Proceedings of Machine Learning Research vol 140:1–41, 2022 1st Conference on Causal Learning and Reasoning

Evidence-Based Policy Learning

Author names withheld

Editor: Under Review for CLeaR 2022

Abstract

The past years have seen the development and deployment of machine-learning algorithms to estimate personalized treatment-assignment policies from randomized controlled trials. Yet such algorithms typically optimize expected outcomes without taking into account that treatment assignments are frequently subject to hypothesis testing. In this article, we explicitly take significance testing of the effect of treatment-assignment policies into account, and consider assignments that optimize the probability of finding a subset of individuals with a statistically significant positive treatment effect. We provide an efficient implementation using decision trees, and demonstrate its gain over selecting subsets based on positive (estimated) treatment effects. Compared to standard tree-based regression and classification tools, this approach tends to yield substantially higher power in detecting subgroups affected by the treatment.

Keywords: Policy learning, heterogeneous treatment effects, randomized experiments, recursive partitioning, causal machine learning.

1. Introduction

Recent years have seen the development of machine-learning algorithms that estimate heterogeneous causal effects from randomized controlled trials. While the estimation of average effects – for example, how effective a vaccine is overall, whether a conditional cash transfer reduces poverty, or which ad leads to more clicks – can inform the decision whether to deploy a treatment or not, heterogeneous treatment effect estimation allows us to decide who should get treated.

Existing algorithms for the assignment of treatment typically optimize welfare without taking hypothesis testing into account. These algorithms aim to maximize realized outcomes, and thus focus on assigning treatment to individuals with positive (estimated) treatment effects. Yet in practice, the deployment of assignment policies often only happens after passing a test that shows the ass ignment produces a positive net effect relative to some status quo. For example, a drug manufacturer may have to demonstrate that the drug is effective on the target population by submitting a hypothesis test to the FDA for approval. When target groups differ not only in their treatment effects, but also in how precisely these treatment effects can be measured in a trial, then this additional testing step may affect the optimal choice of treatment assignment.

In this article, we explicitly take significance testing of the effect of treatment-assignment policies into account. In our baseline setting, we assume that a policy obtained from training data only gets implemented when a hypothesis test on additional hold-out data shows a significant positive effect relative to not assigning anybody to treatment. In the drug example, the manufacturer decides based on a training sample for which subset of individuals to apply for approval, and subsequently submits additional data for that subpopulation to the FDA.

In the main article, we optimize for the *testing goal* of maximizing the probability of passing the test. Such a goal may correspond to a researcher who wants to obtain significance to get published, or the drug manufacturer seeking approval. Under a Normal large-sample approximation, we show

© 2022 .

that this goal is equivalent to maximizing the ratio of mean to standard error of treatment effect estimates in the hold-out data. When conditional variances differ across covariates, then the oracle solution to this goal can differ markedly from an oracle assignment based on which individuals have predictably positive treatment effect.

We discuss an implementation based on greedy optimization of sample analogues, and present an example algorithm based on decision trees. We show that empirical optimization of our target goal approximates an optimal assignment in finites samples, and present approximation bounds for finite support and under a sparsity assumptions on conditional treatment effects and conditional variances.

We illustrate theoretically the benefit of the evidence-based approach by proving that an algorithm that splits groups empirically according to our main target criterion outperforms a naive classification approach when observations vary considerably in their predictable noise.

In a simulation study, we compare the performance of our algorithm that optimizes for this testing goal to tree-based algorithms that target maximal average outcome by assigning based on positive estimated treatment effects. We demonstrate the improvement in performance with respect to our main target metric, and highlight cases in which such an assignment policy tests better than a simple overall t-test in the full training and hold-out sample combined.

Related work. Learning personalized policies from experimental or observational data has been extensively studied in the recent machine learning literature (Dudík et al., 2011; Swaminathan and Joachims, 2015; Nie and Wager, 2017; Kallus and Zhou, 2018). A large volume of works use ideas from the doubly robust literature in semi-parametric inference (Rubin and van der Laan, 2005, 2007; Díaz and van der Laan, 2013; van der Laan and Luedtke, 2014; Kennedy et al., 2017, 2019) so as to construct objectives to be optimized over a space of policies. However, the majority of these works optimize the mean policy value and do not focus on the objective of succeeding in a subsequent test on a fresh sample. In the experimental section we compare to a predominant such approach.

Our work also builds upon work on recursive partitioning and treatment effect estimation in the causal machine learning literature Künzel et al. (2019); Athey and Imbens (2016); Athey et al. (2019); Nie and Wager (2017); Athey and Wager (2017); Zhou et al. (2018); Oprescu et al. (2019); Friedberg et al. (2018); Chernozhukov et al. (2017, 2018b); Foster and Syrgkanis (2019); Bertsimas et al. (2019). We compare to some predominant methods from this line of work in our experiments. This line of work uses data-adaptive machine learning procedures to learn how the effect of a treatment varies with observable characteristics, even when these characteristics are high-dimensional. Unlike the treatment effect literature, we do not care about learning an accurate model of how the effect varies, but rather only learn a model of who-to-treat. Thus we solely need to understand for which subset of the population the treatment effect is positive. Moreover, the majority of the heterogeneous treatment effect literature constructs sub-groups mainly based on some mean heterogeneity criterion. Arguably closest to us is the work of Powers et al. (2018) which analyze a tree algorithm whose criterion uses a standardized heterogeneity of the form $(\tau(S_1)-\tau(S_2))^2/(\sigma(S_1)^2+\sigma(S_2)^2)$ (already proposed in Athey and Imbens (2016), where S_1, S_2 is a candidate split, $\tau(S_i)$ is the mean effect on each sub-group and $\sigma(S_i)$ the standard deviation. However, this criterion tries to identify the modes of heterogeneity and does not target the policy problem directly. Thus it might be creating very fine partitions in regions of the space where the effect is definitely positive. Such sub-grouping is irrelevant from the perspective of policy learning, where all we care about is the sign of the effect.

Our work also relates to recent work on false discovery rate control in multiple and adaptive hypothesis testing Ramdas et al. (2019); Lei et al. (2017). Unlike our setting, the FDR literature typically considers the statistically more daunting task of testing on the same data that was used to construct the adaptive hypothesis. Here we consider a sample splitting setting, where testing will be constructed on a separate data set. Moreover, unlike the FDR literature we construct algorithms for optimizing the out-of-sample power of the test. Arguably closest to our work from the FDR literature is the concurrent work of Duan et al. (2021), which instantiates the FDR line of work on the heterogeneous treatment effect problem and presents an interactive algorithm that controls the false discovery rate of the constructed sub-group to be treated. On the contrary our work focuses on finding an assignment in a non-interactive manner, which maximizes the power of a test on a separate sample. The work of Armstrong and Shen (2015) considers inference on optimal treatment assignments, where the goal is to obtain a subgroup for which all conditional average treatment effects are positive with high confidence. Relative to this work, we instead provide a subgroup that aims to maximize evidence for a positive average subgroup effect on hold-out data, thus sidestepping the multiple-testing challenge in Armstrong and Shen (2015).

Our setup and motivation is very closely related to concurrent and independent work by Talisa and Chang (2021), who also consider the objective of finding subgroups with statistically significant treatment effects. Relative to their approach, we construct recursive-partitioning and model-based solutions that directly optimize for maximizing power. Specifically, we do not only estimate a power-maximizing assignment given a partition of the covariate space, but we also develop methods for optimizing the partition itself for the power criterion. Relatedly, McFowland et al. (2018) consider the problem of identifying most-affected subgroups, which they operationalize as a problem of detecting anomalous patterns. In a similar spirit, Ladhania et al. (2020) consider the selection of subgroups with consistently large treatment effects, which are then tested on separate data.

There has also been recent work on providing robust versions of off-policy learning, which is similar in spirit to our motivation. The work of Dwivedi et al. (2020), presents a stability based objective for estimating sub-group effects and for creating sub-groups with treatment effect heterogeneity, that are stable (i.e. do not vary a lot) and are well-calibrated out of sample, i.e. the average effect of sub-groups for which we declared a large effect has to also be large on out-of-sample. Finally, the concurrent work of Leqi and Kennedy (2021) provides a way of estimating assignment policies that maximize the median policy value, instead of the expected policy value, which typically is more robust to outlier samples. Both of these works could lead to better out-of-sample power of statistical tests, since they lead to more robust sub-group heterogeneity identification. Unlike these works, our work directly targets the out-of-sample power as an objective.

Structure of the article. Section 2 introduces our assumptions about the data, and formalizes different goals when finding treatment assignment policies. In Section 3, we characterize the (infeasible) theoretically optimal solution to the main testing goal, before we feasible implementations in Section 4. We illustrate the utility of this approach theoretically in Section A in the appendix and in a simulation study with hybrid data in Section B. Section C discusses extensions.

2. Setup and Goal

We consider the problem of estimating an assignment policy $a : \mathcal{X} \to \{0, 1\}$ that maps a covariate $x \in \mathcal{X}$ to an assignment $a(x) \in \{0, 1\}$ of being assigned to the treatment group. For a population distribution of $(Y^{(1)}, Y^{(0)}, X)$ with potential outcomes $Y^{(1)}, Y^{(0)}$ and covariates X, we consider

the expected lift

$$u(a) := \mathrm{E}[a(X)\, Y^{(1)} + (1 - a(X))\, Y^{(0)}] - \mathrm{E}[Y^{(0)}] = \mathrm{E}[a(X)\, \tau(X)]$$

under assignment according to a relative to the status quo of assigning everybody to control, where

$$\tau(x) := \mathrm{E}[Y^{(1)} - Y^{(0)} | X = x]$$

denotes the conditional average treatment effect. We learn an assignment policy from a randomized trial with n iid draws (Y_i, W_i, X_i), where treatment $W_i \in \{0, 1\}$ is assigned randomly with probability $\mathrm{E}[W_i] = p \in (0, 1)$, $Y_i = Y_i(W_i)$, and $(Y_i^{(1)}, Y_i^{(0)}, X_i)$ follow the population distribution. Randomization of W_i allows us to identify $\tau(X)$ from the data.

An assignment maximizing the expected lift $u(a)$ is $\tilde{a}(x) := \mathbb{1}\{\tau(X) > 0\}$, suggesting two related standard strategies to estimate an assignment that maximize $u(a)$: First, estimating heterogeneous treatment effects $\tau(x)$ directly by $\hat{\tau}(x)$ and then assigning based on $\hat{a}(x) = \mathbb{1}\{\hat{\tau}(x) > 0\}$, or second, performing cost-sensitive classification, so as to maximize the objective $\frac{1}{n}\sum_{i=1}^{n} a(X_i)\tilde{Y}_i$, based on an un-biased estimate of $Y_i(1) - Y_i(0)$ for each sample i, such as $\tilde{Y} = \frac{W-p}{p(1-p)}(Y - c(X))$ for any centering function c.

Yet in practice, after estimating an assignment $\hat{a}$ that determines who should get treated, we often have to validate that the assignment does indeed improve the outcome $u(\hat{a})$ relative to the status quo. A standard tool for such validation – and frequent constraint to deployment – is a test for statistical significance against the null hypothesis that $u(\hat{a}) \leq 0$. For example, a drug company may only get a drug approved for a subgroup of the population after passing a statistical test that provides evidence of the effectiveness of the drug against a placebo condition. Rather than arguing in favor or against such hypothesis testing, we take it as a given constraint for deployment.

We therefore assume that there is an additional hold-out dataset with N draws (Y_j, W_j, X_j) from the same distribution as the experiment available, on which we perform a hypothesis test of the form

$$\hat{T}(\hat{a}) = \mathbb{1}\left\{ \frac{\hat{u}(\hat{a})}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(\hat{a}))}} \geq z_{1-\alpha} \right\},$$

where $\hat{u}(\hat{a})$ is an estimate of $u(\hat{a})$ and $\sqrt{\widehat{\mathrm{Var}}(\hat{u}(\hat{a}))}$ is an estimate of the measurement noise $\mathrm{Var}(\hat{u}(\hat{a})|\hat{a})$ from the hold-out data, and $z_{1-\alpha}$ is the critical value of a one-sided Normal test with significance level α (that is, the $1 - \alpha$-quantile of a standard Normal).

In this article, we focus on estimators based on the pseudo-outcomes $\tilde{Y}_j = (Y_j - c(X_j))\left(\frac{W_j}{p} - \frac{1-W_j}{1-p}\right)$ for the propensity $p = \mathrm{E}[W]$ and the centered outcome $Y_j - c(X_j)$, yielding

$$\hat{u}(\hat{a}) = \frac{1}{N}\sum_{j=1}^{N} \hat{a}(X_j)\, \tilde{Y}_j = \frac{1}{N}\sum_{j=1}^{N} \hat{a}(X_j)\, (Y_j - c(X_j))\left(\frac{W_j}{p} - \frac{1 - W_j}{1 - p}\right)$$

$$\widehat{\mathrm{Var}}(\hat{u}(\hat{a})) = \frac{1}{N(N-1)}\sum_{j=1}^{N} (\hat{a}(X_j)\, \tilde{Y}_j - \hat{u}(\hat{a}))^2.$$

Note that $\mathrm{E}[\tilde{Y}_j|\hat{a}] = \mathrm{E}[\hat{a}(X)(Y^{(1)}-Y^{(0)})|\hat{a}] = u(\hat{a})$ irrespective of the centering c, but the centering can reduce the variance. In practice, c may be estimated on the training sample. We will justify a specific choice of c in Section 4 below. Similarly, we may replace the true propensity p of treatment by its sample analogue $\hat{p}(x) = \frac{\sum_{j=1}^N W_j}{N}$, in which case we obtain the difference in averages

$$\hat{u}(\hat{a}) = \frac{1}{\sum_{j=1}^N W_j} \sum_{j=1}^N W_j \, \hat{a}(X_j) \, (Y_j - c(X_j)) - \frac{1}{\sum_{j=1}^N (1-W_j)} \sum_{j=1}^N (1-W_j) \, \hat{a}(X_j) \, (Y_j - c(X_j)).$$

We obtain straightforward extensions when propensity scores are non-constant or have to be estimated, and for alternative outcomes $\hat{Y}_j$ (see Section C).

The test above is justified under standard regularity conditions by

$$\frac{\hat{u}(\hat{a})}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(\hat{a}))}} \xrightarrow{d} \mathcal{N}(0,1)$$

under the null hypothesis $u(\hat{a}) = 0$, where we assume throughout that $\hat{a}$ assigns a sufficiently large part of the sample to treatment so the limit remains well-behaved.

In the main article, we consider the goal of maximizing the probability of passing the test, so we want to choose an assignment $\hat{a}$ with maximal expected probability

$$U(\hat{a}) := \mathrm{E}[\hat{T}(\hat{a})|\hat{a}] = \mathrm{P}\left(\frac{\hat{u}(\hat{a})}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(\hat{a}))}} \geq z_{1-\alpha} \middle| \hat{a} \right),$$

where expectation and probability are over the hold-out sample. For an assignment that assigns nobody to treatment, we can in practice assume that p-values are uniformly distributed, yielding baseline utility $U(\hat{a}) = \alpha$. This goal would correspond, for example, to the goal of a drug company to get a drug approved, or the goal of a researcher to find a subset with a significantly positive treatment effect. (By splitting the sample, optimizing for such a goal does not distort inference as long as the hold-out data is only used for the test.)

3. Characterization of Oracle Policies

In the previous section, we motivated the objective function

$$U(a) = \mathrm{E}[\hat{T}(a)] = \mathrm{P}\left(\frac{\hat{u}(a)}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(a))}} \geq z_{1-\alpha} \right)$$

for the expected lift $u(a) = \mathrm{E}[a(X) \, \tau(X)]$ of the assignment. In this section, we solve for their theoretically optimal (but generally infeasible) oracle maximizers when we apply a large-sample approximation for the training sample.

Expanding on the above notation, we write

$$\tau(x) := \mathrm{E}[Y^{(1)} - Y^{(0)}|X = x] = \mathrm{E}\left[\left(\frac{W}{p} - \frac{1-W}{1-p} \right) (Y - c(X)) \middle| X = x \right],$$

$$\sigma^2(x) := \mathrm{Var}\left(\left(\frac{W}{p} - \frac{1-W}{1-p} \right) (Y - c(X)) \middle| X = x \right),$$

where $Y - c(X)$ is the centered outcome (and we assume wlog that c is chosen such that $\mathrm{E}[Y - c(X)] = 0$ in the case of using empirical propensity). We focus on the case where treatment effects $\tau(x)$ are small, $\tau(x) = \frac{\mu(x)}{\sqrt{N}}$. Assuming that treatment effects are of the order of magnitude $\frac{1}{\sqrt{N}}$ allows us to model non-trivial testing in the limit, and expresses the use case where data is so noisy or effects so small that testing is meaningful. In Section 4 below, we extend our approach beyond this asymptotic framework and instead provide finite-sample results.

In the local-to-zero limiting regime and for the estimators outlined above, for any fixed rule a with $\mathrm{E}[a(X)] > 0$ we have that

$$\frac{\hat{u}(a)}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(a))}} \xrightarrow{d} \mathcal{N}\left(\frac{\mathrm{E}[a(X)\,\mu(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}}, 1\right).$$

Since the asymptotic power of the resulting test is thus

$$\mathrm{P}\left(\mathcal{N}\left(t(a), 1\right) > z_{1-\alpha}\right) = \Phi(t(a) - z_{1-\alpha}), \qquad t(a) = \frac{\mathrm{E}[a(X)\,\mu(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}}$$

(with the Normal cdf Φ), maximizing asymptotic power corresponds to maximizing

$$t(a) := \frac{\mathrm{E}[a(X)\,\mu(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}} = \sqrt{N}\frac{\mathrm{E}[a(X)\,\tau(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}}.$$

Theorem 1 (Oracle policy characterization) *Asymptotically as $N \to 0$ and under standard regularity assumptions, the policy that maximize the testing objective solves*

$$a_* = \arg\max_a U(a) = \arg\max_a \frac{\mathrm{E}[a(X)\,\tau(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}}$$

when there are positive treatment effects, and otherwise $a_* \equiv 0$. Note that the optimizer does not depend on the hold-out size N, or the size α of the test.

