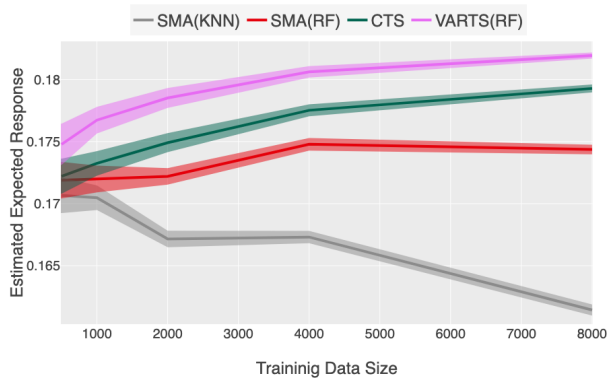


Figure 3: Estimated expected response of each method with the MineThat dataset.

noise level was high. This empirical finding verifies that our VARTS can adjust to a noisy dataset for effective treatment optimization. On the other hand, CTS was negatively affected by the noise level and was outperformed by SMAs when the noise level was high. This implies that CTS may not be applicable to noisy real-world datasets.

Imbalance: As shown in Figure 2 (d), VARTS(ENet) and VARTS(RF) performed stably with small sample treatments and outperformed the others. This result is because our method reduces the variance of its local expected response estimator by utilizing counterfactual predictions; this implies that the treatment assignment probabilities during RCT do not necessarily need to be balanced.

#Features: Figure 2 (e) shows that VARTS performed better than CTS and SMA(RF) even when #features was large. This is because the cardinality of the split set $\mathcal{S}$ increased with #features. CTS's estimator had a large variance or large estimation error. Therefore, there is a greater chance that it chooses a suboptimal split as the number of estimation times increases. On the other hand, VARTS largely removed the effect of the #features because its local expected response estimator was theoretically and empirically proven to achieve a smaller variance and tighter estimation error tail bound.

Worst-case: As shown in Figure 2 (f), VARTS significantly outperformed the other methods with regard to the worst-case performance. This property is essential for real-world applications because accurately evaluating the performance of treatment optimization algorithms is almost impossible in reality, owing to the missing counterfactual outcomes. Therefore, the result demonstrates the stability and applicability of our algorithm.

6 Real-World Experiment

Here we compared the proposed VARTS algorithm with the baseline algorithms using a standard real-world dataset.

6.1 Dataset Description In the experiment, the MineThat Email Campaign Dataset[5] was used. This dataset contains 64000 RCT data of an email advertisement campaign encouraging customers to visit a website of a store. The outcomes were whether or not the customers visited the website [1]. We aimed to optimize the email advertisement allocation and maximize the customers' visits to the website.

6.2 Experimental Setup We compared SMA with KNN, SMA with random forest, CTS, and VARTS with random forest. We conducted 50 simulations with different training/test splits and with different training data sizes. We evaluated the methods in terms of their means and standard errors of the expected response estimated by (3.3) in the test sets.

6.3 Results Here, we present the performance of the methods with the MineThat data. As shown in Figure 3, VARTS performed the best, especially when #Train was small. It obtained an improvement of approximately 2% against the other methods when #Train = 500 and improvements of approximately 13% against SMA(KNN), 4% against SMA(RF), and 1.5% against CTS when #Train = 8000.

Overall, CTS outperformed SMAs; while this result is consistent with those of the previous experiment [5], our method outperformed CTS by a large margin.

