NBER WORKING PAPER SERIES

GENERIC MACHINE LEARNING INFERENCE ON HETEROGENEOUS TREATMENT EFFECTS IN RANDOMIZED EXPERIMENTS, WITH AN APPLICATION TO IMMUNIZATION IN INDIA

Victor Chernozhukov
Mert Demirer
Esther Duflo
Iván Fernández-Val

Working Paper 24678
http://www.nber.org/papers/w24678

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2018, Revised February 2023

This paper was delivered (virtually) by Esther Duflo at the Fischer-Shultz Lecture of the Econometric Society World Congress, 2020. We thank the editor Guido Imbens, three anonymous referees, Susan Athey, Moshe Buchinsky, Denis Chetverikov, Carlos Cineli, Matt Hong, Stella Hong, Steven Lehrer, Siyi Luo, Max Kasy, Sylvia Klosin, Susan Murphy, Whitney Newey, Patrick Power, Victor Quintas-Martinez, Suhas Vijaykumar, and seminar participants at ASSA 2018, Barcelona GSE Summer Forum 2019, Brazilian Econometric Society Meeting 2019, BU, Lancaster, NBER summer institute 2018, NYU, UCLA, Whitney Newey's Contributions to Econometrics conference, and York for valuable comments. Anirudh Sankar provided us with excellent research assistance. We gratefully acknowledge research support from the National Science Foundation, AFD, USAID, and 3ie. An R package that implements the methods in this paper, GenericML, is available on GitHub at https://github.com/mwelz/GenericML. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India
Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val
NBER Working Paper No. 24678
June 2018, Revised February 2023
JEL No. C18,C21,D14,G21,O16

ABSTRACT

We propose strategies to estimate and make inference on key features of heterogeneous effects in randomized experiments. These key features include best linear predictors of the effects using machine learning proxies, average effects sorted by impact groups, and average characteristics of most and least impacted units. The approach is valid in high dimensional settings, where the effects are proxied (but not necessarily consistently estimated) by predictive and causal machine learning methods. We post-process these proxies into estimates of the key features. Our approach is generic, it can be used in conjunction with penalized methods, neural networks, random forests, boosted trees, and ensemble methods, both predictive and causal. Estimation and inference are based on repeated data splitting to avoid overfitting and achieve validity. We use quantile aggregation of the results across many potential splits, in particular taking medians of p-values and medians and other quantiles of confidence intervals. We show that quantile aggregation lowers estimation risks over a single split procedure, and establish its principal inferential properties. Finally, our analysis reveals ways to build provably better machine learning proxies through causal learning: we can use the objective functions that we develop to construct the best linear predictors of the effects, to obtain better machine learning proxies in the initial step. We illustrate the use of both inferential tools and causal learners with a randomized field experiment that evaluates a combination of nudges to stimulate demand for immunization in India.

Victor Chernozhukov
Department of Economics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Mass. 02139
vchern@mit.edu

Mert Demirer
MIT Sloan School of Management
100 Main St
Cambridge, MA 02142
mdemirer@mit.edu

Esther Duflo
Department of Economics, E52-544
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
eduflo@mit.edu

Iván Fernández-Val
Department of Economics
Boston University
270 Bay State Rd
Boston, MA 02215
ivanf@bu.edu

1. Introduction

Randomized Controlled Trials (RCT) and Machine Learning (ML) are arguably two of the most important developments in data analysis methods for applied researchers. RCTs play an important role in the evaluation of social and economic programs, medical treatments and marketing (e.g., Duflo et al., 2007; Imbens and Rubin, 2015). ML is a name attached to a variety of constantly evolving statistical learning methods including Random Forest, Boosted Trees, Neural Networks, Penalized Regression, Ensembles, and Hybrids; see, e.g., Wasserman (2016) for a recent review, and Friedman et al. (2001), Bishop and Nasrabadi (2006), Murphy (2012), Hastie et al. (2015), Goodfellow et al. (2016) and James et al. (2021) for prominent textbook treatments. ML has become a key tool for prediction and pattern recognition problems, surpassing classical methods in high dimensional settings.

At first blush, those two sets of methods may seem to have very different applications: in the most basic randomized controlled experiment, there is a sample with a single treatment and a single outcome. Covariates are not necessary and even linear regression is not the best way to analyze the data (Freedman, 2008; Imbens and Rubin, 2015). In practice however, applied researchers are often confronted with more complex experiments. For example, there might be accidental imbalances in the sample, which require selecting control variables in a principled way. ML tools, such as the lasso method proposed in Belloni et al. (2014, 2017) or the double machine learning method proposed in Chernozhukov et al. (2017), have proven useful for this purpose. Moreover, some complex RCT designs have so many treatment combinations that ML methods may be useful to select the few treatments that actually work and pool the rest with the control groups for statistical power (Banerjee et al., 2019). Finally, researchers and policy makers are often interested in features of the impact of the treatment that go beyond the simple average treatment effect. In particular, very often, they want to know whether the treatment effect depends on covariates, such as gender, age, etc. This heterogeneity is essential to assess if the impact of the program would generalize to a population with different characteristics, and, for economists, to better understand the driving mechanism behind the effects of a particular program. In a review of 189 RCTs published in top economic journals since 2006, we found that 76 (40%) report at least one subgroup analysis, wherein they report treatment effects in subgroups formed by baseline covariates.[1]

One issue with reporting treatment effects split by subgroups, however, is that there might be a large number of potential ways to form subgroups. Often researchers collect rich baseline surveys, which give them access to a large number of covariates: choosing subgroups ex-post opens the possibility of overfitting. To solve this problem, medical journals and the FDA require pre-registering

[1] The papers were published in *Quarterly Journal of of Economics*, *American Economic Review*, *Review of Economics Studies*, *Econometrica* and *Journal of Political Economy*. We thank Karthik Mularidharan, Mauricio Romero and Kaspar Wüthrich for sharing the list of papers they computed for another project.

the sub-sample of interest in medical trials *in advance*. In economics, this approach has gained some traction with the adoption of pre-analysis plans, which can be filed in the AEA registry for randomized experiments. However, restricting the heterogeneity analysis to pre-registered subgroups amounts to throwing away a large amount of potentially valuable information, especially now that many researchers collect large baseline data sets. It should be possible to use the data to discover *ex post* whether there is any relevant heterogeneity in treatment effect by covariates.

To do this in a disciplined fashion and avoid the risk of overfitting, scholars have recently proposed using ML tools. Indeed, ML tools seem ideal for exploring heterogeneity of treatment effects when researchers have access to a potentially large array of baseline variables to form subgroups and few guiding principles on which of those are likely to be relevant. Several recent papers, which we review below, develop methods for detecting heterogeneity in treatment effects. Empirical researchers have taken notice.[2]

This paper develops a generic approach to using any of the available ML tools to predict and make inference on heterogeneous treatment or policy effects. A core difficulty of applying ML tools to the estimation of heterogenous causal effects is that, while they are successful in prediction empirically, it is much more difficult to obtain uniformly valid inference, i.e., inference that remains valid under a large class of data generating processes. In fact, in high dimensional settings, absent strong assumptions, generic ML tools may not even produce consistent estimators of the *conditional average treatment effect* (CATE), the difference in the expected potential outcomes between treated and control states conditional on covariates. Previous attempts to solve this problem focused either on specific tools (for example, the method proposed by Athey and Imbens (2016), which has become popular with applied researchers, and uses trees), or on situations where those assumptions might be satisfied. Our approach to resolving the fundamental impossibilities in non-parametric inference is different. Motivated by Genovese and Wasserman (2008), instead of attempting to get consistent estimation and uniformly valid inference on the CATE itself, we focus on providing valid estimation and inference on *features* of CATE.

We start by building a ML proxy predictor of CATE, and then target features of the CATE based on this proxy predictor. In particular, we consider three objects, which are likely to be of interest to applied researchers and policy makers: (1) *Best Linear Predictor* (BLP) of the CATE on the ML proxy predictor; (2) *Sorted Group Average Treatment Effects* (GATES) or average treatment

[2] In the recent past, several new empirical papers in economics used ML methods to estimate heterogeneous effects. E.g. Hussam et al. (2022) showed that villagers outperform the machine learning tools when they predict heterogeneity in returns to capital. Davis and Heller (2020) predicted who benefits the most from summer internship projects. Deryugina et al. (2019) used the methods developed in the present paper to evaluate the heterogeneity in the effect of air pollution on mortality. Crepon et al. (2021) also built on the present paper to develop a methodology to determine if the impact of two different programs can be accounted for by different selection. The methodological papers reviewed later also contain a number of empirical applications.

effect by heterogeneity groups induced by the ML proxy predictor; and (3) *Classification Analysis* (CLAN) or the average characteristics of the most and least affected units defined in terms of the ML proxy predictor. Thus, we can find out if there is detectable heterogeneity in the treatment effect based on observables, and if there is any, what the treatment effect is for different bins. And finally we can describe which of the covariates are associated with this heterogeneity.

There is a trade-off between more restrictive assumptions or tools and a more ambitious estimation. We address this trade-off by focusing on coarser objects of the function rather than the function itself, but make as little assumptions as possible. This seems to be a worthwhile sacrifice: the objects for which we have developed inference appear to us at this point to be the most relevant, but in the future, one could easily use the same approach to develop methods to estimate other objects of interest. For example, Crepon et al. (2021) used the same technique to construct and estimate a specific form of heterogeneity at the post-processing stage. Even then, as we will see, getting robust and conservative standard errors for heterogeneity requires a larger sample size than just estimating average treatment effects. This reflects a different trade-off: if we do not assume that we can predict *ex ante* where the heterogeneity might be (in which case we can write it down in a pre-analysis plan), power will be lower, and detecting heterogeneity will require a larger sample. This is a consideration that applied researchers will need to keep in mind when designing and powering their experiments.

Our estimation and inference methods rely on sample splitting to avoid overfitting and other inferential non-regularities (from using the model selection as an integral part of machine learning). Conditional on a single data split into a training and a hold-out sample, statistical inference is conceptually straightforward and appealing. Indeed, in this case, statistical inference reduces to the classical inference for linear regression and sample means. Theoretically, if a researcher can credibly pre-commit to a single data split, this gives one clean solution to the inferential problem. However, researchers often consider multiple data splits to demonstrate that the empirical results are robust and not driven by a favorable or unfavorable particular split.[3] To support this approach, we propose quantile-aggregated inference – which aggregates inferential results by taking medians of estimates and medians and other quantiles of upper and lower confidence intervals obtained from different splits. We show that quantile aggregation formally lowers estimation (reporting) risks over a single-split procedure, and we establish its inferential properties.

The proposed approach is generic in that it can be applied in conjunction with any ML method. To compare and select among ML methods, we develop goodness-of-fit measures for the BLP and GATES. We also take one step backward and use these goodness-of -fit measures to build ML proxies that better target the CATE through causal learning. We show that these causal machines

[3] Moreover, using multiple splits and aggregating the results reduces the probability that two researchers working with the same data will arrive at different conclusions.

produce provably better proxies of the CATE than generic (predictive) ML methods. Moreover, by designing the ML to target CATE directly, the post-processing methods that we develop can focus more on providing valid inference and less on correcting biases.