The result holds more generally for continuous relaxations of the assignment problem, to which we return in Section C. In the binary case we study here, $a^2(x) = a(x)$, so the criterion becomes $\mathrm{E}[a(X)\,\tau(X)]/\sqrt{\mathrm{E}[a(X)\,\sigma^2(X)]}$.

4. Implementation Using Empirical Optimization of Sample Analogues

In Section 3, we characterized an oracle solution for the goal of finding an assignment that optimizes for passing a hypothesis test on hold-out data. In principle, one could use any estimation method of treatment effects $\tau(x)$ and variances $\sigma^2(x)$ on the training data to estimate such an assignment. In this section, we study implementations that optimize the sample analogue of the criterion in Theorem 1 in the training data.

Specifically, we now drop the assumption that treatment effects are local to zero, and consider finding an assignment policy a that approximates a maximizer of the population t-score

$$\frac{\mathrm{E}[\hat{u}(a)]}{\sqrt{\mathrm{Var}(\hat{u}(a))}}$$

by maximizing its empirical analogue in the training sample, which is the usual in-sample t-score. We formalize this task below before discussing solutions to the empirical optimization problem and their properties.

4.1. Empirical and Population Objectives

When given access to n samples $\{X_i, Y_i, W_i\}_{i=1}^n$ of features $X_i \in \mathcal{X}$, outcome $Y_i \in \mathbb{R}$ and treatment $W_i \in \{0, 1\}$. We can estimate the mean value of any assignment policy $a : \mathcal{X} \to \{0, 1\}$ in an un-biased manner using the inverse-propensity scoring approach. If we let $p = \mathrm{P}(W = 1)$, be the probability of a treatment assignment in the randomized control trial, then we have that an un-biased estimate of the value of a policy is

$$V_{IPS}(a) := \frac{1}{n} \sum_{i=1}^{n} a(X_i)\, Y_i \left(\frac{W_i}{p} - \frac{1 - W_i}{1 - p} \right). \tag{1}$$

Even though it is un-biased, the latter estimate typically has high variance. Hence, we can typically improve the performance of our estimate by considering the semi-parametric efficient doubly-robust estimate of the value. If we construct a regression estimate $\hat{g}(x, w)$ of the conditional expectation $g(x, w) := \mathrm{E}[Y \mid X = x, W = w]$, then an un-biased, lower variance estimate of the policy value is

$$V_{DR}(a) := \frac{1}{n} \sum_{i=1}^{n} a(X_i)(g(X_i, 1) - g(X_i, 0)) + \frac{1}{n} \sum_{i=1}^{n} a(X_i)\, (Y_i - g(X_i, W_i)) \left(\frac{W_i}{p} - \frac{1 - W_i}{1 - p} \right). \tag{2}$$

We first start by using the observation that equivalently, one can simplify the doubly-robust estimate in the form of an appropriately centered IPS estimate (see e.g. Wu and Gagnon-Bartsch (2018)), i.e.

$$V_c(a) := \frac{1}{n} \sum_{i=1}^{n} a(X_i)\, (Y_i - c(X_i)) \left(\frac{W_i}{p} - \frac{1 - W_i}{1 - p} \right), \tag{3}$$

$$c_0(x) := (1 - p)\, \mathrm{E}[Y \mid X = x, W = 1] + p\, \mathrm{E}[Y \mid X = x, W = 0]. \tag{4}$$

After straightforward numerical manipulations it is easy to show that $V_{c_0}(a) = V_{DR}(a)$. Moreover, observe that the optimal centering function can be written as the minimizer of a weighted least squares regression problem (e.g. Spiess (2018)),[1]

$$\mathrm{E}\left[\left(\frac{W}{p^2} + \frac{1 - W}{(1 - p)^2} \right) (Y - c(X))^2 \right] \tag{5}$$

which can subsequently be estimated by arbitrary machine learning non-parametric methods in a data-adaptive manner.

Thus we can estimate the value of any policy using the IPS formula on a centered outcome $Y_i^c = Y_i - c(X_i)$. Subsequently, we can define the t-statistic for the value of a policy as

$$Z_n(a) := \frac{\texttt{AVG}(\rho_1, \ldots, \rho_n)}{\texttt{STD}(\rho_1, \ldots, \rho_n)} \qquad \rho_i := a(X_i)\, \tilde{Y}_i \qquad \tilde{Y}_i := Y_i^c \left(\frac{W_i}{p} - \frac{1 - W_i}{1 - p} \right), \tag{6}$$

1. The first-order condition for each x is: $\frac{1}{p}\, \mathrm{E}[Y \mid X = x, W = 1] + \frac{1}{1-p}\, \mathrm{E}[Y \mid X = x, W = 0] = \left(\frac{1}{p} + \frac{1}{1-p} \right) c(x)$. Re-arranging yields $c(x) = (1 - p)\, \mathrm{E}[Y \mid X = x, W = 1] + p\, \mathrm{E}[Y \mid X = x, W = 0]$ as desired. Wu and Gagnon-Bartsch (2018) proposes a leave-one-out estimator of average treatment effects in experiments based on residualization by this score. Spiess (2018) argues that this specific score solves the weighted prediction problem in (5). Opper (2021) also considers an implementation for average effects in experiments using cross-fitted centering.

where $\texttt{AVG}(\cdot)$ is simply the average of the n numbers and $\texttt{STD}(\cdot)$ is their standard deviation.

First we re-write our objective as

$$Z_n(a)^2 = \frac{\left(\frac{1}{n}\sum_i \rho_i\right)^2}{\frac{1}{n}\sum_i \rho_i^2 - \left(\frac{1}{n}\sum_i \rho_i\right)^2} = \frac{1}{\frac{\frac{1}{n}\sum_i \rho_i^2}{\left(\frac{1}{n}\sum_i \rho_i\right)^2} - 1}. \tag{7}$$

Thus we see that maximizing $Z_n(a)$ is equivalent to minimizing $\frac{\frac{1}{n}\sum_i \rho_i^2}{\left(\frac{1}{n}\sum_i \rho_i\right)^2}$ or maximizing $V_n(a) = \frac{\frac{1}{n}\sum_i \rho_i}{\sqrt{\frac{1}{n}\sum_i \rho_i^2}}$. Thus, we can re-write the following equivalent objective:

$$V_n(a) = \frac{\frac{1}{n}\sum_{i=1}^n a(X_i)\,\tilde{Y}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n a(X_i)\,\tilde{Y}_i^2}}$$

Moreover, throughout our analysis in this section it will be helpful to define the population analogue $V(a)$ of the latter objective,

$$V(a) = \frac{\mathrm{E}[a(X)\tau(X)]}{\sqrt{\mathrm{E}[a(X)\nu(X;c_0)^2]}},$$

where $\nu(X;c) = \mathrm{E}[\tilde{Y}_i^2 \mid X_i = x]$. We also note the following useful properties that relate the expected value of quantities in the empirical objective with quantities in the population objective.

Lemma 2 *The following properties hold:*

$$\begin{aligned}
\tau(x) :=\ & \mathrm{E}[\tilde{Y}_i \mid X_i = x] = \mathrm{E}[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x] \\
\nu^2(x;c) :=\ & \mathrm{E}[\tilde{Y}_i^2 \mid X_i = x] \\
=\ & \frac{\mathrm{Var}(Y^{(1)} \mid X = x)}{p} + \frac{\mathrm{Var}(Y^{(0)} \mid X = x)}{1-p} + \mathrm{E}[Y^{(1)} - Y^{(0)} \mid X = x]^2 + \frac{(c(x) - c_0(x))^2}{p(1-p)} \\
=:\ & \sigma(x)^2 + \tau(x)^2 + \frac{(c(x) - c_0(x))^2}{p(1-p)} = \nu^2(x;c_0) + \frac{(c(x) - c_0(x))^2}{p(1-p)}
\end{aligned}$$

Moreover, maximizing $V(a)$ is equivalent to maximizing $\frac{\mathrm{E}[a(X)\tau(X)]}{\sqrt{\mathrm{E}[a(X)\sigma(X)^2] + \mathrm{Var}(a(X)\tau(X))}}$.

4.2. A solution for infinite or exponentially-sized $\mathcal{X}$

In Section D in the appendix, we provide a solution to the optimization problem for finitely supported features $\mathcal{X}$ based on submodular optimization. However, we note that in typical applications $\mathcal{X}$ will not be finite but either uncountable or exponential in some implicit representation. For instance if x is represented by p binary features, then $|\mathcal{X}| = 2^p$. Moreover, typically, x will be represented by p continuous features. In this setting we not only need to solve the optimal assignment problem for a set of finite pre-defined sub-groups, but in some sense we need to also construct what these sub-groups are. In other words we need to partition the $\mathcal{X}$ space in a finite set of partitions and then find the optimal assignment with respect to this partition. However, the choice of partitions also needs to happen in a manner that will lead to large t-statistic.

A tree-policy based direct approach. A popular approach in data science for such partitioning is recursive partitioning via binary decision trees Athey and Imbens (2016). Given the success of greedy tree-based heuristics in the statistical estimation literature, we focus on such heuristic solutions to this maximization problem. In particular, we define a greedy tree construction algorithm that at every step investigates the expansion of each leaf into two child nodes by splitting on a particular feature and at a particular threshold. Subsequently, we investigate the combination of the two possible treatment configurations of the two candidate child nodes and evaluate the t-statistic of each such new candidate policy, choosing the candidate split that maximizes the statistic. We note that unlike typical tree criteria, our t-statistic criterion cannot be solely computed based on the samples of a node, since the standard deviation in the denominator is a global policy property. Thus at every candidate split we need to use the whole assignment policy to evaluate the performance of a new split. The details of our heuristic are presented in Algorithm 1.

Input: n samples $\{X_i, Y_i^c, W_i\}_{i=1}^n$ of features $X_i \in \mathbb{R}^p$, centered outcome $Y_i^c \in \mathbb{R}$ and treatment $W_i \in \{0,1\}$ **Input:** minimum score increase $\epsilon_{\min}$, maximum depth $d_{\max}$, minimum number of leaf samples $n_{\min}$ Calculate fraction of treated population $p = \frac{1}{n}\sum_i W_i$ Let $\rho_i = Y_i^c\left(\frac{W_i}{p} - \frac{1-W_i}{1-p}\right)$ and calculate t-stat $Z_0 = \frac{\texttt{AVG}(\rho_1,\ldots,\rho_n)}{\texttt{STD}(\rho_1,\ldots,\rho_n)}$ For each sample $i \in [n]$: initialize assignment $a_i = 1\{Z_0 \geq 0\}$ and set $Z_* = Z_0 1\{Z_0 \geq 0\}$ Initialize root $(id = 0, N = N_0, d = 0, w = 1\{Z_0 \geq 0\})$, containing samples $N = \{1,\ldots,n\}$, at depth $d = 0$ Add root node to a queue, set $n_{nodes} = 1$
while *queue is not empty* **do**
  Pop a node (id, N, d, w) from the head of the queue If $d \geq d_{\max}$: **continue** Set $split = \emptyset$ **for** $j \in \{1,\ldots,p\}$ **do**
    Draw random set of thresholds $\{\theta_1,\ldots,\theta_{\text{mtry}}\}$ uniformly without replacement from $\{X_{i,j}\}_{i\in N}$ **for** $\theta \in \{\theta_1,\ldots,\theta_{mtry}\}$ **do**
      Let $N_\ell = \{i \in N : X_{i,j} < \theta\}$ and $N_r = \{i \in N : X_{i,j} \geq \theta\}$ If $|N_\ell| \leq n_{\min}$ or $|N_r| \leq n_{\min}$: **continue** **for** $c \in \{\ell, r\}$ **do**
        Initialize candidate assignment policy $\tilde{a} = a$ For each sample $i \in N$: update candidate policy $\tilde{a}_i = 1\{i \in N_c\}$ Let $\rho_i = \tilde{a}_i Y_i^c\left(\frac{W_i}{p} - \frac{1-W_i}{1-p}\right)$ and calculate z-stat of new policy: $Z(\tilde{a}) = \frac{\texttt{AVG}(\rho_1,\ldots,\rho_n)}{\texttt{STD}(\rho_1,\ldots,\rho_n)}$ **if** $Z(\tilde{a}) \geq Z_* + \epsilon_{\min}$ **then**
          Set $Z_* := Z(\tilde{a})$ and $a := \tilde{a}$ Set $split := \{(n_{nodes}, N_\ell, d+1, 1\{\ell = c\}), (n_{nodes}+1, N_r, d+1, 1\{r = c\})\}$ Set $feature(id) := j$ and $threshold(id) := \theta$ Set $left(id) = n_{nodes}$ and $right(id) = n_{nodes} + 1$
        end
      end
    end
  end
  If $split = \{N_\ell, N_r\}$, append N_ℓ and N_r at the end of the queue and set $n_{nodes} := n_{nodes} + 2$
end
Output: An assignment policy $\alpha : \mathcal{X} \to \{0,1\}$ implicitly defined by the constructed binary tree.
Algorithm 1: Evidence-Based Policy Tree Learning

A model-based approach. Observe that to identify the optimal policy it suffices to consider policies that offer a uniform treatment assignment for groups of x's for which neither the mean treatment effect $\tau(x)$ or $\nu(x)$ changes.

Lemma 3 (Exact Partitions) *There exists an optimal policy such that all elements in $\mathcal{X}$ with both $\tau(x) = \tau$ and $\nu(x) = \nu$, for some τ and $\nu \geq 0$, receive the same treatment assignment.*

Lemma 4 (Approximate Partitions) *Let $\{P_1, \ldots, P_k\}$ be a partition of the space $\mathcal{X}$, with $p_i = \Pr[X \in P_i]$, and such that, for all $i \in [k]$ (where $[k] := \{1, \ldots, k\}$) and for any $x, x' \in P_i$,*

$$\sqrt{p_i}\,|\tau(x) - \tau(x')| \leq \epsilon, \qquad \sqrt{p_i}\,|\nu^2(x) - \nu^2(x')| \leq \epsilon, \tag{8}$$

and let $\tau_i = \mathrm{E}[\tau(X) \mid X \in P_i]$ and $\nu_i^2 = \mathrm{E}[\nu^2(X) \mid X \in P_i]$. Suppose that a_* is an optimal policy and that $\min_{i\in[k]} \nu_i^2 \geq \underline{\nu}^2$ and $\max_{i\in[k]} \tau_i/\nu_i^2 \leq \bar{\kappa}$. Then there exists an assignment policy a_ϵ, which for every $i \in [k]$, assigns the same treatment to all $x \in P_i$, and which satisfies: $Z(a_\epsilon) \geq Z(a_*) - \frac{2\,k}{\underline{\nu}}\,(1 + \bar{\kappa})\,\epsilon$.

Thus if we can find a partition of the $\mathcal{X}$ space into a finite set of partitions $P_1, \ldots, P_k$, such that for all $i \in [k]$, and for all $x, x' \in P_i$: $\tau(x) = \tau(x') = \tau_i$ and $\nu(x) = \nu(x') = \nu_i$, then observe that it suffices to optimize over policies $a(x) = r(\pi(x))$, where $\pi(x)$ denotes the index of the partition that x lies in. Equivalently, it suffices to identify the subset $S \subseteq [k]$ of partitions to treat, with value $v(S) := \frac{\sum_{i\in S} p_i \tau_i}{\sum_{i\in S} p_i \nu_i^2}$. Subsequently, the reduction to submodular minimization for finite elements $\mathcal{X}$ applies, but now the partitions $\{P_1, \ldots, P_k\}$, take the role of the finitely many elements of $\mathcal{X}$. Hence, a $(1+\epsilon)$-approximate optimal policy can be found with $O\left(\log(1/\epsilon)\right)$ oracle calls to a submodular minimization oracle for decomposable submodular functions, as in Theorem 11.

Now we focus on the finite sample regime and focus on the case when $\mathcal{X} = \{0,1\}^p$. Moreover, we assume that both $\tau(x)$ and $\nu^2(x)$ are r-sparse functions, in the sense that they can be written as: $\tau(x) = \tilde{\tau}(x_R)$ and $\nu^2(x) = \tilde{\nu}^2(x_R)$ for some subset $S \subseteq [p]$ (where $x_R = (x_i)_{i\in R}$)), with $|R| = r$. Furthermore, for simplicity of exposition we assume that $\tau(x), \nu^2(x) \in [-1/2, 1/2]$ and that $\tilde{Y}_i \in [-1, 1]$. Our results can trivially be extended to any finite bound H on the random variables at an extra multiplicative cost of H in our bounds.