7 Conclusion

In this paper, we theoretically show that the previous local expected response estimator used in CTS is biased because of the regularization for variance reduction. To improve the previous method, we propose a variance reduced estimator and a corresponding algorithm. Our theoretical analysis showed that our estimator achieves a smaller variance and tighter estimation error tail bound than the naive one while remaining unbiased for the true expected response. Furthermore, our estimator empirically outperformed the other baselines, especially when the sample size was small. In contrast, a positive value for the n_reg hyperparameter of CTS reduced the variance, but the variance reduction effect was slight and produced a huge bias. Moreover, VARTS demonstrated a state-of-the-art performance for realis-

[5] available at https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html.

tic synthetic datasets with small sample sizes, unbalanced treatment assignments, and high noise levels. In addition, it showed the most stable and best worst-case performance when applied to a range of data generating processes. Finally, experiments on a standard real-world dataset verified the effectiveness and reliability of our method.

As future work, the development of a reliable evaluation metric will be an important topic. VARTS has many hyperparameters, and accurate parameter tuning is required to obtain the optimal policy. An accurate evaluation metric would make VARTS more versatile and improve its performance.

The full version of the paper is available at https://usaito.github.io/files/varts.pdf

References

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

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

[3] Edward Abrahams and Mike Silver. The case for personalized medicine, 2009.

[4] SH Katsanis, Gail Javitt, and Kathy Hudson. A case study of personalized medicine, 2008.

[5] Yan Zhao, Xiao Fang, and David Simchi-Levi. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 588–596. SIAM, 2017.

[6] Yan Zhao, Xiao Fang, and David Simchi-Levi. A practically competitive and provably consistent algorithm for uplift modeling. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1171–1176. IEEE, 2017.

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

[8] Lukasz Zaniewicz and Szymon Jaroszewicz. Support vector machines for uplift modeling. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 131–138. IEEE, 2013.

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

[10] Leo Guelman, Montserrat Guillén, and Ana M Perez-Marin. A survey of personalized treatment models for pricing strategies in insurance. *insurance: Mathematics and Economics*, 58:68–76, 2014.

[11] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. Doubly robust prediction and evaluation methods improve uplift modeling for observational data. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 468–476. SIAM, 2019.

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

[13] Charles Manahan. A proportional hazards approach to campaign list selection. *SAS User Group International (SUGI) 30 Proceedings*, 2005.

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

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

[16] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[17] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[18] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[19] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

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

[21] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

[22] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

Supplementary Material

A Proofs

PROPOSITION 3.1. *The following holds for any $\phi \subset \mathcal{X}$ and $t \in \mathcal{T}$.*

$$V(\phi, t) = \mathbb{E}_{\mathcal{D}_\phi}\left[Z_{naive}(i,t)\right]$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_\phi}\left[Z_{\text{naive}}(i,t)\right] &= \mathbb{E}_{\mathcal{D}_\phi}\left[\frac{Y_i^{\text{obs}}\mathbb{I}\{W_i = t\}}{p^{(t)}}\right] \quad \because (3.2) \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\frac{Y_i^{(t)}}{p^{(t)}}\mathbb{E}_{W_i}\left[\mathbb{I}\{W_i = t\}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}_{\boldsymbol{Y}_i}\left[Y_i^{(t)} \mid \boldsymbol{X}_i \in \phi\right]\right] \\
&= V(\phi, t)
\end{aligned}$$

□

PROPOSITION 3.2. *Suppose $V(\phi', t) \neq V(\phi, t)$, then,*

$$V(\phi', t) \neq \mathbb{E}_{\mathcal{D}_{\phi'}}\left[Z_{cts}(i,t)\right]$$

Proof. Given $V(\phi', t) \neq V(\phi, t)$, there exists a constant value $\mathcal{C} \in \mathbb{R}\backslash\{0\}$ which satisfies the following:

$$V(\phi, t) = V(\phi', t) + \mathcal{C}$$

Hence,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\phi'}}\left[Z_{\text{cts}}(i,t)\right] &= \frac{p^{(t)}}{p^{(t)} + \overline{n_reg}} V(\phi', t) \\
&\quad + \frac{\overline{n_reg}}{p^{(t)} + \overline{n_reg}} V(\phi, t) \\
&= \frac{p^{(t)}}{p^{(t)} + \overline{n_reg}} V(\phi', t) \\
&\quad + \frac{\overline{n_reg}}{p^{(t)} + \overline{n_reg}} \left(V(\phi', t) + \mathcal{C}\right) \\
&\neq V(\phi', t)
\end{aligned}$$