We apply our method to a large-scale RCT of nudges to encourage immunization in the state of Haryana, Northern India. This experiment, an important practical application in its own right, is designed and discussed in Banerjee et al. (2019). Immunization is generally recognized as one of the most effective and cost-effective ways to prevent illness, disability, and diseases. Yet, worldwide, close to 20 million children every year do not receive critical immunizations (Unicef, 2019). While early policy efforts have focused mainly on improving the infrastructure for immunization services, a more recent literature suggests that "nudges" (such as small incentives, leveraging the social network, SMS reminders, social signalling, etc.) may have large effects on the use of those services.[4] This project was a collaboration with the government of Haryana, which was willing to experiment with a combination of nudges, with the goal of choosing the most effective policy and implement it at scale. It built a custom vaccination platform, and ran a large-scale experiment covering seven districts, 140 Primary health centers, 2,360 villages involved in the experiment (including 915 at risk for all the treatments), and 295,038 children in the resulting database. Immunization was very low at baseline: in every single village of the district, the fraction of children whose parents had reported they received the measles vaccine (the last in the sequence) was 39%, and only 19.4% had received the vaccine before the age of 15 months, whereas the full sequence is supposed to be completed in one year. The experiment was a cross randomized design of three main nudges: providing incentives, sending SMS reminders, and seeding ambassadors. It included several variants for each policy: the level and schedule of the incentives, the number of people receiving reminders, and the mode of selection of the ambassadors, leading to a large number (75) of finely differentiated bundles.

Banerjee et al. (2019) developed a methodology to identify the most effective and cost-effective bundle of policies, based on an application of LASSO to a marginal effects specification that imposes some structure on the bundles, and in particular, the idea that policy variants (e.g. level of incentives, or level of coverage of SMS reminders) may be indistinguishable in practice. They found that the most cost-effective policy is to combine "information hubs" (people identified by others as good at diffusing information) and SMS reminders. This is cheap and can be done everywhere. In fact, they showed that this policy is the only one among those tested that would actually save money to the government for each measles shot, while increasing immunization. But the most *effective* policy, i.e., the policy that increases immunization the most, is the combination of incentives, immunization ambassadors, and SMS reminders, which is much more expensive.

[4] See, for example, Banerjee et al. (2010); Bassani et al. (2013); Wakadha et al. (2013); Johri et al. (2015); Oyo-Ita et al. (2016); Gibson et al. (2017); Karing (2018); Domek et al. (2016); Uddin et al. (2016); Regan et al. (2017); Alatas et al. (2019); Banerjee et al. (2021).

Yet, while this policy increases the cost per immunization, the effects are important: the number of monthly measles shots (the last vaccine in the schedule, and thus a marker for full immunization) delivered increases by 3.26, corresponding to 44% of the mean vaccination rate in the control group that got neither SMS nor increasing incentives and information hubs. The government was therefore interested in finding out where the program would be most effective, to implement it only in those places even at the higher cost per immunization.

The pre-analysis plan specified to look for heterogeneity by gender and by "Village-level baseline/national census variables, including assets, beliefs, knowledge, and attitudes towards immunization" but did not identify one or two specific baseline variables to look at. This reflected genuine uncertainty (as is often the case). Many factors can influence policy impact, from attitudes to implementation capabilities to baseline levels, and we did not have a specific theory of where to look. It is precisely the type of context that requires a principled approach to avoid overfitting, and provide a policy-relevant recommendation.[5]

The rest of the paper is organized as follows. Section 2 formalizes the framework, describes our approach and compares it with the existing literature. Section 3 presents identification and estimation strategies for the key features of CATE of interest. Section 4 introduces our inference method that accounts for uncertainty coming from parameter estimation and sample splitting. Section 5 presents the construction of causal machines that can learn CATE better than purely predictive approaches or some existing proposals for causal approaches. Section 6 reports the results of the empirical application and provides detailed implementation algorithms. Section 7 concludes with some remarks. The Appendix gathers proofs of the main theoretical results and additional technical results.

2. Our Agnostic Approach

2.1. **Model and Key Causal Functions.** Let $Y(1)$ and $Y(0)$ be the potential outcomes in the treatment state 1 and the non-treatment state 0 (see Neyman, 1923; Rubin, 1974). Let Z be a possibly high-dimensional vector of covariates that characterize the observational units. The main causal functions are the baseline conditional average (BCA):

$$b_0(Z) := \mathrm{E}[Y(0) \mid Z], \tag{2.1}$$

and the conditional average treatment effect (CATE):

$$s_0(Z) := \mathrm{E}[Y(1) - Y(0) \mid Z] = \mathrm{E}[Y(1) \mid Z] - \mathrm{E}[Y(0) \mid Z]. \tag{2.2}$$

[5] This approach of finding the best treatment and then looking at where it works the best gets closer to the idea of "personalized medicine". Using the same data, Agarwal et al. (2020) go one step further and use a "synthetic intervention" approach to look for the policy that works the best for each kind of village.

Suppose the binary treatment variable D is randomly assigned conditional on Z, with probability of assignment depending on a subvector of stratifying variables $Z_1 \subseteq Z$, namely

$$D \perp\!\!\!\perp (Y(1), Y(0)) \mid Z, \tag{2.3}$$

and the propensity score is known and is given by

$$p(Z) := \mathrm{P}[D = 1 \mid Z] = \mathrm{P}[D = 1 \mid Z_1], \tag{2.4}$$

which we assume is bounded away from zero or one:

$$p(Z) \in [p_0, p_1] \subset (0,1) \;\; a.s. \tag{2.5}$$

This setup is similar to Rosenbaum and Rubin (1983).

The observed outcome is $Y = DY(1) + (1-D)Y(0)$. Under the stated assumption, the causal functions are identified by the components of the regression function of Y given D, Z:

$$Y = b_0(Z) + Ds_0(Z) + U, \;\; \mathrm{E}[U \mid Z, D] = 0, \tag{2.6}$$

that is, $b_0(Z) = \mathrm{E}[Y \mid D = 0, Z]$, and

$$s_0(Z) = \mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z]. \tag{2.7}$$

This regression underlies the use of predictive ML methods that learn $\mathrm{E}[Y \mid D, Z]$ and then estimate CATE using the formula.

Alternatively one can identify CATE using the following two equivalent "causal" regressions:

$$s_0(Z) = \mathrm{E}[HY \mid Z] = \mathrm{Cov}[Y, H|Z]; \tag{2.8}$$

where H is the residualized treatment scaled by its variance:

$$H = H(D, Z) := \frac{D - p(Z)}{p(Z)(1 - p(Z))}, \tag{2.9}$$

also known as the Horvitz-Thompson transform. We mention these alternative strategies here, because as shown in Section 5, they can lead to better ways of approximating $s_0(Z)$ than through the predictive regression (2.6), and our inference tools equally apply to ML methods that try to learn $s_0(Z)$ through either of these relations. In fact, in our empirical analysis, the strategies based on (2.8) measurably outperform the strategies based on (2.7).

2.2. **Estimation and Inference Challenges.** Regardless of the way we try to learn $s_0(Z)$, estimation and inference are challenging in modern high-dimensional settings, because the target function $z \mapsto s_0(z)$ can live in a very complex class. ML methods effectively explore various forms of sparsity to yield "good" approximations to $s_0(z)$. In its simplest form, sparsity reduces the complexity of $z \mapsto s_0(z)$ by assuming that it can be well-approximated by a function that only depends on a low-dimensional subset of z, making consistent estimation possible. As a result, these methods can perform much better than classical methods in high-dimensional settings under sparsity. However, sparsity or, more generally, low complexity of the CATE function s_0, are untestable assumptions that must be used with caution.

Without some form of sparsity, it is hard, if not impossible, to obtain consistent estimators of $z \mapsto s_0(z)$. There are several fundamental reasons as well as large gaps between theory and practice that are responsible for this. One fundamental reason is that ML methods might not even produce consistent estimators of $z \mapsto s_0(z)$ in high dimensional settings. For example, if z has dimension d and the target function $z \mapsto s_0(z)$ is assumed to have p continuous and bounded derivatives, then the worst case (minimax) lower bound on the rate of learning this function from a random sample of size N cannot be better than $N^{-p/(2p+d)}$ as $N \to \infty$, as shown by Stone (1982). Hence if p is fixed and d is also small, but slowly increasing with N, such as $d \geqslant \log N$, then there exists no consistent estimator of $z \mapsto s_0(z)$ generally. Hence, generic ML estimators cannot be regarded as consistent, unless further assumptions are made. Examples of such assumptions include structured forms of linear and non-linear sparsity and super-smoothness.[6]

The problem of obtaining uniformly valid inference on $z \mapsto s_0(z)$ using generic ML methods is even more difficult. While the previous assumptions make consistent adaptive estimation possible (e.g., Bickel et al., 2009), confidence sets that adapt to unknown regularity (smoothness or sparsity) do not exist even for low-dimensional nonparametric problems (Low et al., 1997; Genovese and Wasserman, 2008)[7]. Construction of adaptive confidence bands then requires making additional untestable assumptions.[8]

In this paper, we take an agnostic view. We neither rely on any sparsity or low-complexity assumptions to make the ML estimators consistent, nor impose other stronger conditions to make “traditional” confidence intervals valid. We simply treat ML as providing proxy predictors for the objects of interest.

2.3. Our Approach.

To address the previous challenges, we propose strategies for estimation and inference on *key features* of $s_0(Z)$ rather than on $s_0(Z)$ itself. Because of this difference in focus we can avoid making strong assumptions about the properties of the ML estimators.

Let (M,A) denote a random partition of the set of indices $\{1,\dots,N\}$. The strategies that we consider rely on random splitting of $\text{Data} = (Y_i, D_i, Z_i)_{i=1}^N$ into a main sample, denoted by Data_M

[6] The function $z \mapsto s_0(z)$ is super-smooth if it has continuous and bounded derivatives of all orders.

[7] Let $z \mapsto s_0(z)$ be a target function that lives in an infinite-dimensional class with unknown regularity s (e.g., smoothness or degree of sparsity). Adaptive consistent estimation (resp. inference) for $z \mapsto s_0(z)$ with respect to s is possible if there exists a consistent estimator (resp. valid confidence set) with a rate of convergence (resp. diameter) that changes with s in a (nearly) rate-optimal way.

[8] See, e.g., Giné and Nickl (2010), where self-similarity conditions are used in low-dimensional nonparametric problems.

$= (Y_i, D_i, Z_i)_{i \in M}$, and an auxiliary sample, denoted by $\text{Data}_A = (Y_i, D_i, Z_i)_{i \in A}$. We will sometimes refer to these samples as M and A. After splitting the sample, we carry out two stages:

Stage 1: From the auxiliary sample A, we obtain ML estimators of the baseline and treatment effects, which we call the ML proxy predictors,

$$z \mapsto B(z) = B(z; \text{Data}_A) \text{ and } z \mapsto S(z) = S(z; \text{Data}_A).$$

Here $S(Z)$ is a possibly biased and noisy predictor of $s_0(z)$ and $B(Z)$ is a possibly biased and noisy predictor of $b_0(Z)$ (or other technical "baseline" functions, as we discuss below). We do not require these predictors to be consistent for the true functions.