Importantly, we assume that the variables in R are β-strictly relevant, in the sense that the variance reduction incurred by conditioning on one of the variables in R by an additive factor β, larger than conditioning on any variable in $[p]/R$. More formally:

Definition 5 *For any function $g : \{0,1\} \to [-1/2, 1/2]$, let $V(S; g) = \mathrm{E}[\mathrm{E}[g(X) \mid X_S]^2]$. We say that function g satisfies the (β, r)-strong sparsity condition if there exists a set R of size $|R| = r$, such that, for all $i \in R$ and $j \in [p]/R$ and $T \in [p]/\{i\}$:*

$$V(\{i\} \cup T; g) - V(T; g) \geq V(\{j\} \cup T; g) - V(T; g) + \beta \tag{9}$$

Observing that $\mathrm{E}[\tilde{Y}_i \mid X_i = x] = \tau(x)$ and that $\mathrm{E}[\tilde{Y}_i^2 \mid X_i = x] = \nu^2(x)$, we have the following corollary of the results in Syrgkanis and Zampetakis (2020) (in particular Theorem C.10.2).

Lemma 6 *Suppose that both $\tau(x)$ and $\nu^2(x; c_0)$ satisfy the (β, r)-strong sparsity condition. Suppose that for any $x_R \in \{0,1\}^r$, $\mathrm{P}(X_R = x_R) \geq \zeta/2^r$ for some $\zeta > 0$ and $n = \Omega\left(\frac{2^r \log(p/\delta)}{\beta^2} + \frac{2^r \log(1/\delta)}{\zeta}\right)$. Moreover, suppose that the centering function $\hat{c}$ is estimated based on a separate sample and n is large enough such that it satisfies $\|\hat{c} - c_0\|_2^2 \leq \frac{\beta\, p\,(1-p)}{2}$. Then a greedy level-split regression tree trained on labels $\tilde{Y}_i = (Y_i - \hat{c}(X_i))\left(\frac{W_i}{p} - \frac{1-W_i}{1-p}\right)$ and based on the Breiman criterion and with depth $d \geq r$, satisfies w.p. $1-\delta$, that every x, x' that fall in the same leaf of the tree, have*

$\tau(x) = \tau(x')$. Similarly, a greedy level-split regression tree trained on labels $\tilde{Y}_i^2$ and based on the Breiman criterion and with depth $d \geq r$, satisfies w.p. $1-\delta$, that every x, x' that fall in the same leaf of the tree, have $\nu^2(x; c_0) = \nu^2(x'; c_0)$.

Thus by taking the intersection of the partitions defined by two greedy decision trees with depth $d = r$, we guarantee the property required by Theorem 3. This partition can also be thought as the leafs of a decision tree with depth $d = 2\,r$.

Subsequently for any policy value defined on the leafs of a binary tree of depth d, adaptively chosen based on the data, we want to understand the concentration of the numerator and denominator of our objective V_n. If we let A denote the space of binary policies defined on the leafs of an adaptively built binary decision tree of depth d, then we need to control the uniform deviations: w.p. $1-\delta$,

$$\sup_{a\in A}\left|\frac{1}{n}\sum_{i=1}^{n} a(X_i)\,\tilde{Y}_i - \mathrm{E}[a(X)\tau(X)]\right| \leq \epsilon(n,\delta), \tag{10}$$

$$\sup_{a\in A}\left|\frac{1}{n}\sum_{i=1}^{n} a(X_i)\,\tilde{Y}_i^2 - \mathrm{E}[a(X)\nu^2(X_i;c)]\right| \leq \epsilon(n,\delta). \tag{11}$$

Assuming that $\bar{Y}_i$ is bounded in $[-1,1]$, then by standard results in statistical learning theory, both of the above deviations are satisfied with $\epsilon(n,\delta) = O\left(\sqrt{\frac{VC(\Pi)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$ Shalev-Shwartz and Ben-David (2014), where $VC(\Pi)$ is the VC-dimension of policy space Π. By existing results in statistical learning theory, the VC dimension of a binary classification tree of depth d, with p features is $O\left(d\,2^d\log(p)\right)$ (see Mansour (1997) for the first claim of this fact and e.g. Section A.2 of Syrgkanis and Zampetakis (2020) for a proof, or Lemma 4 of Zhou et al. (2018), or Leboeuf et al. (2020) for a more extensive treatment). Thus we have the following lemma:

Lemma 7 *Let A denote the space of binary policies defined on the leafs of an adaptively built binary decision tree of depth d. Let c be a fixed centering function. Let $v := 4d2^d + 2^{d+1}\log(2p)$. Assume that $|\tilde{Y}| \leq H$ a.s.. Then w.p. $1-\delta$, for all $a \in A$:*

$$\left|\frac{1}{n}\sum_{i=1}^{n} a(X_i)\,\tilde{Y}_i - \mathrm{E}[a(X)\tau(X)]\right| \leq \sqrt{\frac{\mathrm{E}_n[a(X)\nu(X;c)^2]\,2\,v\log(2\,e\,n/\delta)}{n}} + H\frac{7\,v\log(2\,e\,n/\delta)}{3(n-1)} \tag{12}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} a(X_i)\,\tilde{Y}_i^2 - \mathrm{E}[a(X)\nu(X;c)^2]\right| \leq H\sqrt{\frac{\mathrm{E}[a(X)\tilde{Y}^2]\,2\,v\,\log(e\,n/\delta)}{n}} + H\frac{v\,\log(2\,e/\delta)}{3n} \tag{13}$$

Combining all the latter facts we have the following final theorem:

Theorem 8 *Suppose that both $\tau(x)$ and $\nu^2(x)$ satisfy the (β, r)-strong sparsity condition. Suppose that for any $x_R \in \{0,1\}^r$, $\mathrm{P}(X_R = x_R) \geq \zeta/2^r$ for some $\zeta > 0$ and $n = \Omega\left(\frac{2^r\log(p/\delta)}{\beta^2} + \frac{2^r\log(1/\delta)}{\zeta}\right)$. Let $\hat{a}$ be the output of Algorithm 2 with inputs, n samples $\{X_i, Y_i^{\hat{c}}, W_i\}_{i=1}^n$ with a centering model $\hat{c}$ trained on a separate sample, sparsity bound r and target approximation level ϵ. Moreover, suppose*

that the centering function $\hat{c}$ satisfies $\|\hat{c} - c_0\|_2^2 \leq \frac{\beta\, p\,(1-p)}{2}$. Finally suppose that $\nu^2(x; c_0) \geq \underline{\nu}^2$ for all $x \in \mathcal{X}$ nad that $|\tilde{Y}^{\hat{c}}| \leq H$ a.s.. Let $v = 4r2^r + 2^{r+1}\log(2p)$) and:

$$\kappa(n, \eta, \delta) := \sqrt{\frac{2\, v\, \log(2\, e\, n/\delta)}{n}} + \frac{7\, H\, v\, \log(2e\, n/\delta)}{3(n-1)} + \frac{2^{3r/2}}{2(1+\eta)^{3/2}\zeta\, \underline{\nu}^3}\left(\frac{\|\hat{c} - c_0\|_2^2}{p(1-p)} + \frac{2H\, v \log(2e\, n/\delta)}{\eta\, n}\right)$$

Let a_* be an optimal policy for the criterion $V(a)$. Then $\hat{a}$ satisfies w.p. $1-\delta$:

$$\frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}[\hat{a}(X)\nu^2(X; c_0)]}} \geq \sqrt{\frac{1-\eta}{1+\eta}}\frac{1}{1+\epsilon}\frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}[a_*(X)\nu^2(X; c_0)]}} - 2\kappa(n, \eta, \delta) \tag{14}$$

Moreover, the algorithm runs in time polynomial in $\{2^r, \log(1/\epsilon), p, n\}$.

Finally, in the latter theorem we note that we can remove the $\log(n)$ factor in the leading term of the error, if the regression trees for learning $\tau(x)$ and $\nu^2(x)$ are also trained on a separate sample. This will lead to a leading error term of $\sqrt{2^{r+2}\log(2/\delta)/n}$, which could lead to an improvement for large n.

Input: n samples $\{X_i, Y_i^c, W_i\}_{i=1}^n$ of features $X_i \in \mathbb{R}^p$, centered outcome $Y_i^c \in \mathbb{R}$ and treatment $W_i \in \{0, 1\}$ **Input:** Sparsity bound r **Input:** Approximation target $\epsilon > 0$ Calculate fraction of treated population $p = \frac{1}{n}\sum_i W_i$ and let $\tilde{Y}_i := Y_i^c\left(\frac{W_i}{p} - \frac{1-W_i}{1-p}\right)$ Estimate a conditional average treatment effect $\hat{\tau}(x)$ by running a greedy level-splits regression tree construction algorithm with Breiman's criterion (Algorithm 1 of Syrgkanis and Zampetakis (2020)) with labels $\tilde{Y}_i$ and features X_i and depth $d = r$ Estimate a conditional second moment $\hat{\nu}^2(x)$ by running a greedy level-splits regression tree construction algorithm with Breiman's criterion (Algorithm 1 of Syrgkanis and Zampetakis (2020)) with labels $\tilde{Y}_i^2$ and features X_i and depth $d = r$ Let $\{P_1, \ldots, P_k\}$, with $k = 2^{2r}$, be the partition of $\mathcal{X}$ defined by the intersection of the leafs of the two trees, i.e. $x, x' \in P_i$ if and only if x, x' fall in the same leaf on both trees Apply Algorithm 3 with $Q := \{P_1, \ldots, P_k\}$, target approximation level ϵ and value functions $v(P_t) = \frac{1}{n}\sum_{i\in[n]:X_i\in P_t}\tilde{Y}_i^2$ and $w(P_t) = \frac{1}{n}\sum_{i\in[n]:X_i\in P_t}\tilde{Y}_i$ Let S_* be the output of the algorithm **Output:** Assignment policy $\hat{a} : \mathcal{X} \to \{0, 1\}$, defined as: $\hat{a}(x) = 1\{x \in \cup_{t\in S_*} P_t\}$

Algorithm 2: Evidence-Based Policy Learning via Sparse Non-Parametric Model Estimation

5. Conclusion

Treatment policies are often only implemented when the resulting assignment passes a hypothesis test. In this article, we formalize this requirement, and provide a characterization of optimal policies that are optimized for yielding significant evidence for treatment effects on sub-samples. We provide a greedy implementation using decision trees and document the performance of our algorithm in a simulation study.

Our analysis remains limited in a number of ways we currently address in ongoing research. First, for simplicity we focus here on testing on a hold-out sample, but we consider extensions that use cross-validation to use data more efficiently. Second, we focus on a randomized trial, but these ideas apply similarly to observational data under selection on observables. Finally, we consider here a test against the null of no improvement over assigning everybody to control, but the methods extend e.g. to testing against other reference assignments and to testing for heterogeneous treatment effects.

References

Timothy B Armstrong and Shu Shen. Inference on optimal treatment assignments. *SSRN Electronic Journal*, 2015.

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

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Wenruo Bai, Rishabh Iyer, Kai Wei, and Jeff Bilmes. Algorithms for optimizing the ratio of submodular functions. In *International Conference on Machine Learning*, pages 2751–2759. PMLR, 2016.

Dimitris Bertsimas, Nikita Korolko, and Alexander M Weinstein. Identifying Exceptional Responders in Randomized Trials: An Optimization Approach. *INFORMS Journal on Optimization*, 1 (3):187–199, July 2019.

Victor Chernozhukov, Matt Goldman, Vira Semenova, and Matt Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018a.

Victor Chernozhukov, Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*, 2018b.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018c.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *arXiv preprint arXiv:1809.05224*, 8, 2018d.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric Tests for Treatment Effect Heterogeneity. *The review of economics and statistics*, 90(3):389–405, 2008.

Iván Díaz and Mark J van der Laan. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference*, 1(2):171–192, 2013.

Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Interactive identification of individuals with positive treatment effect while controlling false discoveries. *arXiv preprint arXiv:2102.10778*, 2021.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

Esther Duflo, Pascaline Dupas, Thomas Ginn, Grace Makana Barasa, Moses Baraza, Victor Pouliquen, and Vandana Sharma. Hiv prevention among youth: A randomized controlled trial of voluntary counseling and testing for hiv and male condom distribution in rural kenya. *PloS one*, 14(7):e0219535, 2019.

Raaz Dwivedi, Yan Shuo Tan, Briton Park, Mian Wei, Kevin Horgan, David Madigan, and Bin Yu. Stable discovery of interpretable subgroups via calibration in causal studies. *International Statistical Review*, 88(S1):S135–S178, 2020. doi: https://doi.org/10.1111/insr.12427. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12427.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *arXiv preprint arXiv:1807.11408*, 2018.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251, 2018.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

Edward H Kennedy, Scott Lorch, and Dylan S Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143, 2019.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Rahul Ladhania, Amelia Haviland, Neeraj Sood, Edward Kennedy, and Ateev Mehrotra. Learning and Testing Sub-groups with Heterogeneous Treatment Effects:A Sequence of Two Studies. June 2020.

Jean-Samuel Leboeuf, Frédéric LeBlanc, and Mario Marchand. Decision trees as partitioning machines to characterize their generalization properties. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18135–18145. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d2a10b0bd670e442b1d3caa3fbf9e695-Paper.pdf.

Lihua Lei, Aaditya Ramdas, and William Fithian. Star: A general interactive framework for fdr control under structural constraints. *arXiv preprint arXiv:1710.02776*, 2017.

Liu Leqi and Edward H Kennedy. Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*, 2021.

Yishay Mansour. Pessimistic decision tree pruning based on tree size. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 195–201. Morgan Kaufmann, 1997.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *The 22nd Conference on Learning Theory (COLT)*, 2009.

Edward McFowland, III, Sriram Somanchi, and Daniel B Neill. Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection. March 2018.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Isaac M Opper. Improving average treatment effect estimates in small-scale randomized controlled trials, 2021.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941, 2019.

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.

Aaditya K Ramdas, Rina F Barber, Martin J Wainwright, Michael I Jordan, et al. A unified treatment of multiple testing with prior knowledge using the p-filter. *Annals of Statistics*, 47(5):2790–2821, 2019.

Daniel Rubin and Mark J van der Laan. A general imputation methodology for nonparametric regression with censored data, 2005.

Daniel Rubin and Mark J van der Laan. A doubly robust censoring unbiased transformation. *The international journal of biostatistics*, 3(1), 2007.

Alexander Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Jann Spiess. Optimal Estimation when Researcher and Social Preferences are Misaligned, 2018.

Peter Stobbe and Andreas Krause. Efficient minimization of decomposable submodular functions. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/6ea2ef7311b482724a9b7b0bc0dd85c6-Paper.pdf.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on Learning Theory*, pages 3453–3454. PMLR, 2020.

Victor B Talisa and Chung-Chou H Chang. Learning and confirming a class of treatment responders in clinical trials. *Statistics in medicine*, June 2021.

Mark J van der Laan and Alexander R Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome, 2014.

Edward Wu and Johann A Gagnon-Bartsch. The LOOP Estimator: Adjusting for Covariates in Randomized Experiments. *Evaluation Review*, 42(4):458–488, 2018.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:arXiv:1810.04778*, 2018.

Appendix A. Theoretical Illustration

In Section 3 we argued in favor of choosing an assignment that maximizes the oracle criterion $\mathrm{E}[a(X)\,\tau(X)]/\sqrt{\mathrm{E}[a(X)\,\sigma^2(X)]}$. In Section 4, we showed how empirical maximization in the training data approximates an optimal assignment, in the sense that it yields a value of the oracle criterion that is close to the maximum. Here, we demonstrate in a theoretical example how choosing an assignment according to this empirical maximization can improve power relative to a simple assignment based on estimating conditional average treatment effect alone.

Specifically, for illustration we consider a data-generating process with k equally-sized covariate cells $x \in \{1, \ldots, k\}$, with all outcomes bounded. In the training data of size n, cell-wise average treatment effects are

$$\tau_x = \mathrm{E}[Y^{(1)} - Y^{(0)} | X = x] = \mu_x / \sqrt{n},$$

where we again assume local-to-zero effects as in Section 3 to obtain non-degenerate solutions for the power of the test. In the training data, we obtain treatment effect estimates $\hat{\tau}_x$ with asymptotic variance σ_x^2 and variance estimates $\hat{\sigma}_x^2$ for every cell x, where

$$\sqrt{n/k}\,\hat{\tau}_x \xrightarrow{d} \mathcal{N}(\mu_x, \sigma_x^2), \qquad\qquad \hat{\sigma}_x^2 \xrightarrow{p} \sigma_x^2.$$

We consider two policy-learning algorithms on the training data:

(A) A simple classification algorithm that targets the average outcome under assignment, choosing an assignment $\hat{a}^{(A)} \in \{0,1\}^k$ that maximizes

$$\frac{1}{k} \sum_{x=1}^{k} \hat{a}_x^{(A)} \hat{\tau}_x$$

by setting $\hat{a}_x^{(A)} = \mathbb{1}\{\hat{\tau}_x > 0\}$;

(B) An evidence-based algorithm that targets the sample analogue of the testing criterion as in Section 4, choosing an assignment $\hat{a}^{(B)} \in \{0,1\}^k$ that maximizes

$$\frac{\frac{1}{k}\sum_{x=1}^{k} \hat{a}_x^{(B)} \hat{\tau}_x}{\sqrt{\frac{1}{k}\sum_{x=1}^{k} \hat{a}_x^{(B)} \frac{\hat{\sigma}_x^2}{n/k}}}.$$

We assume that we evaluate both policies on a hold-out sample of size $N = \eta n$, where $\eta > 0$ is fixed throughout. We then compare the expected utility

$$\mathrm{E}[U(\hat{a})] = \mathrm{E}\left[\mathrm{P}\left(\frac{\hat{u}(\hat{a})}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(\hat{a}))}} \geq z_{1-\alpha} \,\middle|\, \hat{a}\right)\right]$$

between the two approaches, where we average over the hold-out (outer expectation) and over the training data (inner probability). For illustration, we consider an extreme case – that treatment effects are arbitrary, but noise levels vary so much that only one of them is worth testing in the hold-out.

