Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

February 28, 2024

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Statistical Inference on Heterogeneous Treatment Effects

14

"Never cross a river that is on average four feet deep."

– Nassim Nicholas Taleb [1].

We study estimation and inference on heterogeneous treatment effects. We introduce DML for inference on heterogenous treatment effects. We first review conditional and group average treatment effects as methods to analyse differences in the impact of treatment arising from the value of covariates. We show how these effects can be estimated using OLS. We illustrate the approach using the 401(k) example. We then consider more flexible inference on heterogeneous effects using adaptations of Random Forest methods, known as Causal Forests and illustrate the approach with an application on a large social science experiment studying the effect of the use of the word "welfare" in policy documents, on public perception.

14.1 CATEs under Conditional Exogeneity

We consider the standard setup for analyzing the effect of a binary treatment in the presence of a high-dimensional set of controls Z. Specifically, we have potential outcomes $Y(0)$ and $Y(1)$ and assigned treatment D that obey the conditional exogeneity condition:

We focus on the binary treatment case, but note that the approach readily extends to more general settings.

$$D \perp\!\!\!\perp Y(d) \mid Z.$$

We observe the outcome $Y := Y(D)$, the treatment assignment D, and the high-dimensional set of controls Z.

Our main interest in this section is the Conditional Average Treatment Effect (CATE) defined as

$$\tau_0(X) = \mathrm{E}[Y(1) - Y(0) \mid X],$$

where X is (typically) a low-dimensional subset of covariates Z. We have already seen in prior sections that under conditional exogeneity, the conditional average treatment effect is identified by the conditional average predictive effect (c.f. Theorem 5.2.1), which leads to the simple identification equation:

$$\tau_0(X) = \mathrm{E}[\mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z] \mid X] \qquad (14.1.1)$$

The value of CATE estimation So far in our analysis we have primarily been focusing on average causal effects. However, average effects are not informative of *whom to treat*. At best they can inform uniform policies, where we decide whether to roll out or not a new treatment on the whole population. Such uniform policies can have two major drawbacks. If the average effect is significantly positive and we decide to uniformly deploy the treatment, then there could potentially exist sub-groups in the population for which the treatment can have severe adverse effects. Analogously, if the average effect is significantly negative or a statistical null, then we might choose not to deploy a new policy or treatment. However, there could exist *responder sub-groups* in the population, for which the new treatment can have a significant positive impact. In both cases, by focusing on average causal effects, we are causing harm on sub-groups of the population, either by depriving of or forcing a new treatment.

Conditional average treatment effects allow us to identify such heterogeneities of the effect and discover in a data-driven manner the sub-groups of the population for which the treatment can be harmful or beneficial. Good estimates of the CATE, allows

us to deploy personalized policies; personalizing the offered treatment based on observable characteristics of each unit. For this reason, the study of CATE estimation has become increasingly more widespread, especially in settings where we have rich datasets, with many informative covariates and in many application domains; with a frontrunner domain being digital experimentation, where datasets are rich and personalization is easily implementable and deployable.

The hardness of CATE estimation From a statistical viewpoint, estimation and inference on the CATE is inherently harder than estimation and inference of average effects. So far, most of the policy relevant target parameters that we have been interested in, take the form of some low-dimensional vector valued parameter. This is the first time, where our target causal parameter of interest is actually a function or the value of a function at a particular point. The closest estimation problem to the CATE is that of estimating a Best Prediction rule or a Conditional Expectation Function. Note that even if we had access to both counterfactuals $Y(1), Y(0)$, then estimation of the CATE is as hard as estimating a regression function corresponding to the outcome $Y(1) - Y(0)$. For such problems, thus far we were content at estimating them with respect to the mean-squared-error metric, and at a reasonable rate that decays to zero, potentially slower than the parametric rate of $n^{-1/2}$. On the contrary for most causal effects of interest, we were not really content with simply a mean-squared error rate; we typically sought the ability to construct confidence intervals and were striving for very accurate estimation, most of the times at parametric rates.

For this reason, when it comes to CATE estimation, we will need to re-calibrate our expectations and potentially relax our goals. In this and the next chapter, we will consider four such avenues:

- Target the estimation of the best linear approximation (BLA) of the CATE function, with a set of predefined low-dimensional engineered features. In this case, we can essentially recover all the desiderata of target causal quantities: estimation at parametric rates, confidence intervals for the BLA at a particular point and even simultaneous confidence bands for the BLA at a set of target evaluation points.
- Target inference on other summarizations of CATE such as its tail expectations, the value of a covariate-based

treatment policy, and the value of the optimal such policy. Again, we recover the desiderata of target causal quantities.

- Construct non-parametric confidence intervals for CATE predictions at a particular point, using novel methods (such as Causal Forests), which are practically powerful and marry machine learning techniques with uncertainty quantification, but which are theoretically valid only when X is low-dimensional, and which in practice can be more brittle and are not as *assumption-lean* as inference based on OLS.
- Drop our desire to produce confidence intervals on the CATE function and only require good accuracy of the learned CATE function as captured by the mean-squared-error metric. In this case, we will be essentially treating the CATE problem as a best prediction problem and we will need to develop analogous methods for model selection, ensembling and out-of-sample evaluation. To compensate for the lack of confidence intervals for the CATE predictions, we will develop hypothesis tests that can be performed out-of-sample, that act as validation metrics that measure the quality of the CATE model as whole, as summarized in particular dimensions. For instance, we can test out of sample, whether the model picked up any statistically significant signal of heterogeneity, or if we use the model to prioritize treatment among the population, then will it lead to statistically significant policy gains.
- Drop the emphasis on learning the effect heterogeneity and focus only on the value of personalized policies that come out of our estimation process. In this case, we view CATE only as a means to our goal of designing personalized policies and in that respect we might want to measure the quality of our process, solely based on the personalized policy gains over some baseline, and not on the accuracy of the magnitude of the effect. Note that to learn a good policy, we are primarily interested in learning the sign of the effect and not necessarily its magnitude and appropriately partitioning the population such that the sign of the effect is relatively homogeneous within each sub-group. From this perspective, learning a good policy is more akin to a *classification* problem (classifying for which parts of the population the effect is positive/negative) as opposed to a regression problem and we will investigate such a formal equivalence.

14.2 Inference on Best Linear Approximations

Our main goal is summarizing the potentially complex and high-dimensional treatment effect function, which may depend on the entire vector Z, in terms of a lower-dimensional object X. We may be interested in such summaries for aiding interpretation or for policy reasons where we are interested in effects among particular recipients defined by observable characteristics.

For example, in the context of the 401(K) analysis from previous chapters, we have that Y is a household's total net financial assets, D is 401(k) eligibility status, and Z is the entire set of household characteristics. We might then take X to be income in which case the CATE $\tau_0(X)$ shows the expected effect of 401(k) eligibility on total financial assets for a subject whose income level is X.