Stage 2: We post-process the proxies from Stage 1 to estimate and make inference on features of the CATE function $z \mapsto s_0(z)$ in the main sample M. The key features that we target include:

(1) **Best Linear Predictor** (BLP) of the CATE $s_0(Z)$ on the ML proxy predictor $S(Z)$;

(2) **Sorted Group Average Treatment Effects** (GATES): average of $s_0(Z)$ (ATE) by heterogeneity groups induced by the ML proxy predictor $S(Z)$;

(3) **Classification Analysis** (CLAN): average characteristics of the most and least affected units defined in terms of the ML proxy predictor $S(Z)$.

Our approach is *generic* with respect to the ML method being used, and is *agnostic* about its formal properties.

We use many data splits into main and auxiliary samples to produce robust estimators. We employ quantile aggregation of inference to combine results across splits. Specifically, for point estimation, we report the median of the estimated key features over different random splits of the data. We take medians and other quantiles of many random conditional confidence sets for interval estimation. Finally, we construct p-values by taking medians of many random conditional p-values. We establish the formal inferential properties of this procedure.

2.4. **Relationship to the Literature.** We focus the review strictly on the literatures about estimation and inference on heterogeneous effects and inference using sample splitting.

This work is related to the literature that uses linear and semiparametric regression methods for estimation and inference on heterogeneous effects. Crump et al. (2008) developed tests of treatment effect homogeneity for low-dimensional settings based on traditional series estimators of the CATE. A semiparametric inference method for characterizing heterogeneity, called the sorted effects method, was given in Chernozhukov et al. (2015). This approach does provide a full set of inference tools, including simultaneous bands for percentiles of the CATE, but is strictly limited to the traditional semiparametric estimators of the regression and causal functions. Hansen et al. (2017) proposed a sparsity-based method called "targeted undersmoothing" to perform inference on heterogeneous effects. This approach does allow for high-dimensional settings, but makes strong assumptions on sparsity as well as additional assumptions that enable the targeted undersmoothing. A related approach, which allows for simultaneous inference on many coefficients (for example, inference on the coefficients corresponding to the interaction of the treatment with other variables) was first given in Belloni et al. (2013) using a Z-estimation framework, where the number of interactions can be very large; see also Dezeure et al. (2016) for a more recent effort in this direction, focusing on de-biased lasso in mean regression problems. This approach, however, still relies on a strong form of sparsity assumptions. Zhao et al. (2017) proposed a post-selection inference framework within high-dimensional linear sparse models for the heterogeneous effects. The approach is attractive because it allows for some misspecification of the model.

Another approach is to use tree-based and other methods. Imai and Ratkovic (2013) discussed the use of a heuristic support-vector-machine method with lasso penalization for classification of heterogeneous treatments into positive and negative ones. They used the Horvitz-Thompson transformation of the outcome (e.g., as in Hirano et al., 2003; Abadie, 2005) such that the new outcome becomes an unbiased, noisy version of CATE.[9] Athey and Imbens (2016) made use of the Horvitz-Thompson transformation of the outcome to inform the process of building causal trees, with the main goal of predicting CATE. They also provided a valid inference result on average treatment effects for groups defined by the tree leaves, conditional on the data split into two subsamples: one used to build the tree leaves and the one to estimate the predicted values given the leaves. Like our methods, this approach is essentially assumption-free. Our generic approach is not limited to trees and does hedge the splitting risks, and so it can be applied together with the causal trees. Wager and Athey (2017) proposed a subsampling-based construction of a causal random forest, providing valid pointwise inference for CATE (see also the review in Wager and Athey (2017) on prior uses of random forests in causal settings) for the case when covariates are very low-dimensional (and essentially uniformly distributed).[10] Unfortunately, this condition rules out the typical high-dimensional settings that arise in many empirical problems, especially in current RCTs, where the number of baseline covariates is potentially very large.

Several other references look at model-based strategies for performing inference on CATE, but relative to our approach, the assumptions invoked are quite strong. Semenova and Chernozhukov (2021) used ML to perform inference on the "partial" CATE, $\mathrm{E}[s_0(Z) \mid X]$, where X is a prespecified low-dimensional set of covariates. Specifically, they constructed an estimator of a denoised HT transform of the outcome and projected it using a nonparametric series estimator on the set

[9] Note that using Horvitz-Thompson (HT) transform of outcome, in this and other references, typically gives very noisy signal. One can improve the approach by either including the HT transform interacted with some baseline covariates as regressors in a regression model, as we do in the present paper, or using residualized outcomes in conjunction with HT, as in Semenova and Chernozhukov (2021).

[10] The dimension d is fixed in Wager and Athey (2017); the analysis relies on the Stone's model with smoothness index $\beta = 1$, in which no consistent estimator exists once $d \geqslant \log n$ in the minimax sense.

of low-dimensional prespecified covariates of interest X, whose dimension is much lower than Z.[11] The main advantage of this approach is that it delivers familiar nonparametric inference on partial CATE (even though inference on the full CATE remains intractable). Fan et al. (2022), Zimmert and Lechner (2019), Chernozhukov et al. (2018), and Chernozhukov et al. (2021a) developed kernel versions of this procedure. Related ideas but based on partialling-out (using residualized outcomes and treatment) appear in Semenova et al. (2017), Nekipelov et al. (2022), Nie and Wager (2020), Foster and Syrgkanis (2019), and more recently (per ArXiv appearance) in Kennedy (2020). Relative to the approach taken here, the assumptions made in these papers are very strong, albeit delivering stronger results. For example, in the approach of Semenova and Chernozhukov (2021), one has to specify the baseline covariates X for the partial CATE analysis, which is exactly what we are trying to avoid in our approach. Second, the methods critically rely on the consistency of ML to estimate the nuisance components well. The latter is a strong assumption in high dimensions, as discussed above.

The idea of using a "hold out" sample to validate the result of a ML procedure to discover heterogeneity was suggested in Davis and Heller (2020), who used the method proposed in Wager and Athey (2017) and compared their results to the heterogeneity in a holdout sample. Our inference approach is different because it calls for multiple splits. This procedure itself is also of independent interest and could be applied to many problems, where sample splitting is used to produce ML predictions (e.g., Abadie et al., 2017). Related references include Wasserman and Roeder (2009), and Meinshausen et al. (2009), where the ideas are related, but the details are quite different, as we shall explain below. The premise is the same; however, as in Meinshausen et al. (2009) and Rinaldo et al. (2016) – we should not rely on a single random split of the data and should adjust inference in some way. Our construction of p-values builds upon ideas in Meinshausen et al. (2009), though what we propose is simpler, and our confidence intervals appear to be new. Of course, sample splitting ideas are classical, going back to Hartigan (1969); Kish and Frankel (1974); Barnard (1974); Cox (1975); Mosteller and Tukey (1977), though having been mostly underdeveloped and overlooked for inference, as characterized by Rinaldo et al. (2016). Finally, our inference method shares with the literature on post-selection inference in statistics that the target estimands are random functions depending on a ML proxy (e.g., Fithian et al., 2014; Lee et al., 2016).

3. Main Identification Results and Estimation Strategies

In what follows, we observe Data $:= (Y_i, Z_i, D_i)_{i=1}^N$, consisting of i.i.d. copies of the random vector (Y, Z, D) having probability law P. The data are defined on an underlying probability space

[11] Specifically, the denoised HT transform of outcome is $\tilde{Y} = g(1,Z) - g(0,Z) + H(Y - g(D,Z))$, where $H = (D - p(Z))/[p(Z)(1 - p(Z)]$ and $g(D,Z) = \mathrm{E}(Y \mid D,Z)$. Semenova and Chernozhukov (2021) used ML to estimate $g(D,Z)$ and the propensity score $p(Z)$, in case the latter is unknown.

with measure P. The expectation operator is denoted by E. When we need to emphasize the dependence of P and E on P, we use the notation P_P and E_P. In what follows we condition on Data_A and therefore consider the functions

$$z \mapsto B(z) = B(z; \mathrm{Data}_A) \text{ and } z \mapsto S(z) = S(z; \mathrm{Data}_A).$$

as fixed functions. Alternatively, in this section, we can interpret E as conditional expectation, where we condition on Data_A.

3.1. **Best Linear Predictor of CATE.**

The first inferential target is the best linear predictor of the CATE using the proxy $S(Z)$.

Definition 3.1 (BLP). *The best linear predictor of $s_0(Z)$ by $S(Z)$ is the solution to:*

$$\min_{b_1, b_2} \mathrm{E}[s_0(Z) - b_1 - b_2 S(Z)]^2,$$

which, if exists, is defined as

$$\mathsf{BLP}[s_0(Z) \mid S(Z)] := \beta_1 + \beta_2 (S(Z) - \mathrm{E}S(Z)),$$

where $\beta_1 = \mathrm{E}s_0(Z)$ and $\beta_2 = \mathrm{Cov}[s_0(Z), S(Z)] / \mathrm{Var}[S(Z)]$.

By construction, $\mathsf{BLP}[s_0(Z) \mid S(Z)]$ is an unbiased predictor of $s_0(Z)$, which improves over $S(Z)$ in the mean-squared error sense, that is

$$\mathrm{E}\{s_0(Z) - \mathsf{BLP}[s_0(Z) \mid S(Z)]\}^2 \leqslant \mathrm{E}[s_0(Z) - S(Z)]^2.$$

Indeed, we can quantify the improvement by

$$\mathrm{E}[s_0(Z) - S(Z)]^2 - \mathrm{E}\{s_0(Z) - \mathsf{BLP}[s_0(Z) \mid S(Z)]\}^2 = (1 - \beta_2)^2 \mathrm{Var}[S(Z)] + [\mathrm{E}S(Z) - \mathrm{E}s_0(Z)]^2,$$

which is positive unless $S(Z)$ is an unbiased predictor and, either $\beta_2 = 1$ or $\mathrm{Var}[S(Z)] = 0$.[12] Accordingly, compared to the ML proxy, the BLP can be seen as a refined predictor of the individual CATE, $s_0(Z)$. If $S(Z)$ is a perfect proxy for $s_0(Z)$, then $\beta_2 = 1$. In general, $\beta_2 \neq 1$, correcting for noise in $S(Z)$. If $S(Z)$ is complete noise, uncorrelated to $s_0(Z)$, then $\beta_2 = 0$. Furthermore, if there is no heterogeneity, that is, $s_0(Z) = s$, then $\beta_2 = 0$. Rejecting the hypothesis $\beta_2 = 0$ therefore means that there is both heterogeneity in $s_0(Z)$ and $S(Z)$ is a relevant predictor.

We provide two strategies for identifying and estimating $\mathsf{BLP}[s_0(Z) \mid S(Z)]$.

[12] The previous expression follows from the decompositions $\mathrm{E}[s_0(Z) - S(Z)]^2 = \mathrm{E}[\{s_0(Z) - \mathrm{E}s_0(Z)\} - \{S(Z) - \mathrm{E}S(Z)\} + \{\mathrm{E}s_0(Z) - \mathrm{E}S(Z)\}]^2$ and $\mathrm{E}\{s_0(Z) - \mathsf{BLP}[s_0(Z) \mid S(Z)]\}^2 = \mathrm{E}\{[s_0(Z) - \mathrm{E}s_0(Z)] - [\mathsf{BLP}[s_0(Z) \mid S(Z)] - \mathrm{E}s_0(Z)]\}^2$, using that $\mathrm{E}\mathsf{BLP}[s_0(Z) \mid S(Z)] = \mathrm{E}s_0(Z)$ and $\beta_2 = \mathrm{Cov}(s_0(Z), S(Z)) / \mathrm{Var}(S(Z))$.