Proposition 9 (Avoiding subsets with high noise)

Assume that treatment effects are arbitrary (and positive), with fixed $\mu_x^2 > 0$. Then for every $\delta > 0$ there exist a constant C such that if

$$\min_x \sigma_x^2 \leq \frac{1}{C}, \qquad \max_x \sigma_x^2 \geq C$$

then

$$\mathrm{E}[U(\hat{a}^{(A)})] < \alpha + \frac{1-\alpha}{2} + \delta, \qquad 1 - \delta < \mathrm{E}[U(\hat{a}^{(B)})]$$

for n sufficiently large.

This result illustrates the more general approximation results in the previous section. Specifically, by Theorem 12, we can approximate the value of an optimal assignment up to an error that is of the same order of magnitude as the difference between different treatment assignments in this local-to-zero framework. Moreover, the leading term in the error in Theorem 12 does not depend on the specific assignment or variances. By choosing sufficiently low variances that lead to a probability of rejection that is arbitrarily close to one for an *optimal* policy, we therefore also obtain an assignment with probability of rejection that is arbitrarily close to one from the training dataset with high probability. The simple classification algorithm, on the other hand, will select the bucket with variance $\max_x \sigma_x^2 \geq C$ at least half of the time in large samples, leading to an assignment with rejection probability that can be arbitrarily close to the size α of the test. Taken together, we find some (extreme) conditions that (for α, δ not too large) imply that the evidence-based assignment dominates.

Appendix B. Simulation Study

We evaluate experimentally how our algorithm compares to alternative methods in the policy learning and heterogeneous treatment effect literature. We consider two experimental designs. In the first setting we consider a stylized data generating process that highlights vividly the regime where our algorithm substantially out-performs the alternative approaches in the literature. Subsequently, we consider a semi-synthetic experiment where we use the features and treatment assignments from a real-world dataset from an HIV trial Duflo et al. (2019) that consists of a four-armed randomized trial among youth in western Kenya and assesses the effect on herpes type 2 (HSV-2) prevalence. We find that in both settings our evidenced based policy tree achieves strictly higher or comparable out-of-sample p-values, as compared to the treatment subgroups identified by the alternative methods, across multiple signal-to-noise regimes, with the difference being very stark in some conditions.

Benchmarks. We consider the following benchmark policies/methods:

1. **all**: Treating everyone

2. **cate**: First, estimating a heterogeneous treatment effect $\tau(x)$ based on the R-learner method (double machine learning for heterogeneous effects) Nie and Wager (2017), which is one of the most popular methods for heterogeneous treatment effect estimation and has several favorable formal statistical properties Foster and Syrgkanis (2019); Chernozhukov et al. (2018c).

In particular, we used random forest regression and classification to estimate the residual outcome and residual treatment required in the R-learner algorithm. Moreover, we used a single decision tree as a final stage treatment effect model, to make it comparable to the model that is constructed by our policy tree. Subsequently, we set the policy to be $a(x) = 1\{\tau(x) \geq 0\}$.

3. **cate_const**: Same as **cate**, but now instead of estimating a propensity model in the R-learner algorithm, we use the constant prior propensity, since we are in a randomized trial. This could potentially reduce variance injected by propensity estimation.

4. **evidence**: Our evidence based policy tree algorithm but without centering, i.e. $c(x) = 0$.

5. **evidenceres**: Our proposed evidence-based policy tree solution from Algorithm 1.

6. **policy_submod**: The **evidenceres** method is used to define the sub-groups as the leafs of the tree. Subsequently, submodular minimization (Algorithm 3) is used to find the truly optimal assignment on the leafs.

7. **submod**: Our proposed model-based solution outlined in Algorithm 2.

8. **classifier**, **classifierres**: A very popular policy learning method, where the doubly robust policy value $V_{DR}(a)$ is optimized over the space of tree policies Dudík et al. (2011); Athey and Wager (2021). Observe that this problem can be framed as a weighted classification problem, where the sign of the doubly robust values for each sample $\mu_i = (Y_i - c(X_i)) \left(\frac{W_i}{p} - \frac{1-W_i}{1-p}\right)$ can be used as labels $\hat{y}_i = sign(\mu_i)$, and the absolute value of these values μ_i as sample weight. We used a decision tree classifier to solve these classification problem. The variant **classifier** does not perform optimal centering and simply uses $c(x) = \frac{1}{n}\sum_i y_i$, i.e. the mean of the y_i. The variant **classifierres** performs optimal centering of the y values.

All of our methods are judged by calculating the out-of-sample p-value of the chosen treatment policy a. In other words, we use the optimally centered policy value $V_c(a)$ and calculate the quantity $Z(a)$ defined in Equation (6) on a held-out sample. The machine learning model that is trained to calculate the optimal centering (we used a random forest regressor), was trained on the training sample and was only evaluated on the held-out sample.

B.1. Simple synthetic setting

We first consider a very simple setting that portrays the regime where our method shines compared to these alternative aforementioned approaches. We consider the case of a single feature $X \in [0, 3]$. The feature space is partitioned in three intervals $[0, 1]$, $[1, 2]$, $[2, 3]$. In each interval t, the value of the outcome of each sample i can be described by the following equation:

$$y_i = w_i \cdot \tau_t + \beta_t + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_t^2) \tag{15}$$

The parameters for the three regions are $\tau_t = (0, 2, 1)$, $\beta_t = (0, 10, 100)$, $\sigma_t = (5, 10, 1)$. Thus we see that in the first region the treatment effect is null, and samples from this region should not really be treated. In the second regime the treatment effect is very positive, but samples from that region also have high variance (i.e. unobserved heterogeneity). Thus even though with many samples we would have clear evidence that samples from this regime should be treated, in a small sample regime,

treating this region runs the risk of getting a statistically insignificant result and hence not being able to declare that the treatment is effective. The final regime has a modest treatment effect, but is much less noisier. Moreover, we observe that these three regions operate in orders of magnitude different baseline outcome regimes. Thus without correct centering of the outcomes, we have a very noisy outcome which can hurt both the accuracy of the method.

We exactly recover these intuitive expectations from our experiments. We find that the evidence based algorithm with optimal centering **evidenceres**, significantly out-performs all other benchmarks for a large range of sample sizes. The results are portrayed in Figure 1.

B.2. Semi-synthetic experiments from HIV randomized trial

The HIV randomized trial data Duflo et al. (2019), is a four-arm trial that was deployed among youths in Kenya. See Duflo et al. (2019) for the details of the data collection and evaluation process. We used one of the four binary arms as our treatment and subselected 2456 samples that did not contain NaN values in the treatment or outcome. Each sample contains 44 observable features/characteristics. The outcome of interest was the prevalence of herpes type 2 (HSV-2) virus, two years after the treatment (binary outcome $y_i \in \{0, 1\}$).

We analyzed the performance of our algorithm in a semi-synthetic manner. We used the raw 2456×44 features X_i, and their corresponding treatment w_i form the real data. Subsequently, we trained a ℓ_1-penalized (sparse) logistic regression outcome model, separate on the treated and untreated population, to learn the functions $p_w(x) = \Pr(y = 1 \mid X = x, W = w) = expit(\theta_w^\top x)$. Subsequently, we also trained un-penalized logistic regressions, to de-bias the learned coefficients. The learned coefficients are depicted in Figure 2. We see that there are three significant factors of heterogeneity, with two factors increasing the effect and one factor decreasing the effect. Subsequently, we generated new outcome data from several manipulations of the learned models. In particular, we generated binary outcome data based on the probabilistic models

$$\Pr(y = 1 \mid X = x, W = 1) := \; expit\left(\text{cov-strength} \cdot \theta_1^\top (X_i - \bar{X}) + \text{separation}\right), \tag{16}$$

$$\Pr(y = 1 \mid X = x, W = 0) := \; expit\left(\text{cov-strength} \cdot \theta_0^\top (X_i - \bar{X}) - \text{separation}\right) \tag{17}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. The constant "cov-strength" controls the signal/heterogeneity contained in the features, while the constant "separation" controls the baseline effect of the treatment.

Our experimental results are depicted in Figures 3, 4 and 5. We find that our **evidenceres** and **submod** methods outperform the benchmarks in the medium signal regimes where the signal in the features is relatively low, while being comparable with the alternative methods in the two extreme regimes (very low/very high signal). Morever, the performance of both algorithms is comparable, with **evidenceres** having a slight advantage. Moreover, we see that running submodular minimization at the end of the evidence based policy tree has little to no-effect, implying that the greedy tree algorithm not only found good sub-groups, but the assignment that it constructed was optimal for these sub-groups.

B.3. Semi-synthetic experiments from HIV randomized trial with continuous outcomes

We also simulated a variant of the semi-synthetic experiment in the previous section but now with continuous outcome data. We use the group structure from our synthetic analysis, but we define

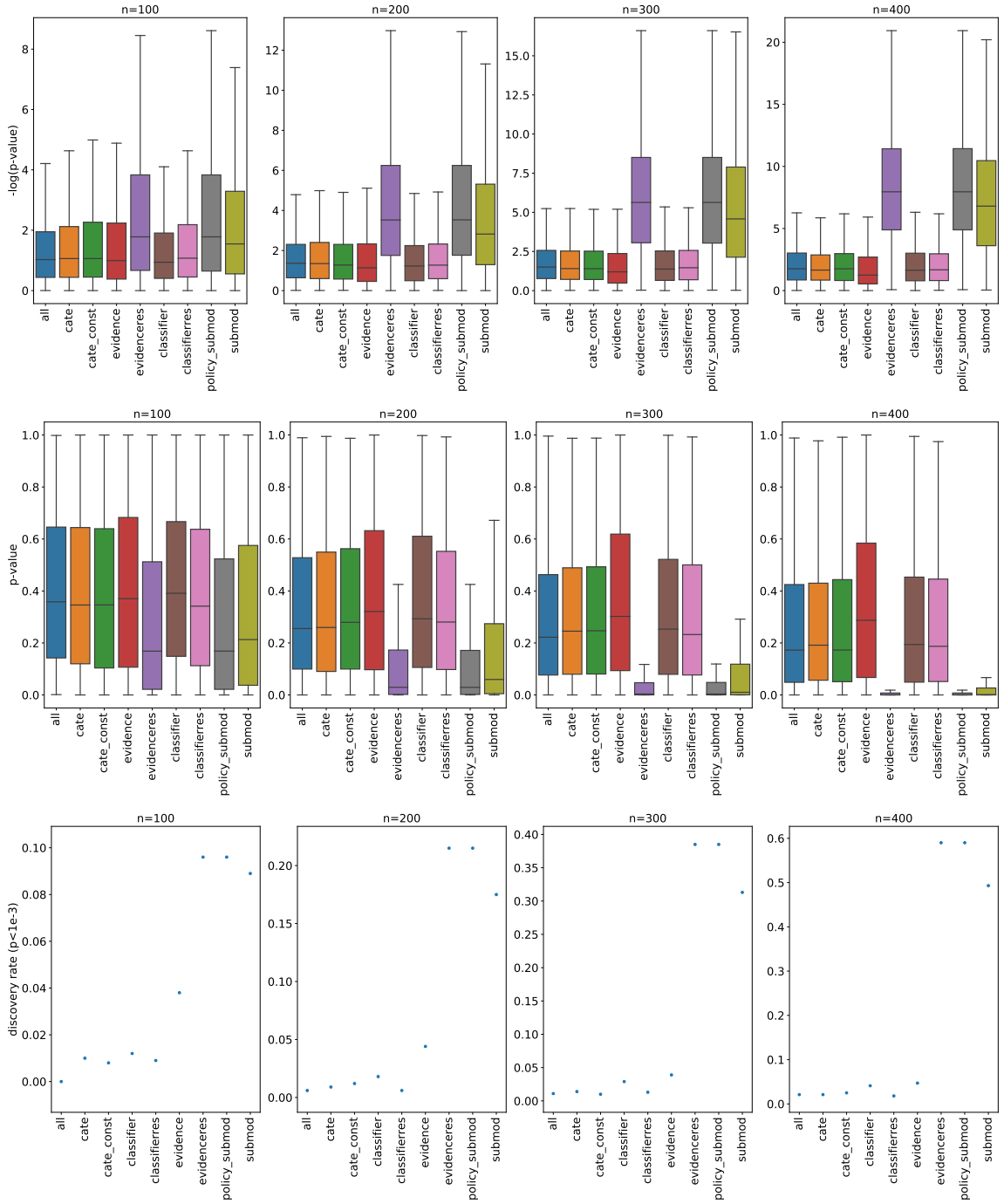

Figure 1: Out-of-sample p-value and negative log of p-value distributions of each method on 1000 experiments. Last figure depicts fraction of 1000 experiments for which out-of-sample p-value $\leq 1e-3$, which we declare as a discovery.

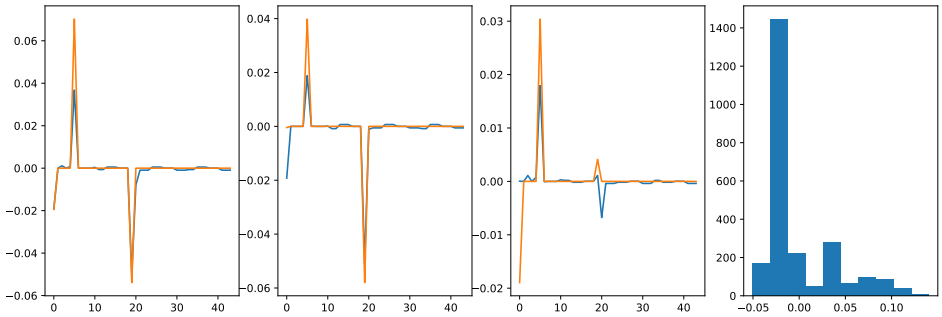

Figure 2: Learned outcome models from the raw data. Left figure is θ_1, second plot is θ_0, third plot is $\theta_1 - \theta_0$ and fourth plot is the histogram of the scores $(\theta_1 - \theta_0)^\top (X_i - \bar{X})$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The orange lines are the de-biased coefficients after the un-penalized logistic regression and the blue ones, the coefficients from the first stage penalized model.

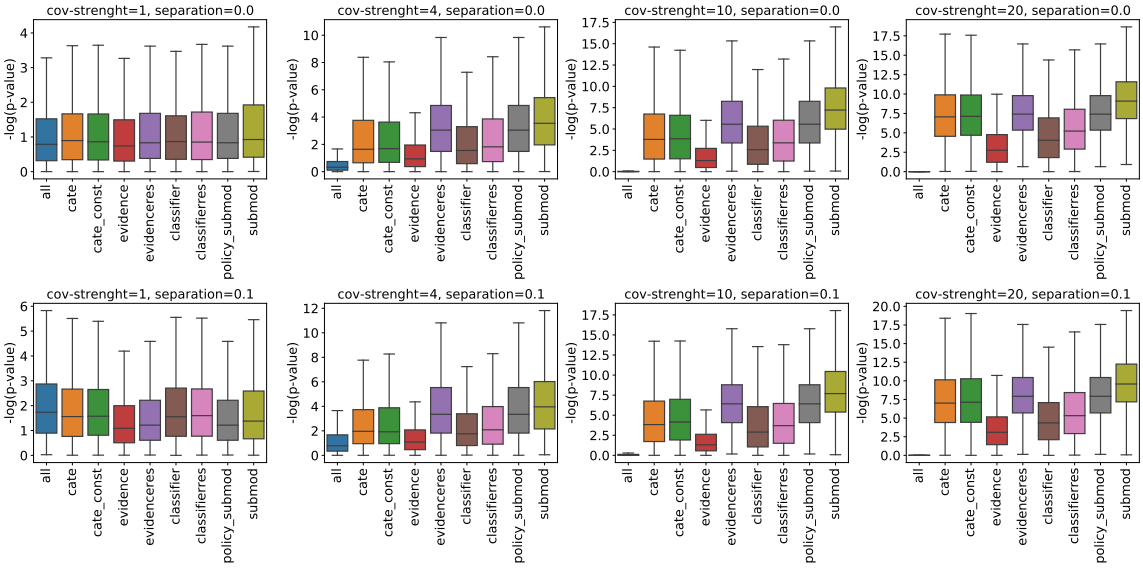

Figure 3: Out-of-sample negative log of p-value distributions of each method on 1000 experiments.