The key to adaptively estimating and potentially performing inference for the CATE is expressing it as a conditional expectation of an unbiased signal:

$$\tau_0(X) = \mathrm{E}[Y(\eta_0) \mid X],$$

where the signal takes the form

$$Y(\eta) = H(\mu)\ (Y - g(D, Z)) + g(1, Z) - g(0, Z),$$

with nuisance parameters $\eta := (\mu, g)$ and

$$H(\mu) := \frac{D}{\mu(Z)} - \frac{1 - D}{1 - \mu(Z)}.$$

Here, $g(D, Z)$ and $\mu(Z)$ are square integrable functions with $\mu(Z)$ taking on values in $[\epsilon, 1 - \epsilon]$ for some $\epsilon > 0$. The true values of these nuisance parameters are $\eta_0 := (\mu_0, g_0)$ defined as

$$\mu_0(Z) := \mathrm{P}(D = 1 \mid Z), \quad g_0(D, Z) := \mathrm{E}[Y \mid Z, D].$$

Importantly, the signal has the Neyman orthogonality property:

$$\partial_\eta \mathrm{E}[Y(\eta_0) \mid X] = 0.$$

Making use of the representation of the CATE as the conditional expectation of $Y(\eta_0)$, we then estimate the CATE using the following steps:

Generic DML for CATE

1. Partition data indices into k folds of approximately equal size: $\{1, ..., n\} = \cup_{k=1}^{K} I_k$. For each fold $k = 1, ..., K$, compute ML estimators $\hat{g}_{[k]}(D, Z)$ and $\hat{\mu}_{[k]}(Z)$ of the best predictors $g_0(D, Z)$ and $\mu_0(Z)$ leaving out the k-th block of data. For any observation $i \in I_k$, define

$$\begin{aligned} Y_i(\widehat{\eta}) &= Y_i(\widehat{\eta}_k) \\ &= H_i(Y_i - \hat{g}_{[k]}(D_i, Z_i)) + \hat{g}_{[k]}(1, Z_i) - \hat{g}_{[k]}(0, Z_i) \end{aligned}$$

where $H_i = \dfrac{D_i}{\hat{\mu}_{[k]}(Z_i)} - \dfrac{1 - D_i}{1 - \hat{\mu}_{[k]}(Z_i)}$.

2. Use low-dimensional or high-dimensional regression methods to regress $Y_i(\hat{\eta})$ on covariate features X_i. If low-dimensional methods are used, inference on CATE can proceed using standard results for low-dimensional methods.

Under regularity conditions, the second step is adaptive, meaning all the learning guarantees and confidence intervals are approximately the same as if we knew the nuisance parameters η_0. This adaptation holds true because of the conditional Neyman orthogonality of $Y(\eta)$. We note that this adaptivity *does not imply* that inferential objects, e.g. confidence intervals, can readily be obtained if high-dimensional methods are used in Step 2. We discuss implementation and inferential issues in more detail in the following sections.

Least Squares Methods for Learning CATEs

Here we focus on using least squares in the second step of the general approach given above.

Consider approximating or summarizing the function $t(x)$ by a linear combination of basis functions:

$$p(x)'\beta_0,$$

where $p(x)$ is d-dimensional dictionary with

$$d \ll n.$$

For example, $p(x)$ could be a vector of group indicators or a vector of orthogonal polynomials or splines.

The parameter β_0 is chosen to minimize the approximation error to the CATE:

$$\min_{\beta} \mathrm{E}(\tau_0(X) - p(X)'\beta)^2.$$

$p(x)'\beta_0$ is thus the best linear predictor for the CATE; that is,

$$\beta_0 = \left(\mathrm{E}\left[p(X)p(X)\right]\right)^{-1} \mathrm{E}\left[p(X)Y(\eta_0)\right].$$

An important, easily interpretable special case is when we choose to use group indicators in forming the basis functions $p(x)$. Specifically, we define group indicators as

$$G_k(X) = 1(X \in R_k),$$

where $R_k's$ are mutually exclusive regions in the covariate space. For example, in the 401(k) example, we may be interested in average treatment effects for observations with household income less than \$10,000, observations with income between \$10,000 and \$20,000, etc. which we could capture by defining $G_1(X) = 1(\text{Income} < \$10,000)$, $G_2(X) = 1(\$10,000 \leq \text{Income} < \$20,000)$, etc. With the group indicators defined, we then set

$$p(X) = (G_1(X), \ldots, G_K(X))'.$$

In this case, the Best Linear Predictor β_0 recovers the GATEs (group average treatment effects).

More generally, $p(x) \in \mathbb{R}^d$ represents a d-dimensional dictionary of series/sieve basis functions – e.g., polynomials or splines – and $p(x)'\beta_0$ corresponds to the best linear approximation to the target function $\tau_0(x)$ in the given dictionary. Under some smoothness conditions, $\pi(x) = p(x)'\beta_0$ will approximate $\tau_0(X)$ as the dimension of the dictionary becomes large, and our inference will target this function.

Taking the approach motivated above to a sample of data, we have that the natural estimator of the best linear predictor of the CATE is

$$p(x)'\widehat{\beta},$$

where $\widehat{\beta}$ is the ordinary least squares estimate of β_0 defined on

the random sample $(X_i, D_i, Y_i)_{i=1}^N$:

$$\widehat{\beta} := \left(\frac{1}{N} \sum_{i=1}^{N} p(X_i) p(X_i)' \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} p(X_i) Y_i(\widehat{\eta}).$$

Semenova et al. [2] derive a complete set of results for the properties of $p(x)'\widehat{\beta}$ as an estimator of the best linear predictor curve $x \mapsto p(x)'\beta_0$. Importantly, these results establish an asymptotic approximation that allows simultaneous inference on all parameters of the best linear predictor curve. The key result verifies that the large sample properties of $\widehat{\beta}$ are the same as those of

$$\bar{\beta} := \left(\frac{1}{N} \sum_{i=1}^{N} p(X_i) p(X_i)' \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} p(X_i) Y_i(\eta_0),$$

when ML tools are used to estimate the nuisance parameter η_0 so long as the ML tools perform sufficiently well. Thus, we can employ standard methods for inference about β_0 and the best linear predictor curve functional $x \mapsto p(x)'\beta_0$.

Specifically, leveraging that $\widehat{\beta}$ and $\bar{\beta}$ have the same large sample properties, we have

$$\hat{\beta} - \beta_0 \sim_a N(0, \hat{\Omega}/N),$$

where

$$\widehat{\Omega} := \widehat{Q}^{-1} \left[\mathbb{E}_n p(X_i) p(X_i)' (Y_i(\widehat{\eta}) - p(X_i)'\widehat{\beta})^2 \right] \widehat{Q}^{-1} \qquad (14.2.1)$$

for $\widehat{Q} = \mathbb{E}_n p(X_i) p(X_i)'$.