Strategy A: Weighted Residual BLP. Consider the weighted linear projection:

$$Y = \alpha_0' X_1 + \alpha_1 (D - p(Z)) + \alpha_2 (D - p(Z))(S - \mathrm{E}S) + \varepsilon, \;\; \mathrm{E}[w(Z)\varepsilon X] = 0, \tag{3.1}$$

where $S := S(Z)$, $w(Z) := \{p(Z)(1-p(Z))\}^{-1}$, $X := (X_1', X_2')'$,

$$X_1 = [1, B(Z), p(Z), p(Z)S(Z)]', \quad X_2 := [D - p(Z), (D - p(Z))(S - \mathrm{E}S)]'.$$

The term $B(Z)$ could be replaced by any "noise-reducing" proxy function. For example, the algorithms of Semenova et al. (2017) and Nie and Wager (2020), targeting the delivery of $S(Z)$, also construct $B(Z)$ that are meant to approximate $\mathrm{E}[Y \mid Z]$ but not $b_0(Z)$; see also Section 5 for other examples of such algorithms. We include X_1 to reduce finite sample noise of the estimators of the BLP parameter based on this strategy.[13]

Note that α_1 and α_2 are identified under weak assumptions. Further, we note that the interaction $(D - p(Z))(S - \mathrm{E}S)$ is orthogonal to $D - p(Z)$ under the weight $w(Z)$, and to all functions of Z such as X_1. Consequently, we obtain the following result that shows that the linear projection (3.1) identifies the BLP.

Theorem 3.1 (BLP Identification A). *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y and X have finite second moments, and* $\mathrm{E}XX'$ *is finite and full rank, which requires* $\mathrm{Var}(S(Z)) > 0$. *Then,* (α_1, α_2) *defined in* (3.1) *identifies the coefficients of the BLP,*

$$\alpha_1 = \beta_1, \quad \alpha_2 = \beta_2.$$

Comment 3.1 (Why not Classical OLS of Y on Proxies?). It is tempting and perhaps more natural to consider the projection equation:

$$Y = \underbrace{\tilde{\alpha}_1 + \tilde{\alpha}_2 B + \tilde{\beta}_1 D + \tilde{\beta}_2 D(S - \mathrm{E}S)}_{\text{BLP of CEF}} + \varepsilon, \quad \mathrm{E}[\varepsilon \tilde{X}] = 0,$$

where $\tilde{X} = [1, B, D, D(S - \mathrm{E}S)]'$. The idea here is the classical one: the ordinary least squares method with Y as the outcome provides the Best Linear Predictor or Approximation to the CEF $\mathrm{E}[Y \mid D, Z]$, even if the latter is nonlinear. Angrist and Pischke (2008) discuss the importance and practical relevance of this property. However, this property does not translate into providing the BLP of CATE $s_0(Z)$.[14] Indeed, even in pure RCTs, while $\tilde{\beta}_1 = \beta_1$ is true, we have that $\tilde{\beta}_2 \neq \beta_2$ in general, and therefore

$$\tilde{\beta}_1 + \tilde{\beta}_2 (S - \mathrm{E}S) \neq \mathsf{BLP}(s_0(Z) \mid S). \tag{3.2}$$

In this case, a sufficient condition for having equality in (3.2) is that (a) $B - \mathrm{E}B$ spans $S - \mathrm{E}S$, or (b) both $b_0(Z)$ and $s_0(Z)$ are orthogonal to $B - \mathrm{E}B$ and $S - \mathrm{E}S$, or (c) $B - \mathrm{E}B$ and $b_0(Z)$ are orthogonal to $S - \mathrm{E}S$. (See Online Appendix A.1 for the proof.) None of these conditions are plausible. ■

[13] Note that X_1 can include other functions of Z. In our experiments, the use of $B(Z)$ strongly improves the precision of estimating BLP (and other quantities such as GATEs and CLAN introduced below).

[14] This answers a question from Joshua Angrist; we are grateful to him for posing it.

The identification result in Theorem 3.1 is constructive. We can base a corresponding estimation strategy on the empirical analog:

$$\begin{aligned} Y_i &= \widehat{\alpha}_0' X_{1i} + \widehat{\alpha}_1 (D_i - p(Z_i)) + \widehat{\alpha}_2 (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i) + \widehat{\varepsilon}_i, \ i \in M, \\ \mathbb{E}_{N,M} &[w(Z_i) \widehat{\varepsilon}_i X_i] = 0, \end{aligned} \tag{3.3}$$

where $X_i = [X_{1i}', X_{2i}']'$, $X_{2i} := [D_i - p(Z_i), (D_i - p(Z_i))(S_i - \mathbb{E}_{N,M} S_i)]'$, and $\mathbb{E}_{N,M}$ denotes the empirical expectation with respect to the main sample, i.e.

$$\mathbb{E}_{N,M} h(Y_i, D_i, Z_i) := |M|^{-1} \sum_{i \in M} h(Y_i, D_i, Z_i).$$

The properties of this estimator, conditional on the auxiliary data, are well known and follow as a special case of Lemma D.1 in the Appendix.

Figure 1 provides two examples. The left panel shows a case where $s_0(Z) = 0$ with zero effect and zero heterogeneity in the CATE, whereas the right panel shows a case where $s_0(Z) = Z$ with strong heterogeneity in the CATE. In both cases, we evenly split 1,000 observations between the auxiliary and main samples, Z follows uniform distribution on $(-1, 1)$, $b_0(Z) = 3Z$, U is standard normal, independently of Z, and Y is generated by (2.6). We obtain the proxy predictor $S(Z)$ by Breiman's random forest, using the ranger implementation in R (Wright and Ziegler, 2017). We also report analogous results for causal random forest based on subsampling.

In the first example, the ML proxy is pure noise by construction, and the BLP post-processor correctly eliminates the noise, producing a CATE prediction that is roughly a 35-39% better approximation to the CATE under the RMSE metric. Furthermore, using our inferential methods of Section 4, we cannot reject the null hypothesis that the BLP is zero. In the second case, under strong heterogeneity, the signal in the ML proxy dominates the noise component. As a result, the BLP does not change the ML proxy drastically, but still gives a meaningful improvement to the CATE under the RMSE metric. These improvements agree with the theoretical arguments given above.

Comment 3.2 (Significance for RCTs.)**.** The first example has implications for the empirical analysis of RCTs. Here we see that one of the best ML algorithms, as per Friedman et al. (2001), can easily suggest a heterogeneous CATE when the treatment is, in fact, a placebo. Placebos (ineffective treatments) are common occurrences in real-world experiments, and our methodology provides a simple way to confirm that the CATE (and not just ATE) is indeed zero in such cases.

Strategy B: HT BLP. This strategy makes use of the Horvitz-Thompson transform H defined in (2.9). It is well known that the transformed response YH provides an unbiased signal about CATE:

$$\mathrm{E}[YH \mid Z] = s_0(Z),$$

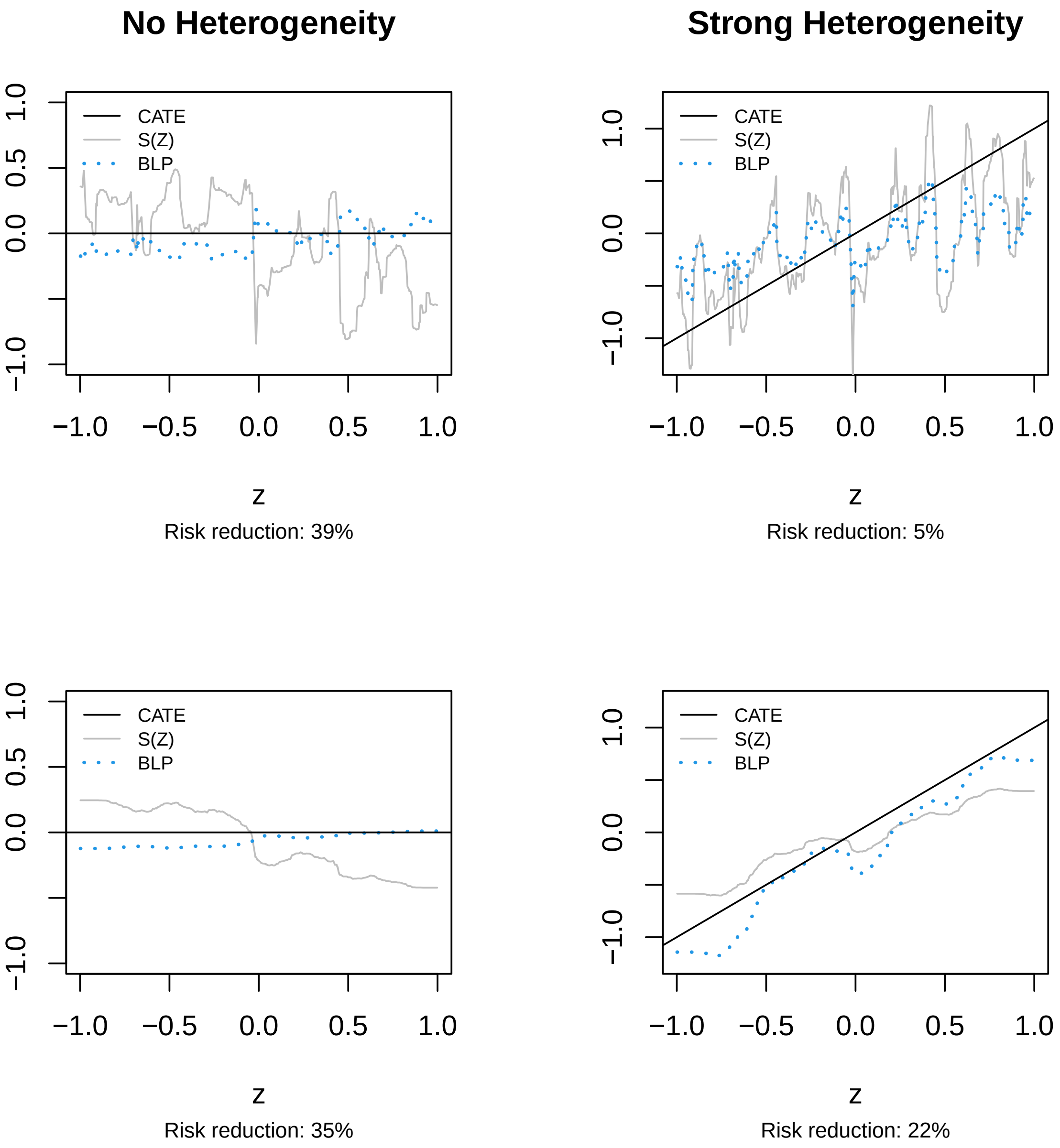


FIGURE 1. BLP Using ML Proxy vs the ML Proxy (Predictive and Causal RF)

NOTES: The CATE is plotted with the solid black line; the proxy predictor $S(Z)$, produced by Random Forest (Causal Random Forest in bottom panels), is plotted with the solid grey (light) line; and the BLP is plotted with the dotted blue line. In both panels, the BLP is less noisy than the ML proxy. In the left panels the BLP improves the RMSE by $35-39\%$; in the right panels BLP improves the RMSE by $5-22\%$.

and it follows by the properties of the best linear predictor that

$$\mathsf{BLP}[s_0(Z) \mid S(Z)] = \mathsf{BLP}[YH \mid S(Z)].$$

Note that $\mathsf{BLP}[s_0(Z) \mid S(Z)]$ is a more precise unbiased predictor of $s_0(Z)$ than YH because, by construction

$$\mathrm{Var}(\mathsf{BLP}[s_0(Z) \mid S(Z)]) = \mathrm{Var}(\mathsf{BLP}[YH \mid S(Z)]) \leqslant \mathrm{Var}(YH).$$

The R-squared of $\mathsf{BLP}[YH \mid S(Z)]$ quantifies the percent reduction in variance of the BLP relative to YH.