□

THEOREM 4.1. *The following holds for any $\phi \subset \mathcal{X}$ and $t \in \mathcal{T}$.*

$$V(\phi, t) = \mathbb{E}_{\mathcal{D}_\phi}\left[Z_{varts}(i,t)\right]$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_\phi}\left[Z_{\text{varts}}(i,t)\right] \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\frac{\left(Y_i^{\text{obs}} - \hat{\mu}_i^{(t)}\right)\mathbb{I}\{W_i = t\}}{p^{(t)}} + \hat{\mu}_i^{(t)}\right] \quad \because (4.7) \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\frac{Y_i^{(t)} - \hat{\mu}_i^{(t)}}{p^{(t)}}\mathbb{E}_{W_i}\left[\mathbb{I}\{W_i = t\}\right] + \hat{\mu}_i^{(t)}\right] \\
&= \mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}_{\boldsymbol{Y}_i}\left[Y_i^{(t)} \mid \boldsymbol{X}_i \in \phi\right]\right] \\
&= V(\phi, t)
\end{aligned}$$

□

THEOREM 4.2. *The following equations hold:*

$$\begin{aligned}
\mathbb{V}_{\mathcal{D}_\phi}\left(Z_{naive}(i,t)\right) &= \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}}\left(\mu_i^{(t)} \mid \boldsymbol{X} \in \phi\right) \\
&\quad + \frac{1 - p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] \qquad \text{(D.1)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}_{\mathcal{D}_\phi}\left(Z_{varts}(i,t)\right) &= \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}}\left(\mu_i^{(t)} \mid \boldsymbol{X} \in \phi\right) \\
&\quad + \frac{1 - p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)} - \hat{\mu}_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] \qquad \text{(D.2)}
\end{aligned}$$

where $\epsilon = \frac{\left(Y_i^{obs} - \mu_i^{(t)}\right)\mathbb{I}\{W_i = t\}}{p^{(t)}}$.

Proof. We prove (D.2), then use it to prove (D.1). The second moment of the random variable (4.7) is:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_\phi}\left[Z_{\text{varts}}^2(i,t)\right] \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\left(\left(\hat{\mu}_i^{(t)} - \mu_i^{(t)}\right)\left(1 - \frac{\mathbb{I}\{W_i = t\}}{p^{(t)}}\right) + \mu_i^{(t)} + \epsilon\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{E}_{\boldsymbol{X}_i}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] \\
&+ \mathbb{E}_{W_i, \boldsymbol{X}_i}\left[\left(\hat{\mu}_i^{(t)} - \mu_i^{(t)}\right)^2\left(1 - \frac{\mathbb{I}\{W_i = t\}}{p^{(t)}}\right)^2 \mid \boldsymbol{X} \in \phi\right] \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{E}_{\boldsymbol{X}_i}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] \\
&+ \frac{1 - p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}_i}\left[\left(\hat{\mu}_i^{(t)} - \mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right]
\end{aligned}$$

Thus, the variance of (4.7) is:

$$\begin{aligned}
&\mathbb{V}_{\mathcal{D}_\phi}\left(Z_{\text{varts}}(i,t)\right) \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[Z_{\text{varts}}^2(i,t)\right] - \left(\mathbb{E}_{\mathcal{D}_\phi}\left[Z_{\text{varts}}(i,t)\right]\right)^2 \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{E}_{\boldsymbol{X}_i}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] \\
&+ \frac{1 - p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}_i}\left[\left(\mu_i^{(t)} - \hat{\mu}_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right] - \left(\mathbb{E}_{\boldsymbol{X}}\left[\mu_i^{(t)} \mid \boldsymbol{X} \in \phi\right]\right)^2 \\
&= \mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}_i}\left(\mu_i^{(t)} \mid \boldsymbol{X} \in \phi\right) \\
&+ \frac{1 - p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}_i}\left[\left(\mu_i^{(t)} - \hat{\mu}_i^{(t)}\right)^2 \mid \boldsymbol{X} \in \phi\right]
\end{aligned}$$