This result can be used to construct uniform confidence bands for

$$x \mapsto p(x)'\beta_0,$$

which can be interpreted as confidence intervals for CATE $x \mapsto \tau_0(x)$ if the approximation error is small.

Application to 401(k) Example

R Notebook for DML on CATE analyzes the ATE of 401(K) conditional on income.

We illustrate estimation of CATEs and GATEs by revisiting the 401(k) example. Here, we consider the effect of 401(k) eligibility on net total financial assets controlling for household characteristics. We consider heterogeneity of this effect as a function of income. We consider two different ways to summarize these heterogeneous effects: GATEs based on coarse income categories

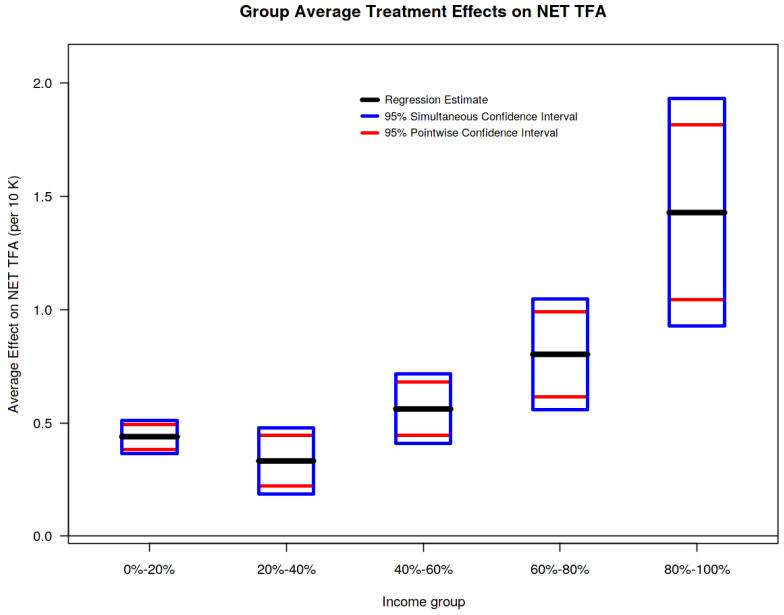

Figure 14.1: Inference on ATE of 401(k) Eligibility by Income Group

and a summary of the CATE given income based on a collection of polynomial terms in log(Income).

We show estimates and confidence bands on GATEs by income groups in Figure 14.1. Here, groups correspond to income quintiles; e.g the first group has households with income smaller than the 20^{th} percentile, the second group has households with income between the 20^{th} and 40^{th} percentile, and so on. Point estimates are provided by the central solid black bands. We represent pointwise confidence bands with the red lines in the interior of the box for each GATE. These bands would be appropriate for inference if one were interested *ex ante* in a single, pre-specified GATE. For example, one might be specifically interested in the eligibility effect among low income individuals and thus focus on the pointwise intervals over the first GATE. Finally, uniform confidence bands are given by the upper and lower bounds of the box for each GATE. These uniform bands provide a coverage guarantee for all five reported GATEs and would be appropriate for inference in settings where one was interested in all five effects and did not *ex ante* have a single specific GATE of interest.

We illustrate using a polynomial in log income to approximate the CATE in Figure 14.2. Point estimates are given by the central black line while the blue lines provide confidence bands. The narrower – dashed – confidence bands are pointwise and would be appropriate for a scenario in which one had a single, pre-specified value of income of interest. The wider confidence bands are uniform, providing a coverage guarantee for the *entire*

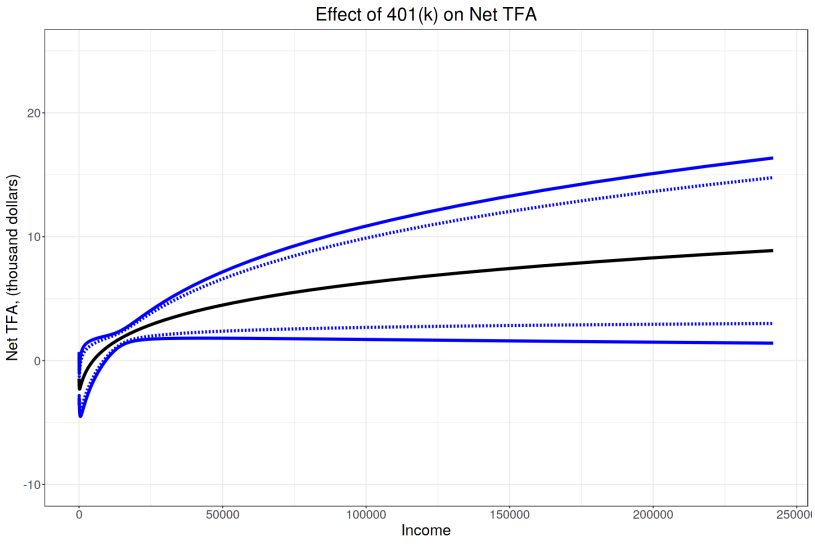

Figure 14.2: Inference on CATE of 401(k) Eligibility Conditional on Log-Income

best linear predictor curve $x \mapsto p(x)\beta_0$. That is, any path for the entire curve that would not be rejected will lie entirely within the uniform confidence band. Finally, note that the coverage guarantee extends to the true CATE function $x \mapsto \tau_0(x)$ if the approximation error of the polynomial to the true CATE is small.

14.3 Personalized Policies and Inference on Their Values

At the opening of this section we hinted that one reason why one might want to estimate a CATE model is so as to deploy a more personalized or contextual policy or to stratify and prioritize the treatment assignment, so as to maximize the outcome of interest. We formalize a personalized treatment policy π as function that given any instance of the variable X returns a probability $\pi(X) \in [0, 1]$ with which we to give treatment or not. Note that if the probability is 1 or 0 it is a deterministic assignment to treat or not treat. We are interested in conducting inference on the value of any policy π and in particular the maximum such possible value.

Given any policy π, we define its value as its gain in the average outcome if we were to follow π's treatment recommendation for everyone in the population compared to treating no one:

$$\begin{aligned} V(\pi) &:= \mathrm{E}[\pi(X)\, Y(1) + (1 - \pi(X))\, Y(0)] - \mathrm{E}[Y(0)] \\ &= \mathrm{E}[\pi(X)\, (Y(1) - Y(0))] \\ &= \mathrm{E}[\pi(X)\, \tau_0(X)]. \end{aligned} \tag{14.3.1}$$

Since we have already seen that the CATE $\tau_0(X)$ can be identified as $\mathrm{E}[Y(\eta_0) \mid X]$, we derive that the policy gains of any candidate policy can be identified as:

$$V(\pi) := \mathrm{E}\left[\pi(X)\,\mathrm{E}[Y(\eta_0) \mid X]\right] = \mathrm{E}[\pi(X)Y(\eta_0)].$$

For a given fixed π, this quantity is akin to the ATE considered in Section 10.3 weighted by a known function $\pi(X)$. In particular, when $\pi(X) = 1$ treats everyone, $V(\pi)$ is the ATE. The policy value is also akin to the GATE considered in the same section: if we scale it up by $1/\mathrm{E}[\pi(X)]$ then it is the GATE among those treated by π. Correspondingly, we can conduct inference on it by following the same recipe as in Section 10.4. This corresponds to the estimate

$$\widehat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)Y(\hat{\eta}_{[k(i)]}),$$

and the statement of Theorem 10.3.1 still applies to this weighted ATE, providing for inference on $V(\pi)$.