The simple linear projection $\mathsf{BLP}[YH \mid S(Z)]$ is completely fine for identification purposes, but can severely underperform in estimation and inference due to lack of precision. We can repair the deficiencies by considering, instead, the linear projection:

$$YH = \mu_0' X_1 H + \mu_1 + \mu_2 (S - \mathrm{E}S) + \varepsilon, \quad \mathrm{E}\varepsilon\tilde{X} = 0, \tag{3.4}$$

where $\tilde{X} := (X_1'H, \tilde{X}_2')'$, $\tilde{X}_2 := (1, S - \mathrm{E}S)'$, and $X_1 = [1, B(Z), p(Z), p(Z)S(Z)]'$ as before. The term X_1 could contain other functions of Z. We include $X_1 H$ in order to *reduce noise*.

The following theorem shows that the linear projection (3.4) also identifies the BLP.

Theorem 3.2 (BLP Identification B)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments, $\tilde{X}$ is such that $\mathrm{E}\tilde{X}\tilde{X}'$ is finite and full rank, which requires* $\mathrm{Var}(S(Z)) > 0$*. Then, (μ_1, μ_2) defined in* (3.4) *identifies the coefficients of the BLP,*

$$\mu_1 = \beta_1, \quad \mu_2 = \beta_2.$$

Theorem 3.2 leads to an estimator defined through the empirical analog:

$$Y_i H_i = \widehat{\mu}_0' X_{1i} H_i + \widehat{\mu}_1 + \widehat{\mu}_2 (S_i - \mathbb{E}_{N,M} S_i) + \widehat{\varepsilon}_i, \ \ i \in M, \ \ \mathbb{E}_{N,M} \widehat{\varepsilon}_i \tilde{X}_i = 0, \tag{3.5}$$

and the properties of this estimator, conditional on the auxiliary data, are well known and given in Lemma D.1.

Comment 3.3 (Comparison of Estimation Strategies)**.** While the two identification strategies are natural, one may wonder whether the two corresponding estimation strategies can be ranked in terms of asymptotic efficiency. We show in Online Appendix A.2 that they produce estimators that are first-order equivalent in large main samples.

3.2. **Sorted Group Average Treatment Effects.**

The second inferential target is the group average treatment effects, where the groups are induced by $S(Z)$.

Definition 3.2 (GATES)**.** *The Sorted Group Average Treatment Effects (GATES) are*

$$\gamma_k := \mathrm{E}[s_0(Z) \mid G_k], \quad k = 1, \ldots, K.$$

where $G_k := \{S \in I_k\}$, with $I_k := [\ell_{k-1}, \ell_k)$ and $-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty$.

Comment 3.4 (Choice of groups)**.** We build the groups to explain as much variation in $s_0(Z)$ as possible. There are many alternatives for creating groups based upon ML tools applied to the auxiliary data. For example, one can group or cluster based upon predicted baseline response as in the "endogenous stratification" analysis (Abadie et al., 2017), or based upon actual predicted treatment effect S. We focus on the latter approach for defining groups, although our identification and inference ideas immediately apply to other ways of defining groups, and could be helpful in these contexts. The causal tree approach of Athey and Imbens (2016) can also be viewed as a GATES analysis, with a specific way of forming groups via recursive partitioning.[15]

Comment 3.5 (GATES as Predictors of CATE)**.** The GATES can also be used as nonlinear predictors of the CATE based on the proxy S, in a similar fashion to the BLP. Indeed, the GATES provide the BLP of CATE using the group indicators $G_k, k = 1,\dots,K$.

We provide two strategies for identifying and estimating the GATES.

Strategy A: Weighted Residual GATES. Consider the weighted linear projection equation:

$$Y = \alpha_0' X_1 + \sum_{k=1}^{K} \alpha_k \cdot [D - p(Z)] \cdot 1(G_k) + \nu, \quad \mathrm{E}[w(Z)\nu W] = 0, \tag{3.6}$$

where $W := (X_1', W_2')'$, X_1 contains a vector of functions of Z, e.g., $X_1 = (B(Z), p(Z)\{1(G_k)\}_{k=1}^K)'$ and $W_2 := (\{[D - p(Z)] \cdot \{1(G_k)\}_{k=1}^K)'$. The presence of $D - p(Z)$ in the interaction $[D - p(Z)] \cdot 1(G_k)$ *orthogonalizes* this regressor relative to all other regressors that are functions of Z, such as X_1. The controls in X_1, as in the BLP estimation, are included to reduce noise.

Theorem 3.3 below shows that the linear projection (3.6) identifies the GATES. We can therefore base an estimation strategy on the empirical analog:

$$Y_i = \widehat{\alpha}_0' X_{1i} + \widehat{\alpha}' W_{2i} + \widehat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[w(Z_i)\widehat{\nu}_i W_i] = 0, \tag{3.7}$$

where $\widehat{\alpha} = (\widehat{\alpha}_1, \dots, \widehat{\alpha}_K)'$. The properties of this estimator, conditional on the auxilliary data, are well known and stated as a special case of Lemma D.1.

Strategy B: HT GATES. Here we employ a linear projection on Horvitz-Thompson transformed variables:

$$YH = \mu_0' X_1 H + \sum_{k=1}^{K} \mu_k \cdot 1(G_k) + \upsilon, \quad \mathrm{E}[\upsilon \tilde{W}] = 0, \tag{3.8}$$

where $\tilde{W} := (X_1' H, \tilde{W}_2')'$, X_1 includes functions of Z, e.g. X_1 the same as above, and $\tilde{W}_2 := [\{1(G_k)\}_{k=1}^K]'$.

[15] Another strand of the literature related to the GATES is the learning policy problem, where a ML method is trained to assign units to treatment and control based on their covariates(e.g., Kitagawa and Tetenov, 2018; Athey and Wager, 2021). This problem can be seen as a GATES analysis with two groups chosen to maximize some function of the CATEs.

Theorem 3.3 shows that the linear projection (3.8) also identifies the GATES. We can therefore base an estimation strategy on the empirical analog:

$$Y_i H_i = \widehat{\mu}_0' X_{1i} H_i + \widehat{\mu}' \tilde{W}_{2i} + \widehat{\upsilon}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[\widehat{\upsilon}_i \tilde{W}_i] = 0, \tag{3.9}$$

where $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_K)'$. The properties of this estimator, conditional on the auxiliary data, are well known and given in Lemma D.1. The resulting estimator has similar performance to the estimator in (3.7), and under some conditions their first-order properties coincide.

We now provide a formal statement of the identification results.

Theorem 3.3 (GATES)**.** *Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments and W and $\tilde{W}$ are such that* $\mathrm{E}WW'$ *and* $\mathrm{E}\tilde{W}\tilde{W}'$ *are finite and have full rank. Consider $\alpha = (\alpha_k)_{k=1}^K$ defined by the weighted regression equation* (3.6) *and $\mu = (\mu_k)_{k=1}^K$ defined by the regression equation* (3.8)*. These coefficients are equal and identify the GATES:*

$$\alpha_k = \mu_k = \gamma_k = \mathrm{E}[s_0(Z) \mid G_k], \quad k = 1, ..., K.$$

Comment 3.6 (Motonicity Restrictions on GATES)**.** Suppose we observe $s_0(Z)$. In this case we can define the ideal GATES as:

$$\gamma_{0k} := \mathrm{E}[s_0(Z) \mid G_{0k}], \quad k = 1, \ldots, K,$$

where $G_{0k} := \{s_0(Z) \in I_{0k}\}$, with $I_{0k} = [\ell_{0,k-1}, \ell_{0,k})$ and $-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty$. By construction the ideal GATES obey the monotonicity restriction:

$$\gamma_{01} \leqslant ... \leqslant \gamma_{0K}.$$

If $S(Z)$ provides a good approximation to $s_0(Z)$, it is reasonable to expect that the GATES also obey the monotonicity restriction: $\gamma_1 \leqslant ... \leqslant \gamma_K$, but there is no guarantee. However, we can always replace $\gamma = \{\gamma_k\}_{k=1}^K$ by the non-decreasing rearrangement (sorted vector) $\gamma^* = \{\gamma_k^*\}_{k=1}^K$, such that γ^* obeys the monotonicity condition $\gamma_1^* \leqslant ... \leqslant \gamma_K^*$. The benefit is that γ^* is always closer to $\gamma_0 = \{\gamma_{0k}\}_{k=1}^K$ than γ in the sense that

$$\|\gamma^* - \gamma_0\|_\infty \leqslant \|\gamma - \gamma_0\|_\infty,$$

where $\|\cdot\|_\infty$ is the sup-norm. This follows from the contraction property of the rearrangement (e.g., Chernozhukov et al., 2009). Therefore, we can always use sorting to better target the ideal GATES. Similarly, when performing estimation, we can replace $\widehat{\gamma} = \{\widehat{\gamma}_k\}_{k=1}^K$ by the their non-decreasing rearrangement (sorted vector) $\widehat{\gamma}^* = \{\widehat{\gamma}_k^*\}_{k=1}^K$, which results in an estimator with lower estimation error in the sense that surely:

$$\|\widehat{\gamma}^* - \gamma_0\|_\infty \leqslant \|\widehat{\gamma} - \gamma_0\|_\infty.$$

3.3. **Classification Analysis.** When the BLP and GATES analyses reveal substantial heterogeneity, it is interesting to know the properties of the subpopulations that are the most and least affected. Here we focus on the "least affected group" G_1 and "most affected group" G_K, where the labels "most" and "least" can be swapped depending on the context.

Definition 3.3 (CLAN). *Let $g(Y,D,Z)$ be a vector of characteristics of an observational unit. The classification analysis (CLAN) is the comparison of the average characteristics of the most and least affected groups:*

$$\delta_1 := \mathrm{E}[g(Y,D,Z) \mid G_1] \quad \textit{and} \quad \delta_K := \mathrm{E}[g(Y,D,Z) \mid G_K].$$

The parameters δ_1 and δ_K are identified because they are averages of variables that are directly observed. The CLAN quantifies the differences between the most and least affected groups and singles out the covariates that are associated with the heterogeneity in the CATE. The CLAN can be extended to comparisons of features other than averages, such as variances, covariances or distributions. In the empirical analysis, we estimate the CLAN parameters by taking averages in M:

$$\widehat{\delta}_1 = \frac{\mathbb{E}_{N,M}[g(Y_i,D_i,Z_i)G_{1,i}]}{\mathbb{E}_{N,M}G_{1,i}} \quad \text{and} \quad \widehat{\delta}_K = \frac{\mathbb{E}_{N,M}[g(Y_i,D_i,Z_i)G_{K,i}]}{\mathbb{E}_{N,M}G_{K,i}}, \tag{3.10}$$

using $G_{k,i} = 1\{S(Z_i) \in I_k\}$, where $I_k = [\ell_{k-1}, \ell_k)$ and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i \in M}$.