We have $\mathbb{V}_{\mathcal{D}_\phi}\left(Z_{\text{varts}}(i,t)\right)$ by replacing $\hat{\mu}_i^{(t)}$ with 0. □

COROLLARY 4.1. (UNBIASEDNESS) *Our local expected response estimator (4.9) is unbiased for the true local expected response:*

$$V(\phi, t) = \mathbb{E}_{\mathcal{D}_\phi}\left[\widehat{V}_{varts}(\phi, t)\right]$$

Proof.

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_\phi}\left[\widehat{V}_{\text{varts}}(\phi,t)\right] &= \mathbb{E}_{\mathcal{D}_\phi}\left[\frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i\in\phi} Z_{\text{varts}}(i,t)\right] \\ &= \frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i\in\phi}\mathbb{E}_{\mathcal{D}_\phi}\left[Z_{\text{varts}}(i,t)\right] \\ &= V(\phi,t) \ \because \text{Theorem 4.1}\end{aligned}$$

□

Corollary 4.2. (Smaller variance) *Our local expected response estimator (4.9) achieves a smaller variance than the naive estimator (3.3) with (4.8).*

$$\mathbb{V}_{\mathcal{D}_\phi}\left(\widehat{V}_{varts}(\phi,t)\right) < \mathbb{V}_{\mathcal{D}_\phi}\left(\widehat{V}_{naive}(\phi,t)\right)$$

Proof. Suppose that $Z_{\text{naive}}(i,t)$ and $Z_{\text{varts}}(i,t)$ are independent bounded random variables. Then, we have:

$$\begin{aligned}&\mathbb{V}_{\mathcal{D}_\phi}\left(\frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i\in\phi} Z_{\text{cts}}(i,t)\right) \\ &= \frac{1}{n_\phi^2}\sum_{i:\boldsymbol{x}_i\in\phi}\mathbb{V}_{\mathcal{D}_\phi}\left(Z_{\text{cts}}(i,t)\right) \\ &= \frac{1}{n_\phi}\Big(\mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}}\left(\mu_i^{(t)} \mid \boldsymbol{X}\in\phi\right) \\ &\quad + \frac{1-p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)}\right)^2 \mid \boldsymbol{X}\in\phi\right]\Big) \\ &\quad \because \text{Theorem 4.2}\end{aligned} \tag{D.3}$$

Similarly we obtain:

$$\begin{aligned}&\mathbb{V}_{\mathcal{D}_\phi}\left(\frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i\in\phi} Z_{\text{varts}}(i,t)\right) \\ &= \frac{1}{n_\phi^2}\sum_{i:\boldsymbol{x}_i\in\phi}\mathbb{V}_{\mathcal{D}_\phi}\left(Z_{\text{varts}}(i,t)\right) \\ &= \frac{1}{n_\phi}\Big(\mathbb{E}_{\mathcal{D}_\phi}\left[\epsilon^2\right] + \mathbb{V}_{\boldsymbol{X}}\left(\mu_i^{(t)} \mid \boldsymbol{X}\in\phi\right) \\ &\quad + \frac{1-p^{(t)}}{p^{(t)}}\mathbb{E}_{\boldsymbol{X}}\left[\left(\mu_i^{(t)}-\hat{\mu}_i^{(t)}\right)^2 \mid \boldsymbol{X}\in\phi\right]\Big) \\ &\quad \because \text{Theorem 4.2}\end{aligned}$$