One measure of the heterogeneity of treatment effects is how much we can deviate from the average effect by carefully tailoring who gets assigned which treatment, that is, how large we can make $V(\pi)$:

$$V^* = \max_{\pi} V(\pi) = \mathrm{E}[\max_{p\in[0,1]} p\,\tau_0(X)] = V(\mathbb{1}\{\tau_0(X) \geq 0\}).$$

The last equality shows that $\pi^*(X) = \mathbb{1}\{\tau_0(X) \geq 0\}$ is an optimal policy (there may be multiple if $\mathrm{P}(\tau_0(X) = 0) > 0$). This suggests we can estimate V^* by following the same recipe as before and treating τ_0 as one more nuisance to estimate and plug into the policy we are evaluating. This, in fact, works well whenever π^* is uniquely optimal because then first order conditions for the optimization problem defining V^* will automatically ensure a zero derivative in τ_0, i.e., Neyman orthogonality as in Section 10.4. Namely, write $V^* = M(\tau_0, \eta_0) = \max_\tau M(\tau, \eta_0)$, where $M(\tau, \eta) = \mathrm{E}[\mathbb{1}\{\tau(X) \geq 0\}Y(\eta)]$. We already know that $\partial_\eta M(\tau_0, \eta_0) = 0$ from the case of evaluating any given policy, seen as a weighted ATE. For the derivative in τ we have

$$\begin{aligned}
&\frac{1}{t}|M(\tau_0 + t\xi, \eta_0) - M(\tau_0, \eta_0)| \\
&\quad = \frac{1}{t}\mathrm{E}[\tau_0(X)(\mathbb{1}\{-t\xi(X) \leq \tau_0(X) < 0 \ \vee \ 0 \leq \tau_0(X) < -t\xi(X)\})] \\
&\quad \leq \mathrm{E}[|\xi(X)|\mathbb{1}\{|\tau_0(X)| \leq t|\xi(X)|\}] \\
&\quad \leq \|\xi\|_2\sqrt{\mathrm{P}(|\tau_0(X)| \leq t|\xi(X)|)},
\end{aligned}$$

which has limit 0 as $t \to 0$ by the continuity of probability and $\mathrm{P}(\tau_0(X) = 0) = 0$ (since otherwise $\mathbb{1}\{\tau_0(X) > 0\}$ would not be almost surely the same as $\pi^*(X)$, contradicting its uniqueness). Thus, it suffices we learn τ_0 at non-parametric rates and plug it into a guess for π^*, which we then evaluate the same as any fixed policy.[1]

1: Just uniqueness of π^* may not be enough satisfy the additional regularity assumptions of Theorem 10.4.1 beyond Neyman orthogonality. We may need to assume not only that $\tau_0(X)$ has no atom at 0 but that it in fact has a bounded density in a neighborhood of 0. Additionally, the norm in which estimates of τ_0 converge is important and interacts with the allowable rate.

While V^* provides insight into how much we can get out of a treatment and covariates X if we are careful about personalizing, it need not give a full picture of the heterogeneity of treatment effects across X. For example, it may well be 0 if treatment effects are all negative, even if they are very heterogeneous. Another lens into heterogeneity may be the value of the optimal policy, when constrained to treat *exactly* q-fraction of the population:

$$\begin{aligned} V_q^* &= \max_{\pi:\mathrm{E}[\pi(X)]=q} \mathrm{E}[\pi(X)\tau_0(X)] \\ &= \max_{\pi} \min_{\lambda} \mathrm{E}[\pi(X)\tau_0(X) + \lambda(q - \pi(X))] \\ &= \min_{\lambda} \max_{\pi} \mathrm{E}[\pi(X)\tau_0(X) + \lambda(q - \pi(X))] \\ &= \min_{\lambda} q\lambda + \mathrm{E}[0 \vee (\tau_0(X) - \lambda)]. \end{aligned} \tag{14.3.2}$$

When, on the other hand, π^* is not at all unique, inference on V^* can be tricky. A solution to this challenging problem is given in [3].

Exchanging the order of max and min in the penultimate line of Eq. (14.3.2) is justified by a result known as Sion's minimax theorem.

We recognize the minimizer of the check loss in Eq. (14.3.2) as the quantile. So the latter minimization is attained at λ equal to the $(1-q)$-th quantile $\mu(\tau_0, q) = \inf\{t : \mathrm{P}(\tau_0(X) > t) \leq q\}$. Thus, any optimal constrained policy has $\pi_q^*(X) = 1$ when $\tau_0(X) > \mu(\tau_0, q)$ and $\pi_q^*(X) = 0$ when $\tau_0(X) < \mu(\tau_0, q)$. The quantity V_q^*/q is exactly the average treatment effect among the q-fraction of a subpopulation with largest values of $\tau_0(X)$, also known as the superquantile or the conditional value at risk. When $\mathrm{P}(\tau_0(X) = \mu(\tau_0, q)) > 0$ this subpopulation may not be unique and there can be different ways of splitting the group with $\tau_0(X) = \mu(\tau_0, q)$ to obtain a subpopulation of fraction exactly q (assuming either an infinite population or a finite population of infinitely divisible units). Notice the quantity V_q^* is still well-defined even in this non-unique case, and that as we vary q, we obtain a full characterization of the distribution of $\tau_0(X)$.

Now, suppose $\mathrm{P}(\tau_0(X) = \mu(\tau_0, q)) = 0$. Then, the constrained optimal policy is uniquely given by $\pi_q^*(X) = \mathbb{1}\{\tau_0(X) \geq \mu(\tau_0, q)\}$ and we have that $V_q^*/q = \mathrm{E}[Y(1) - Y(0) \mid \tau_0(X) \geq \mu(\tau_0, q)]$ is exactly the GATE among those with CATE above the $(1-q)$-th quantile. Moreover, for $q' > q$ with $\mathrm{P}(\tau_0(X) = \mu(\tau_0, q)) = \mathrm{P}(\tau_0(X) = \mu(\tau_0, q')) = 0$, we have that $(V_{q'}^* - V_q^*)/(q' - q)$ is the GATE among those with CATES between the $(1-q')$-th and $(1-q)$-th quantiles.