3.4. **Goodness of Fit Measures for Fitting CATE.** In practical applications it is useful to have goodness-of-fit measures to guide the selection of ML proxies.

For the analysis based on the BLP of CATE, we propose to use:

$$\Lambda := |\beta_2|^2 \mathrm{Var}(S(Z)) = \mathrm{Corr}(s_0(Z), S(Z))^2 \mathrm{Var}(s_0(Z)). \tag{3.11}$$

Maximizing Λ is equivalent to maximizing the correlation between the ML proxy predictor $S(Z)$ and the true score $s_0(Z)$, or equivalent to maximizing the R^2 in the regression of $s_0(Z)$ on $S(Z)$. Therefore, an ML method that attains a higher Λ is a preferred method.

Analogously, for the GATES analysis, we propose to use:

$$\bar{\Lambda} = \mathrm{E}\left(\sum_{k=1}^{K} \gamma_k 1(S \in I_k)\right)^2 = \sum_{k=1}^{K} \gamma_k^2 \mathrm{P}(S \in I_k). \tag{3.12}$$

This is the part of variation of $s_0(z)$, $\mathrm{E}s_0(Z)^2$, explained by $\bar{S}(Z) = \sum_{k=1}^{K} \gamma_k 1(S(Z) \in I_k)$. Hence choosing the ML proxy $S(Z)$ to maximize $\bar{\Lambda}$ is equivalent to maximizing the R^2 in the regression of $s_0(Z)$ on $\bar{S}(Z)$ (without a constant). If the groups $G_k = \{S \in I_k\}$ have equal size, namely $\mathrm{P}(S(Z) \in I_k) = 1/K$ for each $k = 1,...,K$, then

$$\bar{\Lambda} = \frac{1}{K} \sum_{k=1}^{K} \gamma_k^2.$$

Therefore, a ML method that attains a higher $\bar{\Lambda}$ is a preferred method. The empirical versions of the parameters above are:

$$\widehat{\Lambda} = |\widehat{\beta}_2|^2 \mathbb{E}_{N,M}(S_i - \mathbb{E}_{N,M} S_i)^2, \qquad \widehat{\bar{\Lambda}} = \sum_{k=1}^{K} \widehat{\gamma}_k^2 \mathbb{E}_{N,M} 1\{S_i \in I_k\}. \tag{3.13}$$

The choice of the ML method using goodness-of-fit measures does not pose any additional inferential challenge, when there is clearly an ML method that dominates the others – so that we select the best method with probability approaching one. This means that the inferential methods of the next section would not need any further adjustment. When this is not the case, there are several possibilities depending on the scientific reporting objectives. For example, suppose there are two (near) winners and we want to construct a $(1-\alpha)$-confidence set, as in our empirical analysis. Then in the spirit of sensitivity analysis, we can report the union of the $(1-\alpha)$-confidence sets. This ensures the inferential coverage guarantee of $1-\alpha$ continues to apply, if the reader of the empirical report chooses one or the other winner at random. On the other hand, the inferential guarantee needs to be discounted to $1-2\alpha$ via Bonferroni adjustment, if the reader of the empirical report chooses one or the other report depending on the empirical results themselves. We use the first approach because the readers of our empirical analysis are not likely to follow the latter approach.

Finally, if the data sets are big, we could use additional splitting to choose the best-performing ML method, before taking the resulting ML proxies to the main sample.[16]

4. Split-Sample Robust Estimation and Inference Methods

4.1. **Estimation and Inference: The Generic Targets.** Let θ denote a generic target parameter or functional. For example,

- $\theta = \beta_2$ is the BLP slope, the heterogeneity loading parameter;
- $\theta = \mathsf{BLP}[s_0(Z) \mid S(z)] = \beta_1 + \beta_2(S(z) - \mathrm{E}S)$ is the “personalized” predictor of CATE $s_0(z)$;
- $\theta = \gamma_k$ is the GATES for the group G_k;
- $\theta = \gamma_K - \gamma_1$ is the difference in the GATES between the most and least affected groups;
- $\theta = \delta_K - \delta_1$ is the difference in the expectation of the characteristics of the most and least impacted groups in CLAN.

Let (a,m) denote a fixed partition of $\{1,\dots,N\}$. In what follows, we recognize that the estimands depend on the auxiliary sample a,

$$\text{Data}_a := \{(Y_i, D_i, X_i)\}_{i \in a},$$

[16] We also refer to Section 5 which discusses other, more exploratory ideas for building the best ML algorithms for targeting CATE already in the first stage.

used to create the ML proxies $B = B_a$ and $S = S_a$:

$$\theta = \theta_a := \theta(\text{Data}_a).$$

We shall use the notation θ_a when we want to highlight this dependence. Moreover, unlike in Section 3, we made explicit the conditioning on Data_a in the expectations when needed.

Single Split. We begin the discussion of inference conditional on a single split of data induced by the partition $\{(a,m)\}$ of $\{1,...,N\}$ into sets of cardinality $(N-n,n)$. All of the examples admit an estimator $\widehat{\theta}_a$ that is approximately Gaussian, namely as $n \to \infty$ and for any z,

$$\mathrm{P}(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_a) < z \mid \text{Data}_a) \to_P \Phi(z). \tag{4.1}$$

We provide sufficient regularity conditions for (4.1) in Lemma D.1, which may be of independent interest. Indeed, the deployment of single-split inference requires verification of (4.1), but we could not find any reference establishing this condition for least squares estimators.

As a consequence of (4.1), the standard p-values

$$p_a^+ := 1 - \Phi(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_0)), \;\; p_a^- := \Phi(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_0)),$$

for testing the null hypothesis $\theta_a = \theta_0$ against the alternatives $\theta_a > \theta_0$ and $\theta_a < \theta_0$, respectively, are approximately uniform under the null, namely

$$\mathrm{P}(p_a^{\pm} < \alpha \mid \text{Data}_a) = \alpha + o_P(1).$$

As another consequence of (4.1), the standard confidence interval (CI)

$$[L_a, U_a] := [\widehat{\theta}_a \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}_a]$$

covers θ_a with approximate probability $1-\alpha$ conditional on Data_a:

$$\mathrm{P}[\theta_a \in [L_a, U_a] \mid \text{Data}_a] = 1 - \alpha - o_P(1).$$

Thus, we have straightforward inference conditional on a single data split.

Multiple Splits. In practice, researchers often prefer using multiple splits $(a,m)'s$ to reduce estimation risk and demonstrate that the results are robust to how they split the data. Therefore, we need a way to aggregate the results across different splits, and propose quantile aggregation methods and analyze their properties.

Definition 4.1 (Collection of Splits)**.** *Consider the collection $\{(a,m), a \in \mathscr{A}\}$ of partitions of $[N] = \{1,...,N\}$ into auxiliary sets a of cardinality $N-n$ and main sets m of cardinality n. We generate the collection independently of*

$$\text{Data} := (Y_i, D_i, X_i)_{i=1}^N.$$

Different partitions of $[N]$ yield different estimands and estimators. To formalize this randomness, we consider A as a uniform random variable taking values $a \in \mathscr{A}$, that is, $A \sim U(\mathscr{A})$. Therefore, conditional on Data, the estimand θ_A is a random variable. In what follows, we will mainly target our inference on the median value of θ_A. Furthermore, different partitions yield different estimators $\widehat{\theta}_A$ and approximate distributions for these estimators. Therefore, conditional on Data, estimators $\widehat{\theta}_A$, p-values p_A, and intervals $[L_A, U_A]$ are all random variables. We will use quantile aggregation methods to summarize them.

It is useful to recall some definitions of quantiles for discrete variables. For a random variable X with law P_X and k points of support, and index $u \in (0,1)$, the lower and upper u-quantiles are $\underline{\mathrm{Q}}_u(X) := \inf\{x \in \mathbb{R} : \mathrm{P}_X(X \leqslant x) \geqslant u\}$, and $\overline{\mathrm{Q}}_u(X) := \sup\{x \in \mathbb{R} : \mathrm{P}_X(X \geqslant x) \geqslant 1-u)$, respectively. For $w_u := \lfloor uk \rfloor / (\lfloor uk \rfloor + \lceil uk \rceil)$, we define

$$\mathrm{Q}_u(X) := w_u \underline{\mathrm{Q}}_u(X) + (1 - w_u)\overline{\mathrm{Q}}_u(X)$$

as the central quantile.[17] If X is continuous, all three definitions coincide. To define upper, lower, and central medians, we use $u = 1/2$ in the definitions above and M instead of Q_u.

We now formally define the median-aggregated p-value.

Definition 4.2 (Median-Aggregated P-value)**.** *The median p-values for testing one-sided alternative hypotheses are*

$$p^{\pm} = \mathrm{M}(p_A^{\pm} \mid \mathrm{Data}).$$

The two-sided median p-value is $\bar{p} = 2\min(p^+, p^-)$.

Aggregation using lower median p-values was first proposed by Meinshausen et al. (2009) in the context of split-sample hypothesis testing in linear regression with selection. Here we take the central medians, since they are more likely to behave like regular p-values.[18]

We next define the quantile-aggregated point and interval estimators.

Definition 4.3 (Quantile-Aggregated Point and Interval Estimators)**.** *The median point estimator is:*

$$\widehat{\theta} := \mathrm{M}[\widehat{\theta}_A \mid \mathrm{Data}].$$

The β-quantile confidence interval is $[L, U]$, where

$$L := Q_\beta(L_A \mid \mathrm{Data}), \quad U := Q_{1-\beta}(U_A \mid \mathrm{Data}), \quad \beta \leqslant 1/2.$$

We can interpret these definitions as risk-reducing inferential summaries.

[17] For example, the quantile function in R uses this definition (R Core Team, 2022).

[18] For example, the sample lower median of $\{U, 1-U\}$, $U \sim U(0,1)$, obeys $P(\underline{\mathrm{M}} < \alpha) = 2\alpha$ for $\alpha < 1/2$. In contrast, the central median obeys $\mathrm{P}(\mathrm{M} < \alpha) < \alpha$ for $\alpha < 1/2$.

Lemma 4.1 (Risk Contraction)**.** *Consider any fixed target value $\theta' \in \mathbb{R}$. Then $\widehat{\theta}$ is more concentrated near θ' than any single-split generated $\widehat{\theta}_A$:*

$$\mathrm{E}|\widehat{\theta} - \theta'| \leqslant \mathrm{E}|\widehat{\theta}_A - \theta'|. \tag{4.2}$$

Set $\beta = 1/2$. Then the confidence set $[L,U]$ has the same concentration property:

$$\mathrm{E}|U - \theta'| \vee \mathrm{E}|L - \theta'| \leqslant \mathrm{E}|U_A - \theta'| \vee \mathrm{E}|L_A - \theta'|. \tag{4.3}$$

Moreover, for any $\beta \in (0,1/2]$ the width of $[L,U]$ is weakly smaller than the worst-case width of the sets across splits:

$$|U - L| \leqslant \max_{a \in \mathscr{A}} |U_a - L_a|. \tag{4.4}$$

Analogous risk contraction properties hold for the mean aggregation, but we focus on medians for robustness reasons.