The only difference between $\mathbb{V}_{\mathcal{D}_\phi}\left(\widehat{V}_{\text{varts}}(\phi,t)\right)$ and $\mathbb{V}_{\mathcal{D}_\phi}\left(\widehat{V}_{\text{naive}}(\phi,t)\right)$ is the third term in brackets. Hence, this completes the proof when (4.8) holds. □

Lemma A.1. (Hoeffding's Inequality [22]) *Independent bounded random variables $Z_1,...,Z_n$ that take values in intervals of sizes $\zeta_1,...,\zeta_n$ satisfy the following inequality for any $\epsilon>0$.*

$$\mathbb{P}\left(\left|\sum_{i=1}^n Z_i - \mathbb{E}\left[\sum_{i=1}^n Z_i\right]\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n \zeta_i^2}\right)$$

See Theorem 2 in [22] for the proof.

Theorem 4.3. (Estimation error tail bound) *Let $\{(\boldsymbol{x}_i,(y_i^{(0)},y_i^{(1)},\ldots,y_i^{(T-1)}))\}_{i=1}^N$ be an arbitrary set of N independently sampled realizations. Note that $p^{(t)}=\mathbb{P}(W_i=t)$ is the probability of $y_i^{(t)}$ being observed. Then, for any $\delta\in(0,1)$, the following inequalities hold with the probability of at least $1-\delta$:*

$$\left|\widehat{V}_{naive}(\phi,t)-V(\phi,t)\right| \leq \frac{1}{n_\phi}\sqrt{\frac{\log\frac{2}{\delta}}{2}\sum_{i:\boldsymbol{x}_i\in\phi}\rho_i^2}$$

$$\left|\widehat{V}_{varts}(\phi,t)-V(\phi,t)\right| \leq \frac{1}{n_\phi}\sqrt{\frac{\log\frac{2}{\delta}}{2}\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2}$$

where $\rho_i=\frac{y_i^{obs}}{p^{(t)}}$, $\xi_i=\frac{y_i^{obs}-\hat{\mu}_i^{(t)}}{p^{(t)}}$.

Proof. Given the N independently sampled realizations., we put $Z_i=\frac{\left(y_i^{\text{obs}}-\hat{\mu}_i^{(t)}\right)\mathbb{I}\{W_i=t\}}{p^{(t)}}+\hat{\mu}_i^{(t)}$. Accordingly, we have:

$$\begin{aligned}\mathbb{P}\left(Z_i=\frac{\left(y_i^{\text{obs}}-\hat{\mu}_i^{(t)}\right)}{p^{(t)}}+\hat{\mu}_i^{(t)}\right)=p^{(t)}, \\ \mathbb{P}\left(Z_i=\hat{\mu}_i^{(t)}\right)=1-p^{(t)}\end{aligned}$$

Here, we apply Hoeffding's Inequality in Lemma A.1 to the random variables:

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i\in\phi}Z_i-\mathbb{E}\left[\frac{1}{n_\phi}\sum_{i:\boldsymbol{x}_i\in\phi}Z_i\right]\right|\geq\epsilon\right)\leq 2\exp\left(\frac{-2n_\phi^2\epsilon^2}{\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2}\right) \\ &\Leftrightarrow \mathbb{P}\left(\left|\widehat{V}_{\text{varts}}(\phi,t)-V(\phi,t)\right|\geq\epsilon\right)\leq 2\exp\left(\frac{-2n_\phi^2\epsilon^2}{\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2}\right)\end{aligned}$$

We put $2\exp\left(\frac{-2n_\phi^2\epsilon^2}{\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2}\right)=\delta$ and solve this equation for ϵ, yielding:

$$\mathbb{P}\left(\left|\widehat{V}_{\text{varts}}(\phi,t)-V(\phi,t)\right|\leq\frac{1}{n_\phi}\sqrt{\frac{\log\frac{2}{\delta}}{2}\sum_{i:\boldsymbol{x}_i\in\phi}\xi_i^2}\right)\geq 1-\delta$$

Finally, replacing ξ_i with ρ_i completes the proof. □

	CTS estimator	VARTS estimator
base random variable	(3.6)	(4.7)
unbiasedness	× (Proposition 3.2)	✓ (Corollary 4.1)
smaller variance	N/A	✓ (Corollary 4.2)
tighter error bound	N/A	✓ (Theorem 4.3)

Table 2: Comparison of local expected response estimators. Unbiasedness is with regard to the true local expected response. The variance and error bound are compared with those of the naive estimator.