In the latter unique case, we may be tempted to follow the same recipe as before to also estimate V_q^*: plug in estimates of $\tau_0, \mu(\tau_0, q)$ to form a guess of π_q^* and evaluate it as we would any fixed policy. This worked for V^* because π^* maximized value, automatically inducing Neyman orthogonality. However, π_q^* does not globally maximize value, only subject to constraints. In other words, we do have $V^* = M(\tau_0 - \mu(\tau_0, q), \eta_0)$, but generally $\partial M(\tau_0 - \mu(\tau_0, q), \eta_0) \neq 0$.

To recover the same orthogonality-via-optimality as before we need to introduce a cost of violating the constraint. This is exactly what Eq. (14.3.2) does. Namely, writing $V_q^* = M_q(\tau_0, \mu(\tau_0, q), \eta_0) = \min_\lambda \max_\tau M_q(\tau, \lambda, \eta_0)$, where $M_q(\tau, \lambda, \eta) = q\lambda + \mathrm{E}[\mathbb{1}\{\tau(X) \geq \lambda\}(Y(\eta) - \lambda)]$, we will again find that $\partial_\tau M_q(\tau_0, \mu(\tau_0, q), \eta_0) = 0$ and $\partial_\lambda M_q(\tau_0, \mu(\tau_0, q), \eta_0) = 0$ because $\tau_0, \mu(\tau_0, q)$ by first order conditions, and $\partial_\eta M_q(\tau_0, \mu(\tau_0, q), \eta_0) = 0$ as always for weighted ATE estimation. This provides a recipe for estimating V_q^*. The minimax formulation also bestows a special property that if we get τ_0 wrong (or, estimate it too slowly), we will still get a lower bound on V_q^* as long as we estimate a corresponding best-response λ well.[2]

2: Complete details on how to do inference on V_q^* and guarantees thereon are given in [4].

14.4 Non-Parametric Inference for CATEs with Causal Forests

An inherently harder task is performing inference on the value $\tau_0(x)$ at a given point x. Statistically this problem can be even harder than performing inference on the value of a regression function at a particular point x. In fact, one solution casts the problem as such. Note that we already argued that:

$$\tau_0(x) = \mathrm{E}\left[Y(\eta_0) \mid X = x\right] \tag{14.4.1}$$

Thus one approach is to estimate the nuisance parameters η_0 in a cross-fitting manner and then use any flexible regression method that supports prediction intervals and apply it to the regression problem $Y(\hat{\eta}) \sim X$. In low dimensions, many classical approaches, such as kernel regression are applicable and can be invoked. In high-dimensions these methods will struggle to provide any meaningful insight.

An alternative is to use Random Forest based methods that will perform much better in practice. Standard Random Forest approaches typically equally balance bias and variance and hence do not allow for confidence interval construction. Recent work of [5, 6] proposes adaptations of Random Forests that,

in low-dimensions, provably produce asymptotically normal and un-biased predictions and provide theoretically justified construction of confidence intervals. The key ingredients in these adaptations are:

i *honesty*: a separate sample is used to construct the structure of the tree and a separate sample is used to calculate the estimates at the leaf nodes of the tree,

ii *balancedness*: every split should leave at least a $\rho \geq 0.2$ fraction of the samples on each side,

iii *random feature split*: every feature should have a probability of at least π/d to be chosen on each split, where d is the number of features (e.g. this can be achieved by choosing a random feature to split on with probability π),

iv *fully grown*: the tree should be grown fully, such that the number of samples that fall in every leaf should be at most some small constant,

ii *sub-sampling*: unlike typical random forest methods that use bootstrap sub-samples to build each tree (i.e. of the same size as the original samples and drawn with replacement), these adapted forests use sub-samples without replacement and of a smaller size $s \ll n$ than the original sample on each tree (the size of the sub-sample needs to be chosen carefully for the validity of the confidence intervals and should be of the order of $n^{\frac{\alpha d}{\alpha d+1}}$, where $\alpha = \frac{\log(1/\rho)}{\pi \log(1/(1-\rho))}$).

We will refer to any forest construction process that satisfies these properties as an *Honest Random Forest*.

We will refer to Honest Random Regression Forests that are trained on the doubly robust proxy labels $Y(\hat{\eta})$, with cross-fitted estimates of the nuisance functions, as *Doubly Robust Forests*. Based on the results in [6, 7], one can show that the validity of the confidence intervals of Honest Random Regression Forests is maintained even when the labels are biased due to the estimation error of the nuisance parameters $\hat{\eta}$, so long as:

$$\sqrt{n/s}\mathrm{E}[(H(\hat{\mu} - H(\mu_0))\,(\hat{g}(D, Z) - g_0(D, Z)) \mid X = x] \approx 0$$

Note that this requires accuracy of the nuisance estimates with respect to a confidional mean squared error. [7] shows how this can be achieved even in settings where Z is high-dimensional, albeit X remains low-dimensional. In particular, one should expect the size of the confidence interval or the error of the estimate to decay to zero at a rate of $\approx n^{-\frac{1}{2(ad+1)}}$, where d is the dimension of the covariates in X. This result is based on a conditional variant of the Neyman orthogonality property that

is satisfied by the doubly robust proxy labels, in conjunction with the asymptotic normality of predictions that stem from Honest Random Forests.

Finally, under stronger assumptions and for binary covariates, the recent work of [8] also shows that confidence intervals of Honest Regression Random Forests trained with the squared loss criterion and *without the balancedness or random feature split* property, are asympotitcally valid even in high-dimensions, as long as the true regression function $\mathrm{E}[Y \mid X = x]$ is sparse (i.e. only a small constant number of variables are relevant for predicting Y). However, an upper bound on the degree of sparsity (i.e. the number of relevant features) needs to be known. Extending this inference result to high-dimensional continuous features remains and active area of investigation. Similarly, extending this result to simultaneous confidence bands and not just pointwise confidence intervals is another active area of investigation.

While adaptive estimation of the CATE can be obtained fairly generally, it is important to note that X should be low dimensional if we want to obtain confidence intervals or perform hypothesis tests. Genovese and Wasserman (Annals of Stats, 2008) [9] show that there do not exist adaptive confidence bands for estimation of the curve $\tau_0(X)$ except under very restrictive assumptions more generally. They suggest instead to construct adaptive bands that cover a surrogate function π which is close to, but simpler than, τ_0.

In the previous section, where we discuss the use of OLS with low-dimensional X, the surrogate π represents either GATEs or the best linear approximation of the CATE. Inferential guarantees are also available for the case where X is low-dimensional and Random Forests are used. Inferential results for low-dimensional surrogates π based on other methods should also be possible, though we note that GATEs and best linear predictors more generally are readily interpretable and will likely be useful in many settings.