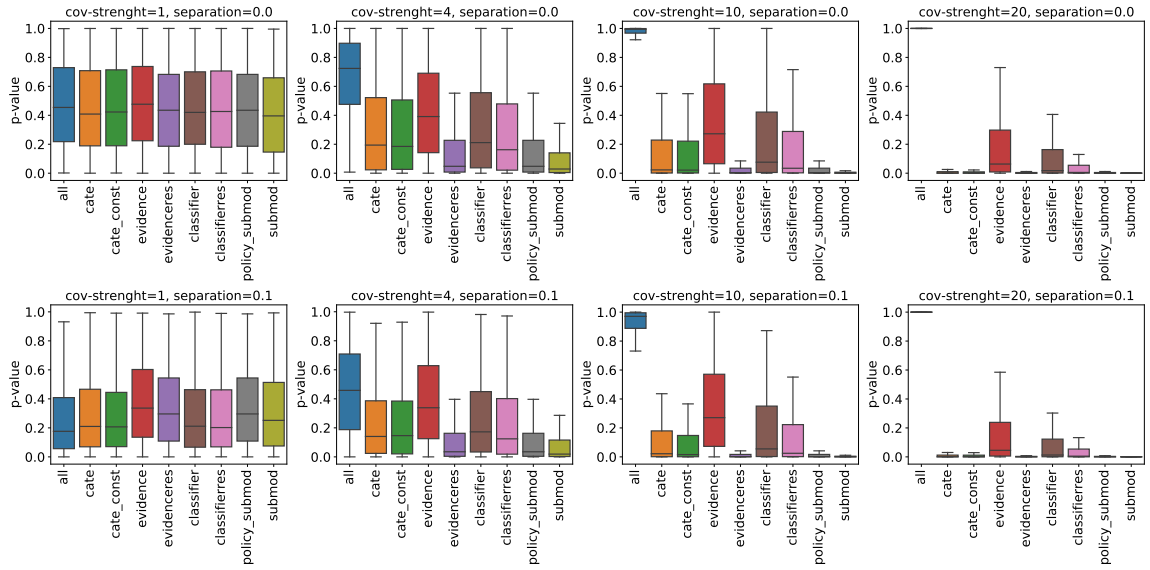

Figure 4: Out-of-sample p-value distributions of each method on 1000 experiments.

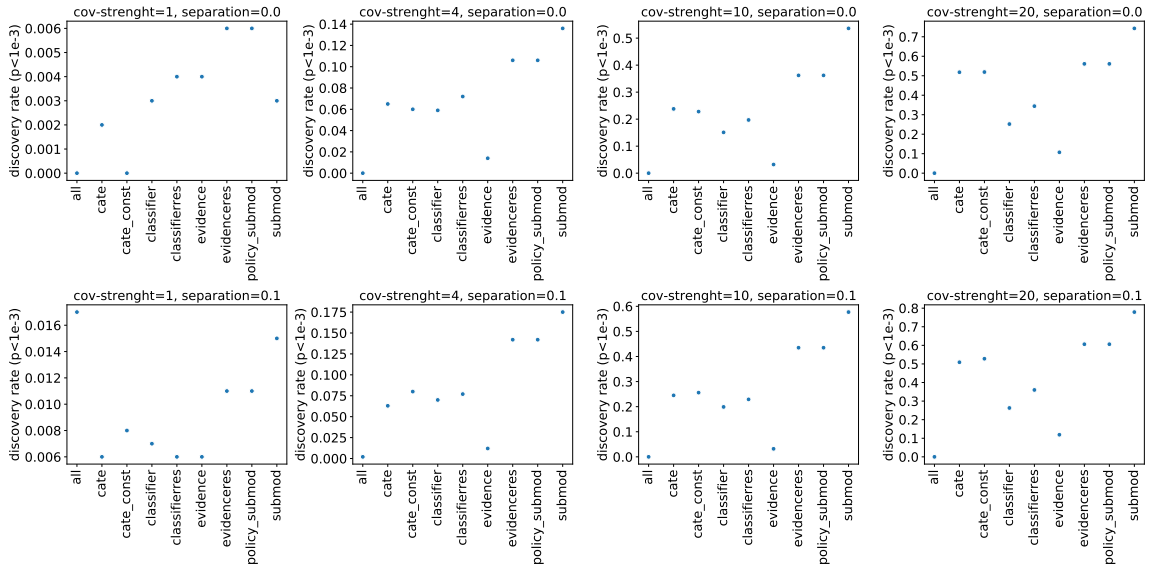

Figure 5: Fraction of 1000 experiments where out-of-sample p-value $\leq 1e-3$, which we declare as a "discovery".

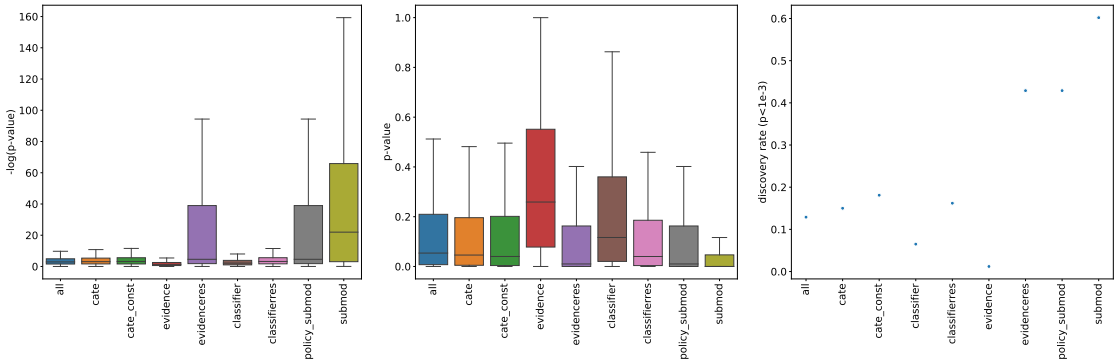

Figure 6: Out-of-sample p-value and negative log of p-value distributions of each method on 1000 experiments. Last figure depicts fraction of 1000 experiments for which out-of-sample p-value $\leq 1e-3$, which we declare as a discovery.

groups based on the real covariates from the HIV trial. Moreover, we use the treatment variable from the real data. The continuous outcome in each experiment is generated as follows:

1. Let $\mu = (-2, 2, 2)$, $\rho = (5, 10, 1)$ and $\beta = (0, 10, 100)$
2. First for each experiment we generate three indices $q = (i, j, k)$ uniformly at random from the set of feature indices $[p]$
3. Let $\theta = (\theta_1, \theta_2, \theta_3)$, where $\theta[t]$ is the median of the $q[t]$ feature across all n samples
4. Then for each sample $i \in [n]$ we generate outcome data based on:

$$Y_i = \tau(X_i)\, W_i + b(X_i) + \epsilon, \quad \epsilon \sim N\left(0, \sigma(X_i)^2\right)$$

$$\tau(X_i) = \sum_{t=1}^{3} \mu[t] \cdot \mathbb{1}\left\{X_i[q[t]] > \theta[t]\right\} \qquad b(X_i) = \sum_{t=1}^{3} \beta[t] \cdot \mathbb{1}\left\{X_i[q[t]] > \theta[t]\right\}$$

$$\sigma(X_i) = \sum_{t=1}^{3} \rho[t] \cdot \mathbb{1}\left\{X_i[q[t]] > \theta[t]\right\}$$

Our experimental results are depicted in Figure 6. We find that our **submod** method outperforms the benchmarks, correctly identifying not to include high-variance sub-groups, even though they might have a slightly positive effect. Moreover, the **evidenceres** method comes a close second, albeit in this setting we see that **submod** is strictly better than the **evidenceres** method.

Appendix C. Extensions

C.1. Continuous relaxation

In the main article, we focussed on obtaining a binary assignment $a : \mathcal{X} \to \{0, 1\}$ that assigns an individual with a given covariate vector x to treatment or not. We can relax the assumption

and consider instead probabilistic assignments $a : \mathcal{X} \to [0, 1]$, where our target is the expected list $u(a) = \mathrm{E}[a(X)\,\tau(X)]$. Testing for $\mathrm{E}[a(X)\,\tau(X)] \geq 0$ can now be understood in two ways: first, as a test of a specific probabilistic assignment; or second, as a test of the null hypothesis that there is no (deterministic) assignment with positive expected effect (even if we do now not construct such a deterministic assignment). The oracle result in Theorem 1 remains applicable in this case. Indeed, since the hold-out data is already given, we do not actually randomize, but instead estimate the weighted estimand $u(\hat{a})$ without adding additional randomness (which would be statistically inefficient). Only in the population would we then randomize for an individual with covariate x based on the probability $\hat{a}(x)$. The continuous relaxation allows for the direct solution of optimal weights in the framework of Section 3:

Proposition 10 (Optimal relaxed policy) *Asymptotically as $N \to 0$, the probabilistic policy maximizing the power of the test on the hold-out sample is given by*

$$a^*(x) \propto \frac{\tau_+(x)}{\sigma^2(x)} \qquad \textit{yielding expected power} \qquad U(a^*) = \Phi\left(\sqrt{N\,\frac{\tau_+^2(X)}{\sigma^2(X)}} - z_{1-\alpha}\right)$$

where $\tau_+(x) = \tau(x)\mathbb{1}_{\tau(x)>0}$. (Note that the absolute scale of the policy does not matter here.)

This particularly tractable formulation implies a criterion that can be used to partition the covariate space recursively using trees with binary splits. Specifically, for $\bar{\tau}_\ell$ the average treatment effect of a leaf and $\bar{\sigma}_\ell^2 = \mathrm{Var}(\hat{\tau}_\ell)$ the variance of the corresponding estimator $\hat{\tau}_\ell$, we aim to find a partition that maximizes

$$\sum_\ell \frac{\bar{\tau}_\ell^2}{\bar{\sigma}_\ell^2},$$

corresponding to weights $\bar{a}_\ell \propto \frac{\bar{\tau}_\ell}{\bar{\sigma}_\ell^2}$ on $\hat{\tau}_\ell$. When we also write $\hat{\sigma}_\ell^2$ for an estimator of $\bar{\sigma}_\ell^2$, this suggests maximizing the sample analogue

$$\sum_\ell \frac{\hat{\tau}_\ell^2}{\hat{\sigma}_\ell^2}$$

over partitions of the data, which can be achieved recursively, and then setting $\hat{a}_\ell \propto \frac{\hat{\tau}_\ell}{\hat{\sigma}_\ell^2}$. For illustration, Figure 7 compares a greedily fitted tree that maximizes the criterion in Theorem 1 over binary assignments with a greedily fitted tree that recursively maximizes this weighted criterion.

C.2. Testing for treatment-effect heterogeneity

In the main article, we focus on finding evidence of subgroups with positive treatment effects. We can use a similar approach to finding evidence of heterogeneous treatment effects. This extension is similar to how Crump et al. (2008); Armstrong and Shen (2015) extend tests for non-constant and positive effects to tests for heterogeneous effects. Specifically, we now consider an assignment

$$a : \mathcal{X} \to \{-1, 0, +1\}$$

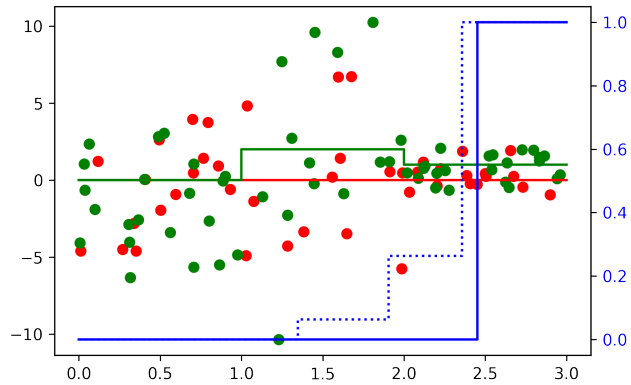

Figure 7: Example of a single tree with $n = 100$ for binary (solid) and weighted (dotted) assignment, where on the blue axis is the probability of assignment to treatment. Green observations are in treatment, and red observations are in control, with solid lines representing conditional means.

where units are assigned to a high-effect ($+1$), neutral (0), and low-effect (-1) group. We then evaluate whether the average treatment effects in the high group are above that of the low group by a hypothesis test. Towards this end, we consider the estimands

$$\Delta_1(a) := \frac{\mathrm{E}[a_+(X)\tau(X)]}{\mathrm{E}[a_+(X)]} - \frac{\mathrm{E}[a_-(X)\tau(X)]}{\mathrm{E}[a_-(X)]} = \mathrm{E}[\tau(X)|a(X) = +1] - \mathrm{E}[\tau(X)|a(X) = -1],$$
$$\Delta_2(a) := \mathrm{E}[a(X)(\tau(X) - \mathrm{E}[\tau(X)])] = \mathrm{Cov}(\tau(X), a(X))$$

where $a_- = (-a)_+ \geq 0$ denotes the negative part. Both $\Delta_1(a) > 0$ and $\Delta_2(a) > 0$ provide evidence of treatment-effect heterogeneity. We focus here on collecting evidence that $\Delta_2(a) > 0$, which itself is maximized for $a^*(x) = \mathrm{sign}(\tau(X) - \mathrm{E}[\tau(X)])$. This test is similar to the regression-based calibration test employed by Chernozhukov et al. (2018a) to evaluate treatment-effect heterogeneity.

A test for $\Delta_2(a) > 0$ is $\hat{T}(a) = \frac{\hat{\Delta}_2(a)}{\sqrt{\widehat{\mathrm{Var}}(\hat{\Delta}_2(a))}}$ where we set

$$\hat{\Delta}_2(a) = \frac{1}{N}\sum_{j=1}^{N}\left(a(X_j) - \frac{1}{N}\sum_{k=1}^{N} a(X_k)\right)\left(\tilde{Y}_j - \frac{1}{N}\sum_{k=1}^{N}\tilde{Y}_k\right)$$
$$\widehat{\mathrm{Var}}(\hat{\Delta}_2(a)) = \frac{1}{N(N-1)}\sum_{j=1}^{N}\left(\left(a(X_j) - \frac{1}{N}\sum_{k=1}^{N} a(X_k)\right)\left(\tilde{Y}_j - \frac{1}{N}\sum_{k=1}^{N}\tilde{Y}_k\right) - \hat{\Delta}_2(a)\right)^2.$$

We now consider a local-to-zero approximation for which $\tau(x) = \bar{\tau} + \frac{\mu(x)}{\sqrt{N}}$. Under regularity conditions,

$$\hat{T}(a) \xrightarrow{d} \mathcal{N}\left(\frac{\mathrm{Cov}(a(X), \mu(X))}{\sqrt{\mathrm{E}[(a(X) - \mathrm{E}[a(X)])^2\sigma^2(X)]}}, 1\right)$$

and we can adapt our above approach to maximize

$$\frac{\text{Cov}(a(X), \tau(X))}{\sqrt{\text{E}[(a(X) - \text{E}[a(X)])^2 \sigma^2(X)]}} = \frac{\text{E}[(a(X) - \text{E}[a(X)])(\tau(X) - \text{E}[\tau(X)])]}{\sqrt{\text{E}[(a(X) - \text{E}[a(X)])^2 \sigma^2(X)]}}$$

either over discrete assignments $a : \mathcal{X} \to \{-1, 0, +1\}$ or the continuous relaxation $a : \mathcal{X} \to \mathbb{R}$.

C.3. Passing the test as a constraint

The testing goal ignores the actual outcome under the policy. We therefore also consider a second goal that incorporates the expected outcome subject to passing the test, namely

$$U'(\hat{a}) = \text{E}[\hat{T}(\hat{a})\, u(\hat{a}) | \hat{a}] = \text{P}\left(\frac{\hat{u}(\hat{a})}{\sqrt{\widehat{\text{Var}}(\hat{u}(\hat{a}))}} \geq z_{1-\alpha} \middle| \hat{a} \right) u(\hat{a}).$$

This goal would correspond for example to a drug manufacturer caring about the health outcomes of a drug subject to approval. It mixes the expected outcome $u(\hat{a})$ with the testing goal $\text{E}[\hat{T}(\hat{a})|\hat{a}]$ to trade off the value of the policy with how easy it is to pass the test. In the approximation of Section 3, the optimal policy would then solve

$$a' = \arg\max_a U'(a) = \arg\max_a \text{E}[a(X)\, \tau(X)]\, \Phi\left(\sqrt{N} \frac{\text{E}[a(X)\, \tau(X)]}{\sqrt{\text{E}[a^2(X)\, \sigma^2(X)]}} - z_{1-\alpha} \right).$$

C.4. Estimation for alternative outcomes and estimated propensity score

In the main article, we considered a treatment effect that is identified in an experiment by $\tau(x) = \text{E}\left[\tilde{Y} \middle| X = x\right] = \text{E}\left[\frac{W-p}{p(1-p)}(Y - c(X)) \middle| X = x\right]$ and can be estimated accordingly. We can extend the same techniques to a family of parameters $\tau(x)$ defined via a linear conditional moment equation:

$$\tau(x) = \text{E}[\psi(Z; h) | X = x] \tag{18}$$

for data Z, target causal parameter $\tau(x)$, and nuisance functions h. For $\tau(x)$ defined in this way, we aim to obtain an assignment that provides a significant test against the null that $u(a) := \text{E}[a(X)\, \tau(X)] \leq 0$. If the moment function satisfies the conditional Neyman orthogonality property Chernozhukov et al. (2018c); Foster and Syrgkanis (2019); Oprescu et al. (2019), and these nuisance estimates are estimated in a cross-fitting manner and are $o(n^{1/4})$ estimable, then their error is asymptotically negligible. In this case, all our algorithms can be directly generalized by simply renaming the pseudo-outcome:

$$\tilde{Y}_i = \psi(Z_i; h) \tag{19}$$

The estimation and optimization algorithms as well as the analysis of their finite-sample performance easily generalizes to this general pseudo-outcome setting. Moreover, Neyman orthogonality for this setting can be easily achieved via an automated-double machine learning process as outlined in the recent work of Chernozhukov et al. (2018d, 2020).

For example, we can generalize the case of conditional average treatment effects to cases where the propensity score is unknown and has to be estimated, in which case $h(w, x) = (c(w, x), p(x))$,

with $c(w,x) = \mathrm{E}[Y \mid W = w, X = x]$, $p(x) = \mathrm{E}[W \mid X = x]$ and $\psi(Y, W, X; (c(\cdot,\cdot), p(\cdot))) = c(1,X) - c(0,X) + (Y - c(W,X)) \left(\frac{W}{p(X)} - \frac{1-W}{1-p(X)}\right)$. Even more straightforwardly, we can also consider the simple prediction setup where we measure the treatment effect (or outcome of interest) $\tilde{Y}$ directly, in which case we write $\tau(x) = \mathrm{E}[\tilde{Y}|X = x]$. Here, $\psi(Y, X) = Y$.