E Detailed Experimental Setups

E.1 Dataset Generating Procedure The synthetic datasets comprised seven scenarios based on the simulation study in [21] with the following elements:

1. The number N of samples in the training set, number p of features, treatment assignment probabilities $p^{(t)}$ for each treatment, and conditional variance σ^2 of Y_i^{obs}. The definitions of these variables depend on the experimental condition (Table 1).

2. The distribution $\mathcal{D}_{\boldsymbol{X}}$ of the feature vectors. Odd-numbered features were drawn independently from a standard Gaussian distribution, and even-numbered features were drawn independently from a Bernoulli distribution with the probability 1/2.

Given these elements, our data generation model is

$$
\begin{aligned}
&\text{for } n = 1, 2, \ldots, N: \\
&\quad \boldsymbol{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\mathbf{X}}, \\
&\quad w_i = \text{Categorical}\left(\{p^{(0)}, \ldots, p^{(T-1)}\}\right), \\
&\quad y_i^{(t)} = \mu\left(\boldsymbol{X}_i\right) + \tau^{(t)}\left(\boldsymbol{X}_i\right),\ \forall t \in \{0, 1, \ldots, T-1\}, \\
&\quad y_i^{\text{obs}} \sim \text{Normal}\left(y_i^{(w_i)}, \sigma^2\right).
\end{aligned}
$$

E.2 Used Functions We used the following functions for μ and τ:

Group 1 :

$$
\begin{aligned}
f_{1\text{-}1}(\boldsymbol{x}) &= 6\mathbb{I}_{\{x_0>1\}} - 6\text{cdf}(-1), \\
f_{1\text{-}2}(\boldsymbol{x}) &= 6\mathbb{I}_{\{x_2>1\}} - 6\text{cdf}(-1), \\
f_{1\text{-}3}(\boldsymbol{x}) &= 6\mathbb{I}_{\{x_4>1\}} - 6\text{cdf}(-1).
\end{aligned}
$$

Group 2 :

$$
\begin{aligned}
f_{2\text{-}1}(\boldsymbol{x}) &= 5x_0, \\
f_{2\text{-}2}(\boldsymbol{x}) &= 2x_2 - 1, \\
f_{2\text{-}3}(\boldsymbol{x}) &= 4x_4 - 2.
\end{aligned}
$$

Group 3 :

$$
\begin{aligned}
f_{3\text{-}1}(\boldsymbol{x}) &= x_1x_3x_5 + 2x_1x_3(1-x_5) + 3x_1(1-x_3)x_5 \\
&+ 4x_1(1-x_3)(1-x_5) + 5(1-x_1)x_3x_5 + 6(1-x_1)x_3(1-x_5) \\
&+ 7(1-x_1)(1-x_3)x_5 + 8(1-x_1)(1-x_3)(1-x_5) - 4.5, \\
f_{3\text{-}2}(\boldsymbol{x}) &= 3x_1x_3x_5 - 2x_1x_3(1-x_5) + 5x_1(1-x_3)x_5 \\
&- 3x_1(1-x_3)(1-x_5) - (1-x_1)x_3x_5 + 7(1-x_1)x_3(1-x_5) \\
&+ (1-x_1)(1-x_3)x_5 - 4(1-x_1)(1-x_3)(1-x_5) - 0.5, \\
f_{3\text{-}3}(\boldsymbol{x}) &= 2x_1x_3x_5 + 4x_1x_3(1-x_5) - 3x_1(1-x_3)x_5 \\
&+ 2x_1(1-x_3)(1-x_5) + 2(1-x_1)x_3x_5 + (1-x_1)x_3(1-x_5) \\
&- 3(1-x_1)(1-x_3)x_5 - (1-x_1)(1-x_3)(1-x_5) - 0.5.
\end{aligned}
$$