Despite these theoretical limitations, forest based approaches are empirically powerful as they tend to identify the most relevant factors that drive treatment effect heterogeneity, while at the same time providing some signal of uncertainty of the prediction. Even though this uncertainty quantification is more brittle than for instance the confidence intervals of an OLS regression, as it depends on many more assumptions and holds only under particular choices of the hyperparameters of the method (which are typically violated in practice; especially

when data-driven hyperparameter tuning is invoked via cross-validation, which is the typical case), honest random forest based approaches still provide a meaningful signal of how uncertain the model is about its CATE predictions, at different regions of the covariate space.

Generalized Random Forests (GRF) An alternative approach is to formulate the CATE problem as solving a local or conditional version of a moment restriction [6, 7]:

$$\mathrm{E}[m(W; \tau_0(x), \eta_0) \mid X = x] = 0 \tag{14.4.2}$$

where $m(W; \tau, \eta)$ is a vector of moment restrictions of the same dimension as τ.

Such Generalized Random Forests are trained so as to maximize the induced heterogeneity in $\hat{\tau}(x)$ with every split. For every node P in some tree of the forest, let $\hat{\tau}_P$ denote our estimate of $\mathrm{E}[m(W; \tau_P, \eta_0) \mid X \in P] = 0$. Such an estimate can be constructed by solving the moment restriction with respect to τ_P using only the samples that fall in node P and using an estimate $\hat{\eta}$ of η_0 based on auxiliary data or in a cross-fitting manner. Let C_1, C_2 denote the child nodes that will be created from some candidate split, with sample sizes n_1 and n_2 correspondingly. Then a proxy criterion that targets maximizing heterogeneity is maximizing $n_1\, \tau_{C_1}^2 + n_2\, \tau_{C_2}^2$. This is one of the criteria typically used in Generalized Random Forests. Moreover, to avoid the computational burden of resolving the moment equation for every candidate split, typically an approximation of the quantities τ_{C_1} and τ_{C_2} is used. In particular, a local linear approximation around the estimate of the parent node is being used and locally updated, i.e. $\tau_{C_1} \approx \tau_P - \frac{1}{n_1} \sum_{i \in C_1} J_P m(W_i; \tau_P, \hat{\eta})$, with $J_P = \frac{1}{n_P} \sum_{i \in P} \partial_\tau m(W; \tau_P, \hat{\eta})$, with n_P being the number of samples in the parent node.

Moreover, the final estimate $\hat{\tau}(x)$ is derived in a manner slightly different than regression forests (albeit it coincides for the case of a regression moment, i.e. $m(W; \tau(x)) = y - \tau(x)$). For more general moments, for every target point x for which we want to predict the CATE, the Random Forest structure is used to construct weights for every other sample $i \in \{1, \ldots, n\}$, that capture the degree of "similarity" of x to X_i. These weights roughly correspond to the fraction of trees in the forest, for which X_i falls in the same leaf node as x, downweighting leafs of larger size. Thus if we have trained a forest with B trees and we let $L_b(x)$ denote the leaf node that a sample with covariates x falls in at tree b and let $|L_b(x)|$ the number of samples in that

leaf, then we have:

$$K(x, X_i) = \frac{1}{B} \sum_{b=1}^{B} \frac{1\{L_b(x) = L_b(X_i)\}}{|L_b(x)|}$$

Then to calculate $\hat{\tau}(x)$, we solve with respect to $\hat{\tau}(x)$, a weighted empirical average version of the moment condition:

$$\sum_{i=1}^{n} K(x, X_i)\, m(W_i; \hat{\tau}(x), \hat{\eta}) = 0 \tag{14.4.3}$$

using the weights that are derived based on the similarity metric induced by the forest structure.

When the forest construction process satisfies the criteria we defined earlier in the section of *honesty*, *balancedness*, *random feature splitting*, *fully grown trees* and *sub-sampling without replacement*, then under similar conditions as in the case of Regression Forests, the prediction of a Generalized Random Forest (and its extension, the Orthogonal Random Forest) can be shown to be asymptotically normal and an asymptotically valid confidence interval construction can be employed. Albeit the same limitations as we described in the regression case, carry over to the confidence intervals produced by these methods.

Causal Forests: a GRF for CATE We describe here an empirically popular variant of causal forests that uses the Generalized Random Forest formulation. Albeit, unlike the Doubly Robust Forest approach, this approach is valid only if $X = Z$ or if we make the stronger further assumption that the high-dimensional CATE function $\delta_0(Z) = \mathrm{E}[Y(1) - Y(0) \mid Z]$, is only a function of the variables X, i.e. $\delta_0(Z) = \tau_0(X)$ and τ_0.

In this case, for a binary treatment, we can write without loss of generality

$$\mathrm{E}[Y \mid D, Z] = \pi_0(Z)\, D + g_0(Z)$$

where $\pi_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z]$ is the conditional average predictive effect (CAPE). Moreover, by conditional exogeneity, the CAPE function π_0 is equal to the high-dimensional CATE function δ_0. Thus, for a binary treatment we can always write the regression equation:

$$Y = \delta_0(Z)\, D + g_0(Z) + \epsilon, \qquad \mathrm{E}[\epsilon \mid D, Z] = 0$$

From this, we can derive that $\mathrm{E}[Y \mid Z] = \delta_0(Z)\mathrm{E}[D \mid Z] + g_0(Z)$.

Subsequently, we can write:

$$Y - \mathrm{E}[Y \mid Z] = \delta_0(X)\,(D - \mathrm{E}[D \mid Z]) + \epsilon$$

Letting $\hat{Y} = Y - \mathrm{E}[Y \mid Z]$ and $\hat{D} = D - \mathrm{E}[D \mid Z]$ and since $Z, \hat{D}$ is uniquely determined by D, Z, we can conclude that the following regression equation holds:

$$\hat{Y} = \delta_0(Z)\,\hat{D} + \epsilon, \qquad \mathrm{E}[\epsilon \mid \hat{D}, X] = 0$$

If we now further assume that $\delta_0(Z) = \tau_0(X)$, and since $X, \hat{D}$ is a subset of $Z, \hat{D}$, we can write the regression equation:

$$\hat{Y} = \tau_0(X)\,\hat{D} + \epsilon, \qquad \mathrm{E}[\epsilon \mid \hat{D}, X] = 0 \tag{14.4.4}$$

From this regression equation we can derive the moment constraint:

$$\mathrm{E}\left[(\hat{Y} - \tau_0(x)\,\hat{D})\,\hat{D} \mid X = x\right] = \mathrm{E}\left[\epsilon\,\hat{D} \mid X = x\right] = 0$$