C.5. Cross-evaluation and bagging

So far, we have focussed on separating the training of a policy from its evaluation. While this may be the most applicable case in many heavily regulated settings, we may gain power by re-using training and hold-out samples by switching their roles. Relatedly, our main implementation is based on a simple tree that could be aggregated to a forest, either across the full sample or even within the training sample alone.

C.6. Testing for alternative reference assignments

Our results generalize to reference assignments other than assigning all observations to control. In the case of a reference assignment $b(X)$,

$$u(a|b) = u(a) - u(b) = \mathrm{E}[(a(X) - b(X))\, \tau(X)].$$

In that case, all results go through if we relabel $Y_b^{(w)} = Y^{(|w-b(x)|)}$ for observations with $X = x$.

Appendix D. Optimal Policy via Submodular Minimization

In the main article, we present a tree-based solution to finding assignments. Here, we provide additional optimization results based on submodular optimization.

Observe that for a finite set $\mathcal{X}$, the empirical optimization problem of the objective $V_n(a)$, is equivalent to finding a subset $S \subset \mathcal{X}$ that maximizes $\frac{\sum_{x\in S} w(x)}{\sqrt{\sum_{x\in S} v(x)}}$ for $v(x) := \frac{1}{n}\sum_{i\in[n]:X_i=x} \tilde{Y}_i^2 \geq 0$ and $w(x) := \frac{1}{n}\sum_{i\in[n]:X_i=x} \tilde{Y}_i$. Without loss of generality, when solving the empirical maximization problem, we can discard features $x \in X$ for which $w(x) \leq 0$, thus when solving the optimization problem it suffices consider the case when $w(x) > 0$, for all $x \in \mathcal{X}$. The latter is equivalent to minimizing the ratio of a monotone submodular function over a monotone modular function. Corollary 3.3 of Bai et al. (2016) shows that the latter can be solved up to a $(1+\epsilon)$-approximation with $O(\log(1/\epsilon))$ calls to a submodular minimization process (which can be solved in polynomial time Schrijver (2000)) that minimizes $\sqrt{\sum_{x\in S} v(x)} - \lambda \sum_{x\in S} w(x)$, which is also submodular, for some appropriately chosen λ in each call. We note however that the result in Bai et al. (2016) requires solving a constrained submodular minimization, subject to the constraint of a non-empty solution. This constraint is non-trivial and not the standard submodular minimization problem solved by existing solvers (covering constraints can render submodular minimization NP-hard). We show here how to modify the result and the algorithm in Bai et al. (2016) to only require an un-constrained submodular minimization oracle.

Theorem 11 *Let $\hat{S}$ be the output of Algorithm 3 with inputs $v(x) := \frac{1}{n}\sum_{i\in[n]:X_i=x} \tilde{Y}_i^2 \geq 0$ and $w(x) := \frac{1}{n}\sum_{i\in[n]:X_i=x} \tilde{Y}_i$ and set $Q = \mathcal{X}_+ := \{x \in \mathcal{X} : w(x) > 0\}$. Assume that $Q \neq \emptyset$ and let*

a_* be the policy implied by $\hat{S}$, i.e. $\hat{a}(x) = 1\{x \in \hat{S}\}$. Then $\hat{S}$ and $\hat{a}$ satisfy that

$$\frac{w(\hat{S})}{\sqrt{v(\hat{S})}} \geq \frac{1}{1+\epsilon} \max_{\emptyset \subset S \subseteq Q} \frac{w(S)}{\sqrt{v(S)}} \qquad \Leftrightarrow \qquad V_n(\hat{a}) \geq \frac{1}{1+\epsilon} \max_{a:\mathcal{X}\to\{0,1\}} V_n(a). \tag{20}$$

Moreover, the algorithm will terminate after $O\left(\log(1/\epsilon) + \log(v(Q)/\min_{x\in Q} v(x))\right)$ iterations.

Theorem 11 provides an approximation guarantee on the empirical criterion. However, for finite $\mathcal{X}$ of size k, a simple uniform deviation bound implies that the empirical objective converges to the population objective for any assignment $a : \mathcal{X} \to \{0, 1\}$, i.e. w.p. $1 - \delta$,

$$\sup_a \left| \frac{1}{i} \sum_{i=1}^{n} a(X_i)\, \tilde{Y}_i - \mathrm{E}[a(X)\tau(X)] \right| \leq \epsilon(n, \delta), \tag{21}$$

$$\sup_a \left| \frac{1}{i} \sum_{i=1}^{n} a(X_i)\, \tilde{Y}_i^2 - \mathrm{E}[a(X)\nu^2(X_i; c)] \right| \leq \epsilon(n, \delta) \tag{22}$$

for any fixed centering function and for $\epsilon(n, \delta) := O\left(\sqrt{\frac{k \log(1/\delta)}{n}}\right)$. Moreover, by Theorem 2,

$$\nu^2(x; c) - \nu^2(x; c_0) = \frac{(c(x) - c_0(x))^2}{p(1-p)}.$$

In fact, we can invoke tighter uniform deviation bounds to get the following theorem:

Theorem 12 *Let $\hat{S}$ be the output of Algorithm 3 with inputs $v(x) := \frac{1}{n}\sum_{i\in[n]:X_i=x} \tilde{Y}_i^2 \geq 0$, $w(x) := \frac{1}{n}\sum_{i\in[n]:X_i=x} \tilde{Y}_i$ and target approximation ϵ and set $Q = \mathcal{X}_+ := \{x \in \mathcal{X} : w(x) > 0\}$ and based on a centering function $\hat{c}$ trained on a separate sample. Assume that $Q \neq \emptyset$ and let $\hat{a}$ be the policy implied by $\hat{S}$, i.e. $\hat{a}(x) = 1\{x \in \hat{S}\}$. Moreover, assume that $\nu^2(X; c_0) \geq \underline{\nu}^2$ and $f(x) > \underline{\pi}$, for all $x \in \mathcal{X}$ and that $|\tilde{Y}| \leq H$ a.s.. Let a_* be any policy (e.g. the policy maximizing the population function V) and*

$$\kappa(n, \eta, \delta) := \sqrt{\frac{2\,k\,\log(2/\delta)}{n}} + \frac{7\,H\,k\,\log(2/\delta)}{3(n-1)} + \frac{1}{2(1+\eta)^{3/2}\underline{\pi}^{3/2}\,\underline{\nu}^3}\left(\frac{\|\hat{c} - c_0\|_2^2}{p(1-p)} + \frac{2H\,k\log(2/\delta)}{\eta\, n}\right).$$

Then $\hat{a}$ satisfies w.p. $1 - \delta$:

$$\frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}[\hat{a}(X)\nu^2(X; c_0)]}} \geq \sqrt{\frac{1-\eta}{1+\eta}} \frac{1}{1+\epsilon} \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}[a_*(X)\nu^2(X; c_0)]}} - 2\kappa(n, \eta, \delta) \tag{23}$$

Moreover, the algorithm will terminate after $O\left(\log(1/\epsilon) + \log(v(Q)/\min_{x\in Q} v(x))\right)$ iterations.

Thus for finite $\mathcal{X}$ we can arbitrarily approximate the optimal assignment policy problem with respect to our objective function. Moreover, the latter function is a special case of a submodular function that can be solved much more efficiently than generic submodular minimization Stobbe and Krause (2010).

Input: A finite set of k elements Q **Input:** Two positive value functions $v : Q \to \mathbb{R}_+$ and $w : Q \to \mathbb{R}_+$ **Input:** Approximation target $\epsilon > 0$ For any $S \subseteq Q$, let $v(S) := \sum_{x \in S} v(x)$ and $w(S) := \sum_{x \in S} w(x)$ Set $\lambda_{\max} := (1+\epsilon)\frac{\sqrt{v(Q)}}{w(Q)}$ and $\lambda_{\min} := \frac{\sqrt{\min_{x \in Q} v(x)}}{w(Q)}$ **while** $\lambda_{\max} \geq (1+\epsilon)\lambda_{\min}$ **do**

Set $\bar{\lambda} := (\lambda_{\max} + \lambda_{\min})/2$ and solve unconstrained submodular minimization problem:

$$\hat{S} := \arg\min_{S \subseteq [m]} \sqrt{v(S)} - \bar{\lambda} w(S) \tag{24}$$

If $\hat{S} = \emptyset$ or $\frac{\sqrt{v(\hat{S})}}{w(\hat{S})} \geq \bar{\lambda}$ set $\lambda_{\min} := \bar{\lambda}$ **else** set $\lambda_{\max} := \bar{\lambda}$

end

Output: Return set $\hat{S} := \arg\min_{S \subseteq Q} \sqrt{v(S)} - \lambda_{\max} w(S)$

Algorithm 3: Evidence-Based Policy Learning via Submodular Minimization for Finite $\mathcal{X}$

This reduction also draws strong connections between our approach and the variance penalization literature in policy learning. In particular, Swaminathan and Joachims (2015) proposes that one maximizes an objective of the form (after bringing it to our notation)

$$Z_{VP}(a) := \sum_{x \in \mathcal{X}} a(x) w(x) - \lambda_n \left(\sum_{x \in \mathcal{X}} a(x) v(x) - \left(\sum_{x \in \mathcal{X}} a(x) w(x) \right)^2 \right) \tag{25}$$

with a vanishing λ_n at a roughly $O(n^{-1/2})$ rate when a belongs to a VC class. We see that unlike variance penalization we balance the standard deviation with the mean policy outcome and we do not necessarily have a vanishing penalty λ. Even closer to our objective is the second moment penalization in Foster and Syrgkanis (2019), which recommends maximizing

$$\sum_{x \in \mathcal{X}} a(x) w(x) - \lambda \sum_{x \in \mathcal{X}} a(x) v(x). \tag{26}$$

Again this objective has the same discrepancy with our objective in that we balance the root of the second moment and we do not necessarily have a vanishing λ. Moreover, in our objective the λ is auto-tuned so as to minimize the t-statistic.

Appendix E. Proofs from Section 3

E.1. Proof of Theorem 1

Proof We first consider the case of a known propensity score p, in which case for every a with $\mathrm{E}[a(X)] > 0$ and with $\tilde{Y} = \left(\frac{W}{p} - \frac{1-W}{1-p}\right)(Y - c(X))$ we have under standard regularity assumptions that

$$\frac{\sqrt{N}}{\sqrt{\mathrm{Var}(a(X)\,\tilde{Y})}}(\hat{u}(a) - u(a)) = \frac{1}{\sqrt{N\ \mathrm{Var}(a(X)\,\tilde{Y})}} \sum_{j=1}^{N} (a(X_j)\,\tilde{Y}_j - \mathrm{E}[a(X)\,\tilde{Y}] \xrightarrow{d} \mathcal{N}(0,1),$$

$$\mathrm{Var}(a(X)\,\tilde{Y}) = \mathrm{Var}(\mathrm{E}[a(X)\,\tilde{Y}|X]) + \mathrm{E}[\mathrm{Var}(a(X)\,\tilde{Y}|X)] = \mathrm{Var}(a(X)\,\tau(X)) + \mathrm{E}[a^2(X)\ \mathrm{Var}(\tilde{Y}|X=x)]$$

$$= \frac{1}{N}\mathrm{Var}(a^2(X)\,\mu(X)) + \mathrm{E}[a^2(X)\ \mathrm{Var}(\tilde{Y}|X=x)] \to \mathrm{E}[a^2(X)\ \mathrm{Var}(\tilde{Y}|X=x)] = \mathrm{E}[a^2(X)\,\sigma^2(X)]$$

since $\mathrm{E}[\tilde{Y}|X=x]=\tau(x)$. Also,

$$N\widehat{\mathrm{Var}}(\hat{u}(a))-\mathrm{Var}(a(X)\,\tilde{Y})=\frac{1}{N-1}\sum_{j=1}^{N}(a(X_j)\,\tilde{Y}_j-\hat{u}(a))^2-\mathrm{Var}(a(X)\,\tilde{Y})\xrightarrow{p}0$$

and thus $N\widehat{\mathrm{Var}}(\hat{u}(a))\xrightarrow{p}\mathrm{E}[a^2(X)\,\sigma^2(X)]>0$. Since also $\sqrt{N}u(a)=\mathrm{E}[a(X)\,\sqrt{N}\tau(X)]=\mathrm{E}[a(X)\,\mu(X)]$, we overall have that

$$\begin{aligned}\frac{\hat{u}(a)}{\sqrt{\widehat{\mathrm{Var}}(\hat{u}(a))}}&=\frac{\sqrt{N}(\hat{u}(a)-u(a))+\sqrt{N}u(a)}{\sqrt{N\widehat{\mathrm{Var}}(\hat{u}(a))}}\\&=\frac{\sqrt{N}(\hat{u}(a)-u(a))+\mathrm{E}[a(X)\,\mu(X)]}{\sqrt{N\widehat{\mathrm{Var}}(\hat{u}(a))}}\xrightarrow{d}\mathcal{N}\left(\frac{\mathrm{E}[a(X)\,\mu(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}},1\right).\end{aligned}$$

For the estimated propensity score $\hat{p}=\frac{\sum_{j=1}^{N}W_j}{N}$, for which $\sum_{j=1}^{N}\left(\frac{W_j}{\hat{p}}-\frac{1-W_j}{1-\hat{p}}\right)=0$,

$$\begin{aligned}&\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\frac{W_j}{\hat{p}}-\frac{1-W_j}{1-\hat{p}}\right)(Y_j-c(X_j))-\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\frac{W_j}{p}-\frac{1-W_j}{1-p}\right)(Y_j-c(X_j)-\mathrm{E}[Y_j-c(X_j)])\\&=\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\underbrace{\left(\left(\frac{W_j}{\hat{p}}-\frac{1-W_j}{1-\hat{p}}\right)-\left(\frac{W_j}{p}-\frac{1-W_j}{1-p}\right)\right)}_{\frac{W_j-\hat{p}}{\hat{p}(1-\hat{p})}-\frac{W_j-p}{p(1-p)}=\frac{W_j-p}{\hat{p}(1-\hat{p})}-\frac{W_j-p}{p(1-p)}+\frac{p-\hat{p}}{\hat{p}(1-\hat{p})}}(Y_j-c(X_j)-\mathrm{E}[Y_j-c(X_j)])\\&=\frac{\sqrt{N}(p-\hat{p})}{\hat{p}(1-\hat{p})}\overbrace{\frac{\sum_{j=1}^{N}(Y_j-c(X_j)-\mathrm{E}[Y_j-c(X_j)])}{N}}^{\xrightarrow{p}0}\\&\quad+\frac{\sqrt{N}(p(1-p)-\hat{p}(1-\hat{p}))}{\hat{p}(1-\hat{p})}\underbrace{\frac{\sum_{j=1}^{N}\frac{W_j-p}{p(1-p)}(Y_j-c(X_j)-\mathrm{E}[Y_j-c(X_j)])}{N}}_{\xrightarrow{p}0}\xrightarrow{p}0\end{aligned}$$

since the factors on the left are asymptotically Normal with mean zero. Hence, the two estimators are asymptotically equivalent. ■

Appendix F. Proofs from Section 4

F.1. Proof of Theorem 2

Proof The first property is the standard derivation of the un-biasedness of the doubly robust estimator with known propensities:

$$\begin{aligned}
\tau(x) :=\ & \mathrm{E}[\tilde{Y}_i \mid X_i = x] \\
=\ & \frac{\mathrm{E}[Y^{(1)} - c(X_i) \mid X_i = x]\,\Pr(W_i = 1 \mid X_i = x)}{p} - \frac{\mathrm{E}[Y^{(0)} - c(X_i) \mid X_i = x]\,\Pr(W_i = 0 \mid X_i = x)}{1-p} \\
=\ & \mathrm{E}[Y^{(1)} - c(X_i) \mid X_i = x] - \mathrm{E}[Y^{(0)} - c(X_i) \mid X_i = x] \\
=\ & \mathrm{E}[Y^{(1)} - Y_i^{(0)} \mid X_i = x]
\end{aligned}$$

$$\begin{aligned}
\nu^2(x;c) :=\ & \mathrm{E}[\tilde{Y}_i^2 \mid X_i = x] = \mathrm{E}[(Y_i^{(1)} - c(X_i))^2 \mid X_i = x]\frac{1}{p} + \mathrm{E}[(Y_i^{(0)} - c(X_i))^2 \mid X_i = x]\frac{1}{1-p} \\
=\ & \frac{\mathrm{Var}(Y^{(1)} \mid X = x)}{p} + \frac{\mathrm{Var}(Y^{(0)} \mid X = x)}{1-p} + \mathrm{E}[Y^{(1)} - Y^{(0)} \mid X = x]^2 \\
& + \frac{(c(x) - c_0(x))^2}{p(1-p)} - 2\,\mathrm{E}[Y^{(1)} - c(X_i)) \mid X_i = x]\frac{c(x) - c_0(x)}{p} \\
& - 2\,\mathrm{E}[Y^{(0)} - c(X_i) \mid X_i = x]\frac{c(x) - c_0(x)}{1-p} \\
=\ & \frac{\mathrm{Var}(Y^{(1)} \mid X = x)}{p} + \frac{\mathrm{Var}(Y^{(0)} \mid X = x)}{1-p} + \mathrm{E}[Y^{(1)} - Y^{(0)} \mid X = x]^2 \\
& + \frac{(c(x) - c_0(x))^2}{p(1-p)} + 2\tau(x)\,(c(x) - c_0(x)) \\
& - 2\tau(x)(c(x) - c_0(x)) \\
=\ & \frac{\mathrm{Var}(Y^{(1)} \mid X = x)}{p} + \frac{\mathrm{Var}(Y^{(0)} \mid X = x)}{1-p} + \mathrm{E}[Y^{(1)} - Y^{(0)} \mid X = x]^2 + \frac{(c(x) - c_0(x))^2}{p(1-p)} \\
=:\ & \sigma(x)^2 + \tau(x)^2 + \frac{(c(x) - c_0(x))^2}{p(1-p)}
\end{aligned}$$

Optimizing:

$$V(a) = \frac{\mathrm{E}[a(X)\tau(X)]}{\sqrt{\mathrm{E}[a(X)\sigma(X)^2] + \mathrm{E}[a(X)\tau(X)^2]}} \tag{27}$$

is equivalent to optimizing:

$$\frac{\mathrm{E}[a(X)\tau(X)]}{\sqrt{\mathrm{E}[a(X)\sigma(X)^2] + \mathrm{E}[a(X)\tau(X)^2] - \mathrm{E}[a(X)\tau(X)]^2}} = \frac{\mathrm{E}[a(X)\tau(X)]}{\sqrt{\mathrm{E}[a(X)\sigma(X)^2] + \mathrm{Var}(a(X)\tau(X))}} \tag{28}$$

■

F.2. Proof of Theorem 11

Proof Let $g_\lambda(S) := \sqrt{v(S)} - \lambda w(S)$. For the initial setting of $\lambda_{\max}$, observe that if we let $S_{\max} := \arg\min_{S\subseteq[m]} g_{\lambda_{\max}}(S)$, then $S_{\max} \supset \emptyset$ and $g_{\lambda_{\max}}(S_{\max}) < 0$, since by choosing $S = [m]$ we achieve $g_{\lambda_{\max}}(S) = \sqrt{v([m])} - \lambda_{\max} w([m]) \leq \epsilon\sqrt{v([m])} < 0 = g_{\lambda_{\max}}(\emptyset)$. Moreover, at any point where we decrease $\lambda_{\max}$ to $\bar{\lambda}$, we have that $\hat{S} \neq \emptyset$ and $\sqrt{v(\hat{S})} < \bar{\lambda} v(\hat{S})$, which implies that $g_{\bar{\lambda}}(S) < 0 = g_{\bar{\lambda}}(\emptyset)$. Thus the property that $S_{\max} \supseteq \emptyset$ is maintained with the new $\lambda_{\max}$. Hence, we conclude that throughout the algorithm this property is maintained.