In what follows, we study the formal inferential guarantees of $[L,U]$. The default choice of β is $1/2$, but we obtain useful theoretical guarantees for other choices $\beta < 1/2$ as well.

Principal Regularity Condition. As the main regularity condition, we assume approximate normality of the split-sample t-statistics:[19]

(R1) There exist a sequence of positive constants $\gamma_N' \searrow 0$ as $(n,N) \to \infty$, such that

$$\sup_{z \in \mathbb{R}} |\mathrm{P}\{\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) < z\} - \Phi(z)| \leqslant \gamma_N'. \tag{4.5}$$

Suppose that the data $\{(Y_i, Z_i, D_i)\}_{i=1}^N$ are generated as i.i.d. copies of (Y,Z,D). In this case, for any $a \in \mathscr{A}$:

$$\mathrm{P}\{\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) < z\} = \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a \in \mathscr{A}} 1(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_a) < z)\right] = \mathrm{P}\{\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_a) < z\},$$

because the expression on the right does not depend on $a \in \mathscr{A}$ under the i.i.d. sampling. This observation simplifies the verification of (R1) for least squares type estimators; see Lemma D.1. While the i.i.d. case is our main focus, the main results in this section rely only on the conditions labeled R, which are likely to hold more generally.

Below we give various theoretical guarantees for our inferential summaries using this condition and adding more conditions to get stronger results.

[19] Here and below we use the standard error $\widehat{\sigma}_A$ instead of the theoretical standard deviation σ_A in all statements, but we can exchange the two if $\widehat{\sigma}_A/\sigma_A \to_P 1$, which holds under typical conditions, e.g. Lemma D.1.

4.2. **Hypothesis Testing with Multiple Splits.** We start the analysis by testing homogeneous hypotheses $\theta_A = \theta_0$, which imply that θ_a does not vary with a. Suppose, for example, that we want to test that the slope of the BLP is zero with probability one, $\beta_{2A} = 0$, against the alternative $\beta_{2A} > 0$ with positive probability. This problem amounts to both testing the heterogeneity in CATE and the relevance of the ML score S_A as a predictor. Another interesting hypothesis is whether $\beta_{2A} = 1$, with probability one, that is, whether S_A is well-calibrated and needs no post-processing.

More generally, suppose we are testing the hypothesis

$$H_0 : \theta_A = \theta_0, \tag{4.6}$$

with probability one, against $H_1^+ : \theta_A > \theta_0$ with positive probability. Testing using the median p-value p^+ will have power against the null when the majority of θ_a's violate the null, so we can interpret the rejection accordingly. Similarly, we can test against $H_1^- : \theta_A < \theta_0$ or $H_1 : \theta_A \neq \theta_0$ with positive probability.

Below we establish the properties of the median p-values under (R1). To get the sharpest results, we can invoke a concentration condition for approximate medians:

(R2) For all $z = \Phi^{-1}(\alpha)$, where the nominal level of interest α is in some closed sub-interval of $(0, 1/4)$, and some sequences of positive constants $\gamma_N'' \searrow 0$ and non-negative constants $\varepsilon_n \searrow 0$:

$$\begin{aligned} \mathrm{P}\left(\mathrm{Q}_{.5-\varepsilon_N}(\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A)|\mathrm{Data}) < z\right) &\leqslant \mathrm{P}\left(\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A) < z\right) + \gamma_N'', \\ \mathrm{P}\left(\mathrm{Q}_{.5-\varepsilon_N}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)|\mathrm{Data}) < z\right) &\leqslant \mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) < z\right) + \gamma_N''. \end{aligned} \tag{4.7}$$

This condition states that the approximate median over-the-splits t-statistics tend to concentrate more than any single-split t-statistic. This condition is intuitive, but it is hard to give general primitive conditions for it.[20] We believe (4.7) is quite plausible. We verified it for typical values of $\alpha < 1/4$ numerically in various experiments that mimic empirical applications, and were unable to find any counterexample.

Theorem 4.1 (Uniform Validity of Median-Aggregated P-Value)**.** *Suppose that the null hypothesis H_0 in* (4.6) *holds with probability one. Let p be either of $\{p^+, p^-, \bar{p}\}$. (i) Suppose that approximate normality (R1) holds, then*

$$\mathrm{P}(2p < \alpha) \leqslant \alpha + o(1),$$

[20] When the t-stats are independent, then their median concentrates in a fixed interval around 1/2 with probability approaching 1 exponentially fast; see DiCiccio et al. (2020). Therefore, the approximate median concentration condition holds. This happens when m's are non-overlapping and some further homogeneity conditions hold. On the other hand, suppose the p-values are the same asymptotically, then the inequality in the concentration condition binds, but does not fail. This situation is not common in our context, though.

where the $o(1)$ depends only on γ'_N. (ii) Suppose in addition that the median concentration condition (R2) holds with $\varepsilon_N = 0$, then

$$\mathrm{P}(p < \alpha) \leqslant \alpha + o(1),$$

where the $o(1)$ depends only on γ'_N and γ''_N.

Therefore under the median concentration condition, the median p-values have the standard property. Without the median concentration condition, the median p-values need to be multiplied by 2. However, based on our computational experiments, median p-values are conservative even for the nominal level α (once $\alpha < 1/4$), mainly due to the concentration property holding with $\gamma''_N < 0$. We therefore do not recommend multiplying by 2; see also DiCiccio et al. (2020) for a similar point. The exact form of $o(1)$, given in the proof, allows us to convert the results into those holding uniformly in a set of probability measures P. The proof of the first result partly relies on the idea of Meinshausen et al. (2009) to use Markov inequality to bound quantiles of an arbitrary collection of marginally uniform random variables.

4.3. **Prediction Intervals with Multiple Splits.** Outside of the settings with homogeneity, the estimand θ_A is a random variable, and we might be interested in characterizing its typical values. Our first approach serves this purpose, and is connected to conformal/permutation inference that is commonly used for predicting unobserved outcomes (Hoeffding, 1952; Vovk et al., 2005; Barber et al., 2022).

Here our goal is to have a prediction interval for θ_A, and the challenge we face is that θ_A is not observed directly, which places us outside the standard conformal setting. However, for each $a \in \mathscr{A}$, we have a (random) confidence interval $[L_a, U_a]$ that has the covering property:

$$\mathrm{P}(\theta_a < L_a) \leqslant \alpha/2 + o(1), \quad \mathrm{P}(\theta_a > U_a) \leqslant \alpha/2 + o(1). \tag{4.8}$$

This condition is implied by the basic regularity condition (R1) in our context.

We take the quantile-aggregated confidence interval $[L, U]$ as our prediction interval for θ_A.

Theorem 4.2 (Properties of the Prediction Interval)**.** *Suppose that (4.8) holds. Then*

$$\mathrm{P}(\theta_A < L) \leqslant \beta + \alpha/2 + o(1), \quad \mathrm{P}(\theta_A > U) \leqslant \beta + \alpha/2 + o(1),$$

where the $o(1)$ terms are the same as in (4.8). Therefore, $\mathrm{P}(\theta_A \in [L, U]) \geqslant 1 - 2\beta - \alpha - 2o(1)$.

We can use the prediction interval $[L, U]$ to characterize the “majority” of the central values of the random target θ_A that one could get from sample splitting. For this we set $\beta = .25$ and ”small” $\alpha = o(1)$, then $MP = [L, U]$ has the property:

$$\mathrm{P}(\theta_A \in MP) \geqslant .5 - o(1). \tag{4.9}$$

That is, MP contains majority of central values of θ_A. On the other hand, if we set $\beta = 1/2$, we get a median aggregated interval

$$MI = [L, U].$$

In this setting, we can think of MI as predicting the median of θ_A with small margin of error, if $\alpha = o(1)$:

$$\mathrm{P}(\theta_A < L) \leqslant 1/2 + o(1) \text{ and } \mathrm{P}(\theta_A > U) \leqslant 1/2 + o(1). \tag{4.10}$$

The latter gives us a useful interpretation of median confidence intervals, and notably this interpretation applies under the weakest possible regularity condition (R1) in our setting.

4.4. **Confidence Intervals for Median Parameter with Multiple Splits.** Instead of predicting the "majority" of θ_A, we may focus the inference on a single target.

Definition 4.4 (Inferential Target)**.** *Our inferential target is the median estimand:*

$$\theta^* = \mathrm{M}[\theta_A | \mathrm{Data}].$$

The choice of the target has the intuitive appeal of representing a typical value. Moreover, in many cases, θ_A will concentrate around its median value θ^*, making it an even more natural target. What follows is the principal regularity condition that covers this concentration scenario.

(R3) For some positive sequences of constants $r_N \searrow 0$ and $\gamma_N''' \searrow 0$ as $(n, N-n) \to \infty$,

$$\mathrm{P}\left(\widehat{\sigma}_A^{-1} |\theta^* - \theta_A| > r_N\right) \leqslant \gamma_N'''. \tag{4.11}$$

The concentration condition above is a convenient property for the interpretability of the inference. Here the rate of concentration of θ_A around θ^* should be faster than the rate $\widehat{\sigma}_A$ of $\widehat{\theta}_A$ estimating θ_A. Thus, the concentration condition implicitly requires the auxiliary set a to be large and the main set m to be small compared to a; so that randomness in the inferential target is small compared to the size of the estimation error $\widehat{\sigma}_a$ in the main sample, which typically is proportional to $n^{-1/2}$.

The condition (R3) is high-level; we demonstrate the plausibility of this condition for the BLP parameter in Appendix B.2 using notions of estimation and algorithmic stability. Estimation stability implies that θ_A concentrates around a fixed value $\theta_\bullet$, in which case the median also concentrates around $\theta_\bullet$. Estimation stability follows from the ML proxy S_A being consistent for some fixed proxy function $s_\bullet$, but not necessarily consistent for the true CATE. This condition can be readily verified using statistical learning theory, as we do in Section 5 for causal learners of CATE. The algorithmic stability condition is strictly weaker than estimation stability, though not as readily available. Our use of these stability criteria is inspired by similar ideas in Wager et al. (2016), Chernozhukov et al. (2021b), and Chen et al. (2022), applied to a different context. Appendix B.2 discusses all of this further.

The following results summarize the properties of the proposed median confidence interval under various conditions.

Theorem 4.3 (Properties of the Confidence Interval for θ^*)**.** *Let $\beta = 1/2$. (i) Suppose that (R1) and (R3) hold. Then,*

$$\mathrm{P}(\theta^* \in [L,U]) \geqslant 1 - 2\alpha - o(1),$$

where $o(1)$ depends only on γ'_N, γ'''_N and r_N. (ii) Suppose in addition that (R2) holds with $\varepsilon_N = 2\sqrt{\gamma'''_N}$. Then,

$$\mathrm{P}(\theta^* \in [L,U]) \geqslant 1 - \alpha - o(1),$$

where $o(1)$ depends only on $\gamma'_N, \gamma'''_N, \gamma''_N$ and r_N. (iii) In either case, the event $\theta^* \in [L,U]$ implies $|\theta^* - \widehat{\theta}| \leqslant |U - L|$.

Under the strongest assumptions, the target θ^* is covered with a probability of at least $1 - \alpha - o(1)$. Under the minimal set of assumptions, the coverage probability is $1 - 2\alpha - o(1)$. In our numerical results, the confidence intervals tend to be conservative even under the minimal condition, with coverage exceeding $1 - \alpha$. Therefore, using $1 - \alpha$ as the nominal level is our recommended choice based on this evidence.