Note that this moment equation is a conditional analogue of the Normal Equation that we used in the PLR model, where we used the equation

$$\mathrm{E}(\hat{Y} - \theta_0\,\hat{D})\,\hat{D} = 0$$

to estimate the constant treatment effect under a partially linear model $\mathrm{E}(Y \mid D, X) = \theta_0 D + g(Z)$. Now that the coefficient associated with D is allowed to vary with X, we can estimate the heterogeneous coefficient by solving the same moment but conditional on X, i.e.

$$\mathrm{E}\left[(\hat{Y} - \tau_0(x)\,\hat{D})\,\hat{D} \mid X = x\right] = 0 \tag{14.4.5}$$

Note that the above method falls in the general framework that can be handled by Generalized Random Forests and their extension, the Orthogonal Random Forests. We can estimate $\hat{\tau}(x)$ by estimating the nuisance function $\eta_0 = (p_0, q_0)$, where $p_0(Z) = \mathrm{E}(D \mid Z)$ and $q_0(Z) = \mathrm{E}(Y \mid Z)$ in a cross-fitting manner, letting $\check{Y} = Y - \hat{p}(Z)$, $\check{D} = D - \hat{q}(Z)$ and then applying the Generalized Random Forest method with moment equation:

$$m(W; \tau(x), \hat{\eta}) = (\check{Y} - \tau(x)\,\check{D})\,\check{D}$$

The formal analysis of the validity of the confidence intervals of this approach can be found in [6] for the case when $X = Z$ and is low-dimensional, in which case, one does not need to account for the errors in $\hat{\eta}$, as long as a constant offset is also added to

the moment equation, solving the vector of moments:

$$m(W; \tau(x), \beta(x), \hat{\eta}) = (\check{Y} - \tau(x)\,\check{D} - \beta(x)) \begin{pmatrix} \check{D} \\ 1 \end{pmatrix}$$

A formal analysis of the case when $X \subseteq Z$ and Z can be high-dimensional (or when the constant term is not added to the moment equation) can be found in [7], which also accounts for the impact of the nuisance estimation errors.

Example 14.4.1 (Forests in the 401k Example) We revisit the 401(k) example that we used in the previous section and apply the forest based methods for CATE estimation. In this case, we used all the variables for heterogeneity (i.e. $X = Z$) and let the forest methods identify the relevant dimensions of heterogeneity in a more data-driven manner. We applied both the Doubly Robut Forest and the Causal Forest approach. For the nuisance estimates, in all cases, we used gradient boosted forests and estimated the nuisances in a cross-fitting manner with 5-fold cross-fitting.

In Figure 14.3 we depict the predictions and confidence intervals of the Doubly Robust Forest method, where the x-axis corresponds to income (while other co-variates are fixed to their overall median values). In Figure 14.4 we depict the analogous plot for the Causal Forest method. We find that both methods identify a similar CATE and that this CATE is inline with the intuitive property that the effect of 401(k) eligibility on net financial assets is larger for larger incomes. Moreover, unlike the BLP estimates, we find that the forest based estimates, behave more reasonably at the extreme ends of the income distribution as they do not extrapolate linearly and identify, in a data-driven manner, a more sigmoid effect curve, between $\approx \$5k$ and $\approx \$22k$. The results are almost identical for the two methods. Moreover, the confidence intervals are informative that the CATE prediction is quite uncertain at the upper extreme part of the income distribution were samples are much more spread out and there is a long tail. Finally, when looking at measures of feature importance for Random Forests, income was identified as the most important feature.

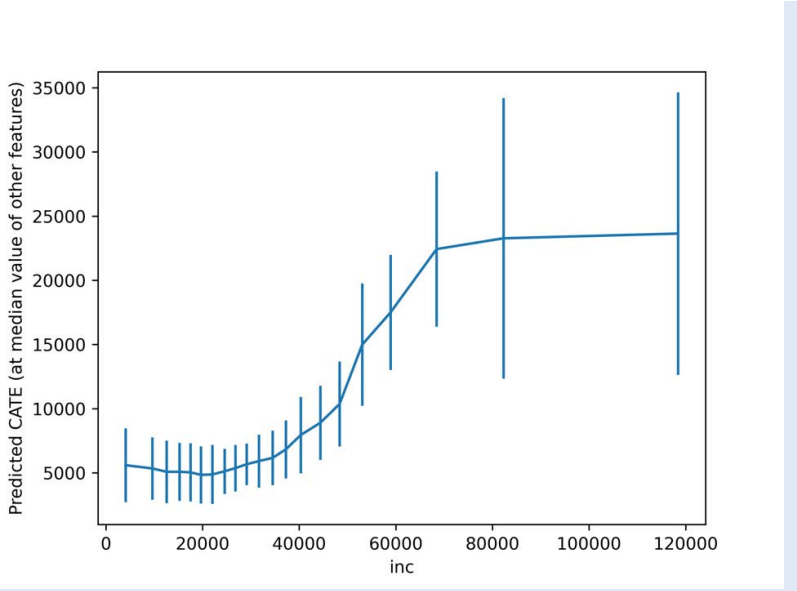

Figure 14.3: Doubly Robust Forest in the 401k example.

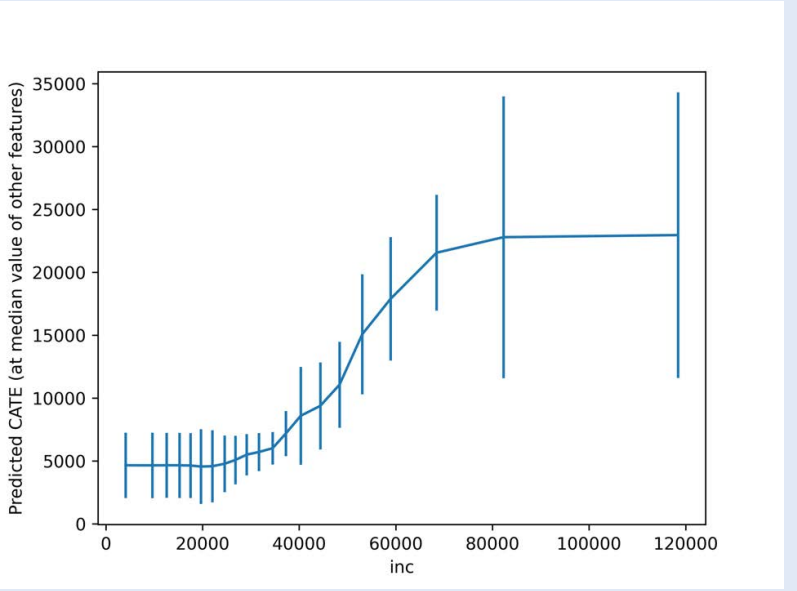

Figure 14.4: Causal Forest in the 401k example.