Moreover, we have that at the initial setting of $\lambda_{\min}$ that

$$\min_{S\subseteq[m]} g_{\lambda_{\min}}(S) = \min_{S\subseteq[m]} \sqrt{v(S)} - \frac{\sqrt{\min_{x\in\mathcal{X}} v(x)}}{w([m])} w(S) \geq \min_{\emptyset\subset S\subseteq[m]} \sqrt{v(S)} - \sqrt{\min_{x\in\mathcal{X}} v(x)} \geq 0. \tag{29}$$

Moreover, at any point where we change $\lambda_{\min}$ to $\bar{\lambda}$ we know that either $\hat{S} = \emptyset$, in which case $\min_S g_{\bar{\lambda}}(S) = 0$ or that $\sqrt{v(\hat{S})} \geq \bar{\lambda} w(\hat{S})$, which implies that $g_{\bar{\lambda}}(\hat{S}) = \min_S g_{\bar{\lambda}}(S) \geq 0$. Thus at all points during the algorithm we have the property that $\min_{S\subseteq[m]} g_{\lambda_{\min}}(S) \geq 0$.

We now look at the values of the aforementioned quantities at the termination of the algorithm. From the second property we have that for any $S \supset \emptyset$ $g_{\lambda_{\min}}(S) \geq 0$. Since $w(S) > 0$ for any $S \supset \emptyset$, we then have that: $\min_{S\supset\emptyset} \frac{\sqrt{v(S)}}{w(S)} \geq \lambda_{\min}$. By the termination criterion of the algorithm we have that: $\lambda_{\min} \geq \frac{1}{1+\epsilon}\lambda_{\max}$. Finally, by the first property we have that $\hat{S} \supset \emptyset$ and $g_{\lambda_{\max}}(\hat{S}) < 0$, where $\hat{S}$ is the set returned by the algorithm. Thus we have that:

$$\frac{\sqrt{v(\hat{S})}}{w(\hat{S})} < \lambda_{\max} \leq (1+\epsilon)\lambda_{\min} \leq (1+\epsilon) \min_{S\supset\emptyset} \frac{\sqrt{v(S)}}{w(S)} \tag{30}$$

Equivalently, since $v(S) > 0$, for any $S \supset \emptyset$, we also have that:

$$\frac{w(\hat{S})}{\sqrt{v(\hat{S})}} > \frac{1}{\lambda_{\max}} \geq \frac{1}{1+\epsilon}\frac{1}{\lambda_{\min}} \geq \frac{1}{1+\epsilon} \max_{S\supset\emptyset} \frac{w(S)}{\sqrt{v(S)}} \tag{31}$$

Observe that after every period, we have that $\lambda_{\max} - \lambda_{\min}$ decreases by a factor of 2. Thus after $O\left(\log_2\left(\frac{\sqrt{v([m])}}{\epsilon\sqrt{\min_{x\in\mathcal{X}} v(x)}}\right)\right)$ iterations, we will have that $\lambda_{\max} - \lambda_{\min} \leq \epsilon\frac{\sqrt{\min_{x\in\mathcal{X}} v(x)}}{w([m])} \leq \epsilon\lambda_{\min}$. Thus $\lambda_{\max} \leq (1+\epsilon)\lambda_{\min}$ and the algorithm will terminate. ■

F.3. Proof of Theorem 12

Proof Let A be the space of all assignments from $\mathcal{X}$ to $\{0,1\}$ and observe that $|A| = 2^k$. Moreover, for any random variable X, let $\mathrm{E}_n[X] = \frac{1}{n}\sum_{i=1}^n X_i$, denote the empirical average. By Bennett's inequality and empirical Bennett's inequality Maurer and Pontil (2009), we have that w.p. $1 - 4\delta$,

for all $a \in A$:

$$\begin{aligned}
\left|\mathrm{E}_n[a(X)\tilde{Y}] - \mathrm{E}[a(X)\tilde{Y}]\right| &\leq \sqrt{\frac{\mathrm{E}_n[a(X)\tilde{Y}^2]\, 2\, k\, \log(2/\delta)}{n}} + H\frac{7\, k\, \log(2/\delta)}{3(n-1)} \\
\left|\mathrm{E}_n[a(X)\tilde{Y}^2] - \mathrm{E}[a(X)\tilde{Y}^2]\right| &\leq \sqrt{\frac{\mathrm{E}[a(X)\tilde{Y}^4]\, 2\, k\, \log(1/\delta)}{n}} + H\frac{k\, \log(2/\delta)}{3n} \\
&\leq H\sqrt{\frac{\mathrm{E}[a(X)\tilde{Y}^2]\, 2\, k\, \log(1/\delta)}{n}} + H\frac{k\, \log(2/\delta)}{3n}
\end{aligned}$$

Thus if we let $\epsilon_1(n,\delta) := \sqrt{\frac{2\, k\, \log(2/\delta)}{n}} + H\frac{7\, k\, \log(2/\delta)}{3(n-1)}$ we have that:

$$\frac{\mathrm{E}_n[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[\hat{a}(X)\tilde{Y}^2]}} \leq \frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[\hat{a}(X)\tilde{Y}^2]}} + \epsilon_1(n,\delta)$$

Moreover, observe that by the AM-GM inequality and thew aforementioned Bennett's inequality, for any $\eta \in (0,1)$:

$$\begin{aligned}
\mathrm{E}_n[a(X)\tilde{Y}^2] - \mathrm{E}[a(X)\tilde{Y}^2] &\geq -\eta\, \mathrm{E}[a(X)\tilde{Y}^2] - \frac{2H\, k\, \log(2/\delta)}{\eta\, n} \\
\implies \mathrm{E}_n[a(X)\tilde{Y}^2] &\geq (1-\eta)\, \mathrm{E}[a(X)\tilde{Y}^2] - \frac{2H\, k \log(2/\delta)}{\eta\, n}
\end{aligned}$$

If we let $\epsilon_2(n,\eta,\delta) := \frac{2H\, k \log(2/\delta)}{\eta\, n}$, then we have that:

$$\begin{aligned}
\frac{\mathrm{E}_n[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[\hat{a}(X)\tilde{Y}^2]}} &\leq \frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{(1-\eta)\, \mathrm{E}[\hat{a}(X)\tilde{Y}^2] - \epsilon_2(n,\eta,\delta)}} + \epsilon_1(n,\delta) \\
&\leq \frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{(1-\eta)\, \mathrm{E}[\hat{a}(X)\nu^2(X;c)] - \epsilon_2(n,\eta,\delta)}} + \epsilon_1(n,\delta) \\
&\leq \frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{(1-\eta)\, \mathrm{E}[\hat{a}(X)\nu^2(X;c_0)] - \epsilon_2(n,\eta,\delta)}} + \epsilon_1(n,\delta) \\
&\leq \frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{(1-\eta)\, \mathrm{E}[\hat{a}(X)\nu^2(X;c_0)]}} + \frac{1}{2\,(1-\eta)^{3/2}\underline{\pi}^{3/2}\,\underline{\nu}^3}\epsilon_2(n,\eta,\delta) + \epsilon_1(n,\delta)
\end{aligned}$$

Similarly, we have that:

$$\begin{aligned}
\frac{\mathrm{E}_n[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[a_*(X)\tilde{Y}^2]}} &\geq \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[a_*(X)\tilde{Y}^2]}} - \epsilon_1(n,\delta) \\
&\geq \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{(1+\eta)\,\mathrm{E}[a_*(X)\tilde{Y}^2] + \epsilon_2(n,\eta,\delta)}} - \epsilon_1(n,\delta) \\
&\geq \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{(1+\eta)\,\mathrm{E}[a_*(X)\nu^2(X;\hat{c})] + \epsilon_2(n,\eta,\delta)}} - \epsilon_1(n,\delta) \\
&\geq \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{(1+\eta)\,\mathrm{E}[a_*(X)\nu^2(X;\hat{c})] + \epsilon_2(n,\eta,\delta)}} - \epsilon_1(n,\delta) \\
&\geq \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{(1+\eta)\,\mathrm{E}[a_*(X)\nu^2(X;c_0)] + \frac{\|\hat{c}-c_0\|_2^2}{p(1-p)} + \epsilon_2(n,\eta,\delta)}} - \epsilon_1(n,\delta) \\
&\geq \frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{(1+\eta)\,\mathrm{E}[a_*(X)\nu^2(X;c_0)]}} - \frac{1}{2(1+\eta)^{3/2}\underline{\pi}^{3/2}\,\underline{\nu}^3}\left(\frac{\|\hat{c}-c_0\|_2^2}{p(1-p)} + \epsilon_2(n,\eta,\delta)\right) - \epsilon_1(n,\delta)
\end{aligned}$$

Finally, since:

$$\frac{\mathrm{E}_n[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[\hat{a}(X)\tilde{Y}^2]}} \geq \frac{1}{1+\epsilon}\frac{\mathrm{E}_n[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[a_*(X)\tilde{Y}^2]}}$$

Thus if we denote with:

$$\begin{aligned}
\kappa(n,\eta,\delta) &:= \frac{1}{2(1+\eta)^{3/2}\underline{\pi}^{3/2}\,\underline{\nu}^3}\left(\frac{\|\hat{c}-c_0\|_2^2}{p(1-p)} + \epsilon_2(n,\eta,\delta)\right) - \epsilon_1(n,\delta) \\
&= \sqrt{\frac{2\,k\,\log(2/\delta)}{n}} + \frac{7\,H\,k\,\log(2/\delta)}{3(n-1)} + \frac{1}{2(1+\eta)^{3/2}\underline{\pi}^{3/2}\,\underline{\nu}^3}\left(\frac{\|\hat{c}-c_0\|_2^2}{p(1-p)} + \frac{2H\,k\log(2/\delta)}{\eta\,n}\right)
\end{aligned}$$

We have that:

$$\frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[\hat{a}(X)\nu^2(X;c_0)]}} \geq \sqrt{\frac{1-\eta}{1+\eta}}\frac{1}{1+\epsilon}\frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}[a_*(X)\nu^2(X;c_0)]}} - 2\kappa(n,\eta,\delta)$$

Finally, noting that $\mathrm{E}[a(X)\tilde{Y}] = \mathrm{E}[a(X)\tau(X)]$ for all $a \in A$, yields the result. ■

F.4. Proof of Theorem 3

Proof Suppose a is an optimal policy and that we have two such elements $x, x' \in \mathcal{X}$, such that $a(x) = 1$ and $a(x') = 0$, while $\nu(x) = \nu(x') = \nu$ and $\tau(x) = \tau(x') = \tau$. First observe that for any x with $\tau(x) \leq 0$, any optimal policy must assign $a(x) = 0$. Thus we restrict attention to the case when $\tau > 0$. Let $f(x) = \Pr[X = x]$. Then we can write our objective as:

$$Z(a) = \frac{A + a(x)f(x)\tau}{\sqrt{B + a(x)f(x)\nu^2}}$$

where $A, B \geq 0$, since an optimal assignment never allocates to any $x \in \mathcal{X}$ with $\tau(x) \leq 0$.

We first consider the case where $B > 0$, i.e. when x is not the only element that is assigned the treatment. Since by removing x from assignment or adding x' to the assignment, we are deteriorating the objective, we must have that:

$$\begin{aligned} \frac{A + f(x)\tau}{\sqrt{B + f(x)\nu^2}} &> \frac{A}{\sqrt{B}} \\ \frac{A + f(x)\tau}{\sqrt{B + f(x)\nu^2}} &> \frac{A + (f(x) + f(x'))\tau}{\sqrt{B + (f(x) + f(x'))\nu^2}} \end{aligned} \tag{32}$$

Consider the function:

$$F(u) := \frac{A + u\tau}{\sqrt{B + u\nu^2}}$$

By the Equation (32), we must have that $F(0) < F(f(x))$ and $F(f(x)) > F(f(x) + f(x'))$. Since for $B > 0$, $F(u)$ is a smooth twice differentiable function in $[0, 1]$ and we have the latter inequalities, it means that there must be a point $u_* \in (0, f(x) + f(x'))$ at which it achieves a maximum. By the first order optimality condition at that point, we must have that:

$$u_*\nu^2\tau + 2B\tau - A\nu^2 = 0 \implies u_* = \frac{A\nu^2 - 2B\tau}{\nu^2\tau}$$

Since $u_* > 0$, we also have that $A\nu^2 - 2B\tau > 0 \implies A\nu^2 > 2B\tau$. Moreover, by the second order optimality condition we must have that:

$$u_*\nu^2\tau + 4B\tau - 3A\nu^2 \geq 0 \implies B\tau - A\nu^2 \geq 0 \implies A\nu^2 \leq B\tau$$

But for $\tau \geq 0$ and $B \geq 0$, the latter two inequalities cannot hold simultaneously, leading to a contradiction. Thus either removing x or adding x' must be a weakly improving change, leading to an optimal policy where x and x' are given the same assignment. Repeatedly applying the above argument, we can arrive to an assignment policy where any pair of x, x' that have $\nu_n(x) = \nu_n(x')$ and $\tau_n(x) = \tau_n(x')$ are assigned the same treatment.

Now consider the edge case when x is the only element assigned by the policy a. In this case since adding the element x' to the assignment leads to a deterioration we have:

$$\frac{f(x)\tau}{\sqrt{f(x)\nu^2}} > \frac{(f(x) + f(x'))\tau}{\sqrt{(f(x) + f(x'))\nu^2}} \implies \sqrt{f(x)} > \sqrt{f(x) + f(x')} \tag{33}$$

which is a contradiction. ■

F.5. Proof of Theorem 4

Proof Consider a partition $i \in [k]$, such that a subset $T_i \subseteq P_i$, with density t_i is treated, and the complement $U_i = P_i/T_i$, with density u_i is not treated by a_*. Let τ_i be the mean value of $\tau(x)$ for all elements in P_i, i.e. $\tau_i = \frac{1}{p_i}\sum_{x\in\mathcal{X}} f(x)\tau(x)$. Similarly, let ν_i^2 be mean value of $\nu(x)^2$. By our assumption, we know that for all $x \in P_i$: $p_i|\tau(x) - \tau_i| \leq \epsilon$ and $p_i|\nu^2(x) - \nu_i^2| \leq \epsilon$. Moreover,

denote with $\tau_{i,0}$ and $\nu^2_{i,0}$ the corresponding means on the untreated sub-population U_i and $\tau_{i,1}, \nu^2_{i,1}$ the means on the treated sup-population T_i. By our assumption we also have that for all $x \in P_i$ and for any $z \in \{0,1\}$: $p_i|\tau(x) - \tau_{i,z}| \leq \epsilon$ and $p_i|\nu^2(x) - \nu^2_{i,z}| \leq \epsilon$.

Let a_* be an optimal policy. We can write our objective as:

$$Z(a_*) = \frac{A + t_i\tau_{i,1}}{\sqrt{B + t_i\nu^2_{i,1}}}$$

where $A, B \geq 0$, since an optimal assignment never allocates to any $x \in \mathcal{X}$ with $\tau(x) \leq 0$. Moreover, A, B are quantities that do not depend on how the policy behaves for any $x \in P_i$.