Group 4 :

$$
\begin{aligned}
f_{4\text{-}1}(\boldsymbol{x}) &= x_0 + x_2 + x_4 + x_6 + x_7 + x_8 - 0.5, \\
f_{4\text{-}2}(\boldsymbol{x}) &= x_0 - x_2 + x_4 + x_5 - x_6 - x_8, \\
f_{4\text{-}3}(\boldsymbol{x}) &= x_0 - x_2 + x_3 + x_4 - x_6 - x_8.
\end{aligned}
$$

Group 5 :

$$
\begin{aligned}
f_{5\text{-}1}(\boldsymbol{x}) &= 4\mathbb{I}_{\{x_0>1\}}\mathbb{I}_{\{x_2>0\}} + 4\mathbb{I}_{\{x_4>1\}}\mathbb{I}_{\{x_6>0\}} + 2x_7x_8 - 4\text{cdf}(-1), \\
f_{5\text{-}2}(\boldsymbol{x}) &= 4\mathbb{I}_{\{x_2>1\}}\mathbb{I}_{\{x_4>0\}} + 4\mathbb{I}_{\{x_6>1\}}\mathbb{I}_{\{x_8>0\}} + 2x_4x_5 - 8\text{cdf}(-1), \\
f_{5\text{-}3}(\boldsymbol{x}) &= 4\mathbb{I}_{\{x_0>1\}}\mathbb{I}_{\{x_8>0\}} + 4\mathbb{I}_{\{x_2>1\}}\mathbb{I}_{\{x_6>0\}} + 2x_2x_3 - 8\text{cdf}(-1).
\end{aligned}
$$

Group 6 :

$$
\begin{aligned}
f_{6\text{-}1}(\boldsymbol{x}) &= \frac{1}{\sqrt{2}}\left(x_0^2 + x_1 + x_2^2 + x_3 + x_4^2 + x_5 + x_6^2 + x_7 + x_8^2 - 7\right), \\
f_{6\text{-}2}(\boldsymbol{x}) &= \frac{1}{\sqrt{2}}\left(2x_1 + x_2^2 + 2x_3 + 2x_5 + x_6^2 + 2x_7 - 5.5\right), \\
f_{6\text{-}3}(\boldsymbol{x}) &= \frac{1}{\sqrt{2}}\left(x_0^2 + 4x_1 + 4x_3 + x_4^2 + 4x_5 + x_6^2 + 4x_7 + x_8^2 - 11\right).
\end{aligned}
$$

Group 7 :

$$
\begin{aligned}
f_{7\text{-}1}(\boldsymbol{x}) &= \frac{1}{\sqrt{2}}\left(f_{3-1}(\boldsymbol{x}) + f_{4-1}(\boldsymbol{x})\right), \\
f_{7\text{-}2}(\boldsymbol{x}) &= \frac{1}{\sqrt{2}}\left(f_{3-2}(\boldsymbol{x}) + f_{4-2}(\boldsymbol{x})\right), \\
f_{7\text{-}3}(\boldsymbol{x}) &= \frac{1}{\sqrt{2}}\left(f_{3-3}(\boldsymbol{x}) + f_{4-3}(\boldsymbol{x})\right).
\end{aligned}
$$

where $\boldsymbol{x} = [x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9]^\top$, and $\text{cdf}(\cdot)$ is the cumulative distribution function of a stan-

dard Gaussian distribution.

Table 3: Specifications of the seven scenarios.