Empirical Example: The "Welfare" Experiment

We illustrate the estimation of CATEs with forests, with an empirical application on studying the effects of the word "welfare" on the support of people for government programs. Starting in the 1980s, the General Social Survey (GSS) started including a question around satisfaction with public spending. What is more important, the GSS conducted a randomized controlled trial where the respondents where assigned one of two variations of the same question at random. Both variations had the same meaning and introduction, albeit one was asking about satisfaction of the respondent with respect to government spending for "welfare programs", while the other variation was

phrased as government spending for "assistance to the poor". This small variation has been found to have substantial average effect on the response and several studies have attempted to parse out treatment effect heterogeneity.

In this section we applied Causal Forests and Doubly Robust Forests on a dataset collected from such GSS surveys from 1986 to 2010, as described in [10]. The dataset consists of 12907 samples containing i) the variant of the question that was assigned to the participant (with $D = 1$ corresponding to "welfare programs" and $D = 0$ corresponding to "assistance to the poor), ii) their numerical level of satisfaction response to the question (Y) and 42 features (X) that contain many characteristics of the respondent related to gender, income, education, family size and marital status, race, political views and occupation sector. The average treatment effect based on a simple two-means estimate is -0.3681 as reported in Figure 14.5.

	coef	**std err**	**P> |z|**	**[0.025**	**0.975]**
const	0.4798	0.006	0.000	0.467	0.492
D	-0.3681	0.007	0.000	-0.383	-0.354

Figure 14.5: Average treatment effect in welfare experiment.

We constructed a Causal Forest and a Doubly Robust Forest using all the 42 variables for treatment effect heterogeneity and as controls. We used gradient boosting regression with cross-fitting to calculate the nuisance functions required for each of the forests. The hyperparameters of the nuisance estimators were selected based on cross validation. Subsequently, we looked at the most important feature in the forest, as measured by a feature importance criterion that roughly corresponds to the average reduction in the splitting criterion, every time that the feature was used for splitting. The most important feature came out to be political views, both in the Causal Forest and in the Doubly Robust Forest. Subsequently, in Figure 14.6 and Figure 14.7 we report the heterogeneous effect for each value of the polviews covariate and imputing all other covariates at their median value. The point estimate and the corresponding 5%-95% confidence intervals that are provided by the forest methods are depicted.

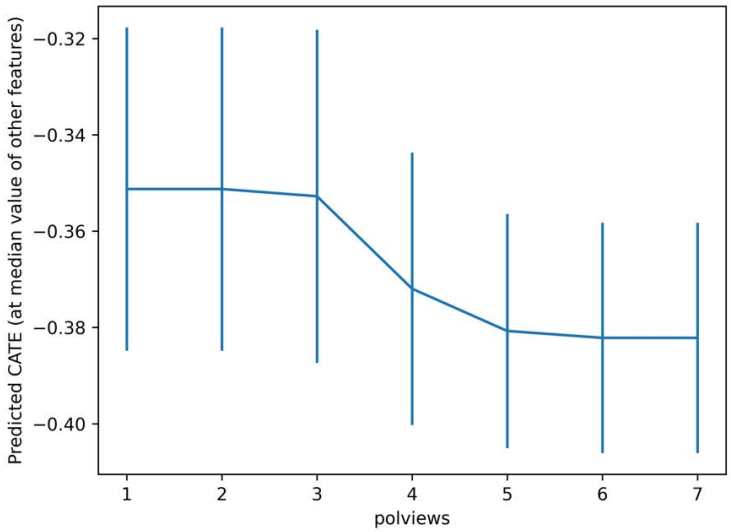

Figure 14.6: R-Learner based causal forest in the welfare example.

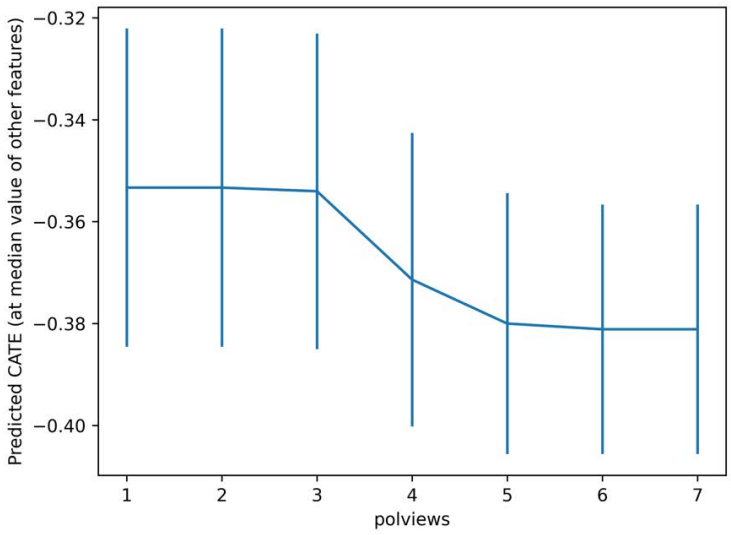

Figure 14.7: Doubly Robust honest regression forest in the welfare example.

Notebooks

- R Notebook for DML on CATE analyzes ATE of 401(K) conditional on income.
- Python Notebook for CATE Inference analyzes CATE of welfare experiment and for 401k experiment with Best Linear Predictors of CATE and with Random Forest and Causal Forest based methods.

Bibliography

[1] Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House Group, 2007 (cited on page 363).

[2] Vira Semenova and Victor Chernozhukov. 'Debiased machine learning of conditional average treatment effects and other causal functions'. In: *The Econometrics Journal* 24.2 (2021), pp. 264–289 (cited on page 370).

[3] Alexander R Luedtke and Mark J Van Der Laan. 'Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy'. In: *Annals of statistics* 44.2 (2016), p. 713 (cited on page 374).

[4] Nathan Kallus. 'Treatment effect risk: Bounds and inference'. In: *Management Science* 69.8 (2023), pp. 4579–4590 (cited on page 375).

[5] Stefan Wager and Susan Athey. 'Estimation and Inference of Heterogeneous Treatment Effects using Random Forests'. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242 (cited on page 375).

[6] SUSAN ATHEY, JULIE TIBSHIRANI, and STEFAN WAGER. 'GENERALIZED RANDOM FORESTS'. In: *The Annals of Statistics* 47.2 (2019), pp. 1148–1178 (cited on pages 375, 376, 378, 380).

[7] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. 'Orthogonal random forest for causal inference'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4932–4941 (cited on pages 376, 378, 381).

[8] Vasilis Syrgkanis and Manolis Zampetakis. 'Estimation and Inference with Trees and Forests in High Dimensions'. In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3453–3454 (cited on page 377).

[9] Christopher Genovese and Larry Wasserman. 'Adaptive confidence bands'. In: *Annals of Statistics* 36.2 (2008), pp. 875–905 (cited on page 377).

[10] Donald P Green and Holger L Kern. 'Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees'. In: *Public opinion quarterly* 76.3 (2012), pp. 491–511 (cited on page 383).