We first consider the case where $B > 0$, i.e. when P_i is not the only partition that contains elements that are assigned the treatment. Since by removing T_i from assignment or adding U_i to the assignment, we are deteriorating the objective by more δ, we must have that:

$$\begin{aligned} \frac{A + t_i\tau_{i,1}}{\sqrt{B + t_i\nu^2_{i,1}}} &> \frac{A}{\sqrt{B}} + \delta \\ \frac{A + t_i\tau_{i,1}}{\sqrt{B + t_i\nu^2_{i,1}}} &> \frac{A + t_i\tau_{i,1} + u_i\tau_{i,0}}{\sqrt{B + t_i\nu^2_{i,1} + u_i\nu^2_{i,0}}} + \delta \end{aligned} \tag{34}$$

By the aforementioned approximation properties we have that:

$$\begin{aligned} \frac{A + t_i\tau_{i,1}}{\sqrt{B + t_i\nu^2_{i,1}}} &\leq \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu^2_i}} + \frac{t_i(\tau_{i,1} - \tau_i)}{\sqrt{B + t_i\nu^2_i}} + \frac{A + t_i\tau_{i,1}}{\sqrt{B + t_i\nu^2_{i,1}}\sqrt{B + t_i\nu^2_i}}\left(\sqrt{B + t_i\nu^2_{i,1}} - \sqrt{B + t_i\nu^2_i}\right) \\ &= \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu^2_i}} + \frac{t_i(\tau_{i,1} - \tau_i)}{\sqrt{B + t_i\nu^2_i}} + \frac{A + t_i\tau_{i,1}}{\sqrt{B + t_i\nu^2_{i,1}}\sqrt{B + t_i\nu^2_i}} \frac{t_i\left(\nu^2_{i,1} - \nu^2_i\right)}{\sqrt{B + t_i\nu^2_{i,1}} + \sqrt{B + t_i\nu^2_i}} \\ &= \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu^2_i}} + \frac{\sqrt{t_i}(\tau_{i,1} - \tau_i)}{\sqrt{\nu^2_i}} + \frac{\tau_{i,1}}{\nu^2_i}\frac{\sqrt{t_i}\left(\nu^2_{i,1} - \nu^2_i\right)}{\sqrt{\nu^2_i}} \\ &\leq \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu^2_i}} + \frac{\epsilon}{\underline{\nu}} + \frac{\bar{\kappa}}{\underline{\nu}}\epsilon \\ &\leq \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu^2_i}} + \frac{\epsilon}{\underline{\nu}}(1 + \bar{\kappa}) \end{aligned}$$

Similarly, we have that:

$$\frac{A + t_i\tau_{i,1} + u_i\tau_{i,0}}{\sqrt{B + t_i\nu^2_{i,1} + u_i\nu^2_{i,0}}} \geq \frac{A + p_i\tau_i}{\sqrt{B + p_i\nu^2_i}} - \frac{2\epsilon}{\underline{\nu}}(1 + \bar{\kappa})$$

Thus for $\delta \geq \frac{2\epsilon}{\underline{\nu}}(1+\bar{\kappa})$, we can combine the latter inequalities to conclude that:

$$\begin{aligned} \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu_i^2}} &> \frac{A}{\sqrt{B}} \\ \frac{A + t_i\tau_i}{\sqrt{B + t_i\nu_i^2}} &> \frac{A + p_i\tau_i}{\sqrt{B + p_i\nu_i^2}} \end{aligned} \tag{35}$$

Using an argument identical to that in the proof of Theorem 3, we can show that the latter leads to a contradiction. Thus either assigning all of the partition P_i to the treatment or removing all of the partition P_i from the treatment, can lead to a new policy which achieves and objective $Z(a)$ at least an additive δ smaller than the optimal policy.

Repeating this process for every partition $i \in [k]$, we can arrive at a new policy that achieves an objective that is at most an additive $k\delta$ smaller than that of the optimal policy and in which all partitions have a uniform assignment to either treatment or non-treatment. ■

F.6. Proof of Theorem 6

Proof The result for τ is an immediate consequence of Theorem C.10.2 of Syrgkanis and Zampetakis (2020). For the case of ν^2 we need to make a small adjustment to account for the fact that $\hat{c} \neq c_0$. In particular, even though $\nu^2(x; c_0)$ might satisfy the (β, r)-strong sparsity condition, the function $\nu^2(x; \hat{c})$ might not. Moreover, we want to guarantee that $\nu^2(x; c_0) = \nu^2(x'; c_0)$ for every x, x' that fall in the leaf of the constructed tree and not that $\nu^2(x; \hat{c}) = \nu^2(x; \hat{c})$.

However, we note that the main argument in Theorem C.10.2 of Syrgkanis and Zampetakis (2020) is robust to small errors. In particular, Lemma C.1 of Syrgkanis and Zampetakis (2020), shows that the empirical splitting criterion $\bar{V}_n(S; \hat{c})$ for every possible split up to depth d, is within an $\epsilon(n, \delta)$ of it's population counterpart w.p. $1 - \delta$, for $\epsilon(n, \delta) = O\left(\sqrt{\frac{2^d d \log(p\, d) + \log(1/\delta))}{n}}\right)$. Moreover, the population criterion is:

$$\bar{V}(S; \hat{c}) := \mathrm{E}[\mathrm{E}[\nu^2(X; \hat{c}) \mid X_S]^2] \tag{36}$$

where we overloaded notation and denoted with $\bar{V}(S; c)$ the function $\bar{V}(S; \nu(\cdot; c))$. Since by Theorem 2 $\nu^2(x; \hat{c}) = \nu^2(x; c_0) + \frac{(\hat{c}(x) - c_0(x))^2}{p(1-p)}$, we thus have that:

$$\begin{aligned} \left|\bar{V}(S; \hat{c}) - \bar{V}(S; c_0)\right| = &\ \left|\mathrm{E}\left[\mathrm{E}[\nu^2(X; c) \mid X_S]^2 - \mathrm{E}[\nu^2(X; c) \mid X_S]^2\right]\right| \\ \leq &\ 2\,\mathrm{E}\left[\left|\mathrm{E}[\nu^2(X; c) \mid X_S] - \mathrm{E}[\nu^2(X; c) \mid X_S]\right|\right] \\ \leq &\ 2\,\mathrm{E}\left[\frac{(\hat{c}(x) - c_0(x))^2}{p(1-p)}\right] = \frac{2\|\hat{c} - c_0\|_2^2}{p(1-p)} \end{aligned}$$

Thus as long as $\epsilon(n, \delta) + \frac{2\|\hat{c} - c_0\|_2^2}{p(1-p)} < \frac{\beta}{2}$, then we have that w.p. $1-\delta$, at every step of the greedy splitting algorithm we will only be choosing a variable in the relevant set R for the function $\nu^2(x; c_0)$,

since for any $i \in R$ and $j \in [p]/R$ and for any $T \subseteq [p]/\{i\}$:

$$\begin{aligned}
\bar{V}_n(\{i\} \cup T; \hat{c}) &\geq \bar{V}(\{i\} \cup T; \hat{c}) - \epsilon(n, \delta) \\
&\geq \bar{V}(\{i\} \cup T; c_0) - \epsilon(n, \delta) - \frac{2\|\hat{c} - c_0\|_2^2}{p(1-p)} \\
&\geq \bar{V}(\{j\} \cup T; c_0) + \beta - \epsilon(n, \delta) - \frac{2\|\hat{c} - c_0\|_2^2}{p(1-p)} \\
&> \bar{V}(\{j\} \cup T; c_0) + \beta/2 \\
&\geq \bar{V}(\{j\} \cup T; \hat{c}) + \beta/2 - \frac{2\|\hat{c} - c_0\|_2^2}{p(1-p)} \\
&\geq \bar{V}_n(\{j\} \cup T; \hat{c}) + \beta/2 - \epsilon(n, \delta) - \frac{2\|\hat{c} - c_0\|_2^2}{p(1-p)} \\
&> \bar{V}_n(\{j\} \cup T; \hat{c})
\end{aligned}$$

Thus variable i will be chosen over variable j. ■

F.7. Proof of Theorem 7

Proof By Section A.2 of Syrgkanis and Zampetakis (2020), the VC dimension of A is at most $v = 4d2^d + 2^{d+1}\log(2p)$. Thus the growth number of this class by Sauer's lemma is at most $\tau(n) = (e\,n/v)^v$. Hence, the ℓ_∞ empirical covering number of A on n samples is at most $\tau(n) = (e\,n/v)^v$. Thus we have by the empirical and population Bennett inequality of Maurer and Pontil (2009) that w.p. $1-\delta$:

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n a(X_i)\,\tilde{Y}_i - \mathrm{E}[a(X)\tilde{Y}]\right| &\leq \sqrt{\frac{\mathrm{E}_n[a(X)\nu(X;c)^2]\,2\,v\log(2\,e\,n/\delta)}{n}} + H\frac{7\,v\log(2\,e\,n/\delta)}{3(n-1)} \\
\left|\mathrm{E}_n[a(X)\tilde{Y}^2] - \mathrm{E}[a(X)\tilde{Y}^2]\right| &\leq \sqrt{\frac{\mathrm{E}[a(X)\tilde{Y}^4]\,2\,v\,\log(e\,n/\delta)}{n}} + H\frac{v\,\log(2e\,n/\delta)}{3n} \\
&\leq H\sqrt{\frac{\mathrm{E}[a(X)\tilde{Y}^2]\,2\,v\,\log(e\,n/\delta)}{n}} + H\frac{v\,\log(2\,e/\delta)}{3n}
\end{aligned}$$

Finally, observe that $\mathrm{E}[a(X)\tilde{Y}] = \mathrm{E}[a(X)\tau(X)]$ and $\mathrm{E}[a(X)\tilde{Y}^2] = \mathrm{E}[a(X)\nu^2(X;c)]$. ■

F.8. Proof of Theorem 8

Proof By Theorem 6, we have that w.p. $1-\delta$, the partitions defined by the leafs of the two trees will contain x's that all have the same value of $\tau(x)$ and $\nu^2(x;c_0)$. Thus the policy space that is being searched over in the last step of the algorithm contains an optimal policy a_*. Moreover, combining Theorem 7 and Theorem 2 and a proof almost identical to that of Theorem 12 (replacing $\underline{\pi}$ with $\zeta/2^r$ and $k\log(2/\delta)$ with $v\log(2\,e\,n/\delta)$, where $v = 4r2^r + 2^{r+1}\log(2p)$), we can show that if we let:

$$\kappa(n,\eta,\delta) := \sqrt{\frac{2\,v\,\log(2\,e\,n/\delta)}{n}} + \frac{7\,H\,v\,\log(2e\,n/\delta)}{3(n-1)} + \frac{2^{3r/2}}{2(1+\eta)^{3/2}\zeta\,\underline{\nu}^3}\left(\frac{\|\hat{c}-c_0\|_2^2}{p(1-p)} + \frac{2H\,v\log(2e\,n/\delta)}{\eta\,n}\right)$$

then $\hat{a}$ satisfies w.p. $1-\delta$:

$$\frac{\mathrm{E}[\hat{a}(X)\tilde{Y}]}{\sqrt{\mathrm{E}_n[\hat{a}(X)\nu^2(X;c_0)]}} \geq \sqrt{\frac{1-\eta}{1+\eta}}\frac{1}{1+\epsilon}\frac{\mathrm{E}[a_*(X)\tilde{Y}]}{\sqrt{\mathrm{E}[a_*(X)\nu^2(X;c_0)]}} - 2\kappa(n,\eta,\delta) \tag{37}$$

■

Appendix G. Proofs from Section A

G.1. Proof of Theorem 9

Proof For $\alpha, \delta > 0$ let K be such that $2^k\,\Phi(-K) < \delta/3$, L such that $\Phi(-L+z_{1-\alpha}) < \delta/3$, and $\varepsilon > 0$ such that $\Phi(\varepsilon - z_{1-\alpha}) < \alpha + \delta$. Fix $(\mu_x)_{x=1}^k$ with $\mu_x > 0$ for all x, and let $C := \max\left\{\frac{2K+L}{\min_x \mu_x}, \frac{2k\max_x \mu_x}{\varepsilon}\right\}$. For any fixed $(\sigma_x^2)_{x=1}^k$ with $\min_x \sigma_x^2 \leq \frac{1}{C}, \max_x \sigma_x^2 > C$ and any fixed assignment $a \in \{0,1\}^k$, we have that

$$\begin{aligned}
&\frac{\frac{1}{k}\sum_{x=1}^k a_x\hat{\tau}_x}{\sqrt{\frac{1}{k}\sum_{x=1}^k a_x\frac{\hat{\sigma}_x^2}{n/k}}} - \frac{\sum_{x=1}^k a_x\mu_x}{\sqrt{\sum_{x=1}^k a_x\sigma_x^2}} \\
&= \frac{\sum_{x=1}^k a_x(\sqrt{n}\hat{\tau}_x - \mu_x)}{\sqrt{\sum_{x=1}^k a_x\hat{\sigma}_x^2}} + \frac{\sum_{x=1}^k a_x\mu_x}{\sqrt{\sum_{x=1}^k a_x\sigma_x^2}\sqrt{\sum_{x=1}^k a_x\hat{\sigma}_x^2}}\underbrace{\left(\sqrt{\sum_{x=1}^k a_x\hat{\sigma}_x^2} - \sqrt{\sum_{x=1}^k a_x\sigma_x^2}\right)}_{\xrightarrow{p}0} \xrightarrow{d} \mathcal{N}(0,1)
\end{aligned} \tag{38}$$

as $n \to \infty$ on the training sample. Since there are only 2^k assignments,

$$\begin{aligned}
&\mathrm{P}\left(\max_{a\in\{0,1\}^k}\left|\frac{\frac{1}{k}\sum_{x=1}^k a_x\hat{\tau}_x}{\sqrt{\frac{1}{k}\sum_{x=1}^k a_x\frac{\hat{\sigma}_x^2}{n/k}}} - \frac{\sum_{x=1}^k a_x\mu_x}{\sqrt{\sum_{x=1}^k a_x\sigma_x^2}}\right| < K\right) \\
&\geq 1 - \sum_{a\in\{0,1\}^k}\mathrm{P}\left(\left|\frac{\frac{1}{k}\sum_{x=1}^k a_x\hat{\tau}_x}{\sqrt{\frac{1}{k}\sum_{x=1}^k a_x\frac{\hat{\sigma}_x^2}{n/k}}} - \frac{\sum_{x=1}^k a_x\mu_x}{\sqrt{\sum_{x=1}^k a_x\sigma_x^2}}\right| \geq K\right) > 1 - \frac{\delta}{2}
\end{aligned}$$

for n sufficiently large. Hence, with probability at least $1-\delta/2$ and for large n,

$$\frac{\sum_{x=1}^k \hat{a}_x^{(B)}\mu_x}{\sqrt{\sum_{x=1}^k \hat{a}_x^{(B)}\sigma_x^2}} \geq \max_{a\in\{0,1\}^k}\frac{\sum_{x=1}^k a_x\mu_x}{\sqrt{\sum_{x=1}^k a_x\sigma_x^2}} - 2K \geq \frac{\min_x \mu_x}{\min_x \sigma_x} - 2K \geq C\min_x \mu_x - 2K \geq L.$$

By (38) applied to the test sample, on that sample

$$\frac{\frac{1}{k}\sum_{x=1}^k \hat{a}_x^{(B)}\hat{\tau}_x}{\sqrt{\frac{1}{k}\sum_{x=1}^k \hat{a}_x^{(B)}\frac{\hat{\sigma}_x^2}{n/k}}} > z_{1-\alpha}$$

with probability at least $1-\delta/2$ for N sufficiently high. Taken together, we reject with probability at least $1-\delta$ once n, N are sufficiently high.

On the other hand, for the simple classifier, $\mathrm{P}(\hat{a}_x^{(1)} = 1) > \frac{1}{2}$ for all x and n sufficiently large, including for x some x with $\sigma_x > C$. In that case,

$$\frac{\sum_{x=1}^k \hat{a}_x^{(A)} \mu_x}{\sqrt{\sum_{x=1}^k \hat{a}_x^{(A)} \sigma_x^2}} \leq \frac{k \max_x \mu_x}{C} < \varepsilon/2$$

and we therefore reject with overall probability at most

$$\frac{1}{2} + \frac{1}{2}\Phi(\varepsilon - z_{1-\alpha}) < \frac{1+\alpha}{2} + \delta$$

as n, N sufficiently large. ■

Appendix H. Proofs from Section C

H.1. Proof of Theorem 10

Proof U in Theorem 1 is monotonically increasing in

$$\frac{\mathrm{E}[a(X)\,\tau(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}}.$$

By $a(x) \geq 0$ and the Cauchy–Schwartz inequality,

$$\mathrm{E}[a(X)\,\tau(X)] \leq \mathrm{E}[a(X)\sigma(X)\,\tau_+(X)/\sigma(X)] \leq \sqrt{\mathrm{E}[a^2(X)\sigma^2(X)]}\sqrt{\mathrm{E}[\tau_+^2(X)/\sigma^2(X)]}$$

and hence

$$\frac{\mathrm{E}[a(X)\,\tau(X)]}{\sqrt{\mathrm{E}[a^2(X)\,\sigma^2(X)]}} \leq \sqrt{\mathrm{E}[\tau_+^2(X)/\sigma^2(X)]},$$

where the upper bound does not depend on a and is attained at $a(x) = \tau_+(X)/\sigma^2(X)$. ■