Scenarios	1	2	3	4	5	6	7
$\mu(\cdot)$	$f_{4\text{-}1}$	$f_{3\text{-}1}$	$f_{6\text{-}1}$	$f_{2\text{-}1}$	0	$f_{1\text{-}1}$	$f_{5\text{-}1}$
$\tau^{(0)}(\cdot)$	$-f_{1\text{-}1}$	$-f_{2\text{-}1}$	$-f_{3\text{-}1}$	$-f_{4\text{-}1}$	$-f_{5\text{-}1}$	$-f_{6\text{-}1}$	$-f_{7\text{-}1}$
$\tau^{(1)}(\cdot)$	$f_{1\text{-}1}$	$f_{2\text{-}1}$	$f_{3\text{-}1}$	$f_{4\text{-}1}$	$f_{5\text{-}1}$	$f_{6\text{-}1}$	$f_{7\text{-}1}$
$\tau^{(2)}(\cdot)$	$-f_{1\text{-}2}$	$-f_{2\text{-}2}$	$-f_{3\text{-}2}$	$-f_{4\text{-}2}$	$-f_{5\text{-}2}$	$-f_{6\text{-}2}$	$-f_{7\text{-}2}$
$\tau^{(3)}(\cdot)$	$f_{1\text{-}2}$	$f_{2\text{-}2}$	$f_{3\text{-}2}$	$f_{4\text{-}2}$	$f_{5\text{-}2}$	$f_{6\text{-}2}$	$f_{7\text{-}2}$
$\tau^{(4)}(\cdot)$	$-f_{1\text{-}3}$	$-f_{2\text{-}3}$	$-f_{3\text{-}3}$	$-f_{4\text{-}3}$	$-f_{5\text{-}3}$	$-f_{6\text{-}3}$	$-f_{7\text{-}3}$
$\tau^{(5)}(\cdot)$	$f_{1\text{-}3}$	$f_{2\text{-}3}$	$f_{3\text{-}3}$	$f_{4\text{-}3}$	$f_{5\text{-}3}$	$f_{6\text{-}3}$	$f_{7\text{-}3}$

E.3 Hyper-parameter Selection Table 4 and 5 summarize the searching space of the hyper-parameters for each dataset. Standard methods of hyper-parameter tuning, such as cross-validation, are not directly applicable to uplift modeling because the real-world problems have a realization from only one potential outcome. Therefore, we used only the feature vectors ($\boldsymbol{x}_i$), the observed outcome (y_i^{obs}), and the treatment assignment (w_i) during the parameter-tuning procedure, because all of them can be used in real-world settings. The set of hyper-parameters maximizing the estimated expected response on the validation data was selected.

Note that for the synthetic experiment, we relied on general 3fold cross-validation for the parameter selection. On the other hand, for the real-world datasets, we used the software, *Optuna*[6]. n_estimators and min_samples_leaf of the algorithms based on the tree structure (i.e., SMA(RF), CTS, and VARTS) were fixed at 100 and 10, respectively for all of the datasets.

Table 4: Hyper-parameter searching space for the synthetic data.

methods	tuned parameter	space
SMA(KNN)	n_neighbors	$\{10, 15, 20, 25\}$
SMA(ENet)	l2-regularization parameter	$\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$
SMA(RF)	max_depth	$\{5, 10, 15, 20\}$
CTS	max_depth	$\{5, 10, 15, 20\}$
	n_reg	{0, 1, 2, 3 }
VARTS	max_depth	$\{5, 10, 15, 20\}$

Table 5: Hyper-parameter searching space for the MineThat data.

methods	tuned parameter	space
SMA(KNN)	n_neighbors	[20, 50]
SMA(RF)	max_depth	[5, 20]
CTS	max_depth	[5, 20]
	n_reg	{0, 1, 2, 3 }
VARTS(RF)	max_depth of VARTS	[5, 20]
	max_depth of RF	[5, 20]

[6] https://optuna.org