Comment 4.1 (Robustness of the Coverage Property)**.** In our numerical results, the coverage property is satisfied even if only (R1) holds. This suggests that it may be possible to establish the coverage property under much weaker conditions. In particular, the coverage property holds without the concentration conditions (R2) and (R3) if

$$\mathrm{P}\left(|\mathrm{M}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) \mid \text{Data})| > z\right) \leqslant \mathrm{P}\left(|\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A)| > z\right) + \gamma'''_N. \tag{4.12}$$

Perhaps surprisingly, this property does hold in numerical experiments even when θ_A does not concentrate around θ^*; see, e.g. Figure 2. However, formally demonstrating this property proved difficult and remains an unresolved problem for future research.

4.5. **Other Issues: Stratified Splitting, Small Variation of Proxies.**

The idea of stratified sample splitting is to balance the proportions of treated and untreated units in both *a* and *m* samples so that the proportion of treated units is equal to the experiment's propensity scores across strata. This balance potentially improves the performance of the inferential algorithms. Stratified sampling formally requires us to replace the i.i.d. assumption with an i.n.i.d. assumption (independent but not identically distributed observations). The inference results continue to apply as long as the conditions (R1), (R2), (R3) hold. We conjecture that these conditions continue to be plausible under stratified splitting.

Another issue is that the analysis may generate proxy predictors S that have little variation, so we can think of them as "weak". This causes some target parameters to be weakly identified, e.g.,

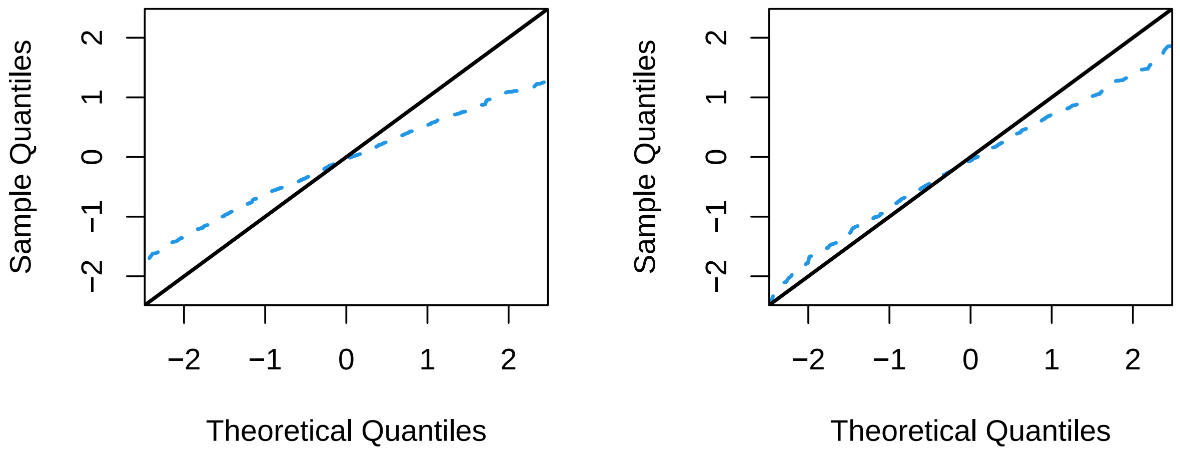


FIGURE 2. A simple Monte-Carlo experiment illustrating inferential robustness with and without concentration conditions.

NOTES. This example shows that the actual quantiles of the statistic $\mathrm{M}(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) \mid \text{Data})$ are conservatively bounded by those of $N(0,1)$ with concentration and without concentration. The estimand θ_A is generated from $U(0,K)$, with $K = 1/\sqrt{N}$ in the left panel (almost homogeneous case) or with $K = 10$ in the right panel (strong heterogeneous case). The estimator $\widehat{\theta}_A$ is generated as $\theta_A + 1/|M|\sum_{i\in M}\varepsilon_i$, where ε_i's are i.i.d. exponential random variables centered to have mean zero. The main sample indices M are randomly drawn from $\{1,...,N\}$ without replacement, with $N = 600$ and the subsample size $n/N = 1/3$. $\widehat{\sigma}_A$ is the classical standard error for the sample mean. In the left figure we get 99.5% coverage, and in the right 98.2% coverage for the nominal level of 95%. The results are based on 100 splits, and 1,000 replications.

the BLP parameter, leading to the potential breakdown of the basic normal approximation (4.5), which our inferential results rely on. To avoid this issue, we can add small noise to the proxies (jittering) so that inference results go through.

5. FURTHER CONSIDERATION: CAUSAL MACHINES THAT LEARN CATE BETTER

Our main proposal so far is to take proxies from any first stage black box machine and post-process them to better target CATE and perform inference on functionals of CATE, such as the BLP and GATES. But, can we design the machines to target CATE directly in the first stage? If we can, then the post-processing methods of the previous section would mostly focus on providing inference, and less on correcting biases of the first stage inputs. Building on Athey and Imbens (2016), we propose two types of such causal machines, and connect them to the emerging literature on orthogonal machine learning, such as Nie and Wager (2020), Semenova et al. (2017), and Foster and Syrgkanis (2019), among others.

5.1. **Focusing ML Methods on CATE in Stage 1.** We propose two options, taking ideas from our stage 2 analysis to stage 1. Specifically, we can train ML proxies in the auxiliary sample based on either:

(A) Minimizing $w(Z)$-weighted square prediction errors of Y using B and $(D-p(Z))S$;
(B) Minimizing square prediction errors of YH on BH and S;

where $B(Z)$ is now a technical "baseline" function of covariates Z, as described below, whose role is to reduce noise in the learning problem.

Definition 5.1 (Causal Learners for Stage 1)**.** *We can solve either of:*

$$(B,S) \in \arg\min_{b\in\mathscr{B}, s\in\mathscr{S}} \sum_{i\in A} w(Z_i)[Y_i - b(Z_i) - \{D_i - p(Z_i)\}s(Z_i)]^2, \tag{A}$$

$$(B,S) \in \arg\min_{b\in\mathscr{B}, s\in\mathscr{S}} \sum_{i\in A} [Y_iH_i - b(Z_i)H_i - s(Z_i)]^2, \tag{B}$$

where $w(Z) = [p(Z)(1-p(Z))]^{-1}$*, and* $\mathscr{B}$ *and* $\mathscr{S}$ *are functional parameter spaces.*

We can refer to the first causal learner as the weighted residual (WR) learner, and the second causal learner as the HT learner. Examples of parameter spaces include spaces of linear functions generated by a set of dictionary transformations of Z, reproducing kernels, linear combinations of decision trees, neural networks, and others. In (A) the parameter spaces are meant, but not required, to contain the functions $z \mapsto \tilde{b}_0(Z) := b_0(z) + p(z)s_0(z)$ and $z \mapsto s_0(Z)$. In (B) the parameter spaces are meant, but not required, to contain the functions $z \mapsto \bar{b}_0(Z) := b_0(Z) + (1-p(z))s_0(z)$ and $z \mapsto s_0(Z)$.

Both (A) and (B) improve over the standard predictive learners that predict Y using the best approximation to $\mathrm{E}[Y \mid D,Z]$ in a given class, but not necessarily the best approximation to the CATE $s_0(Z)$ itself, and may be of independent interest. Moreover, the loss functions in (A) and (B) are also helpful for validation purposes, and choosing the best or aggregating classes of ML methods for targeting the CATE function.

The proposal (B) generalizes and refines the strategy of Athey and Imbens (2016) of predicting YH using (a tree form of) S by introducing denoising by B. Semenova and Chernozhukov (2021) developed a related HT strategy that applies series/sieve learners to the denoised HT-transformed outcome, but it explicitly relies on consistent estimation of the regression function, unlike our approach. We further discuss connections of proposal (A) to the unweighted residual learners of Semenova et al. (2017), Nie and Wager (2020), and others below.[21]

[21] Both of our proposals appeared to be new around the first circulation of this paper as ArXiv:1712.04802. See links to the recent literature below, which proposed related, but different ideas. Relative to the initial version of the

Theorem 5.1 (Oracle Properties of the Population Objective Functions)**.** *Suppose that Y, $b(Z)$, $s(Z)$, and $w(Z)$ are square integrable. (1) Then, the expectations of the loss functions in* (A) *and* (B) *are*

$$\mathrm{E}w(Z)[Y-b(Z)-(D-p(Z))s(Z)]^2 = \mathrm{E}[s_0(Z)-s(Z)]^2+C_{1b}, \tag{5.1}$$

$$\mathrm{E}[YH-b(Z)H-s(Z)]^2 = \mathrm{E}[s_0(Z)-s(Z)]^2+C_{2b}, \tag{5.2}$$

where $C_{1b} := \mathrm{E}[w(Z)(\tilde{b}_0(Z)-b(Z))^2]+C_1$ and $C_{2b} := \mathrm{E}[w(Z)(\bar{b}_0(Z)-b(Z))^2]+C_2$ for some constants C_1 and C_2. (2) Therefore, the minimizers, say $s_\bullet(Z)$, of the left-hand sides of (5.1) and (5.2) over $s \in \mathscr{S}$, if exist, also minimize the oracle loss function $\mathrm{E}[s_0(Z)-s(Z)]^2$ *over the same set.*

Theorem 5.1 shows that the minimizers of the two loss functions provide the best approximation in the mean-square sense to the actual CATE function $s_0(Z)$ in the class $\mathscr{S}$. This property occurs even though we do not observe $s_0(Z)$, and such performance is usually qualified as "oracle." A sufficient condition for the existence of minimizers is that the set $\mathscr{S}$ be convex and closed in the $L^2(P)$ norm.

We illustrate the benefits of using the causal learning objectives (A) and (B) in Figure 3. In the two panels, we compare the CATE learners derived from the standard predictive Random Forest and Neural Network with the Causal Learners from Random Forests and Neural Network that solve the objective functions (A) and (B). We find that the causal learners are better at approximating the actual CATE function, thereby providing better proxies for CATE. The improvements in the RMSE of approximating the CATE provided by the causal learners range from 21% to 55%.

Likewise, the left panel of Figure 4 shows that we can improve the standard predictive OLS by the Causal OLS that solves the objective functions (A) and (B). Here we estimate linear models in Z for the baseline function and CATE. The improvement in the RMSE of approximating the CATE provided by causal OLS is about 35%. This finding might be of interest to researchers using OLS in empirical work.

Finally, the right panel of Figure 4 shows that one can improve the Causal Forest by a causal boosting step that solves the objective functions (A) and (B) by looking for a shallow forest deviation away from the cross-fitted Causal Forest proxy. The improvements in the RMSE of approximating the true CATE provided by this step are $54-63\%$. The explanation for this improvement is that the Causal Forest, while explicitly targeting CATE, actually solves a different objective than (A) or (B).[22] We also verified that this improvement only applies when the propensity score is not constant, like in this example.

paper, the current version contributes with several formal learning properties of (A) and (B). We are grateful to the referees for suggesting that we develop these properties.

[22] In our understanding it performs a residual learning approach, but not the weighted residual learning approach, which makes the method under-perform relative to the Forest Causal Learner based upon (A) or (B).

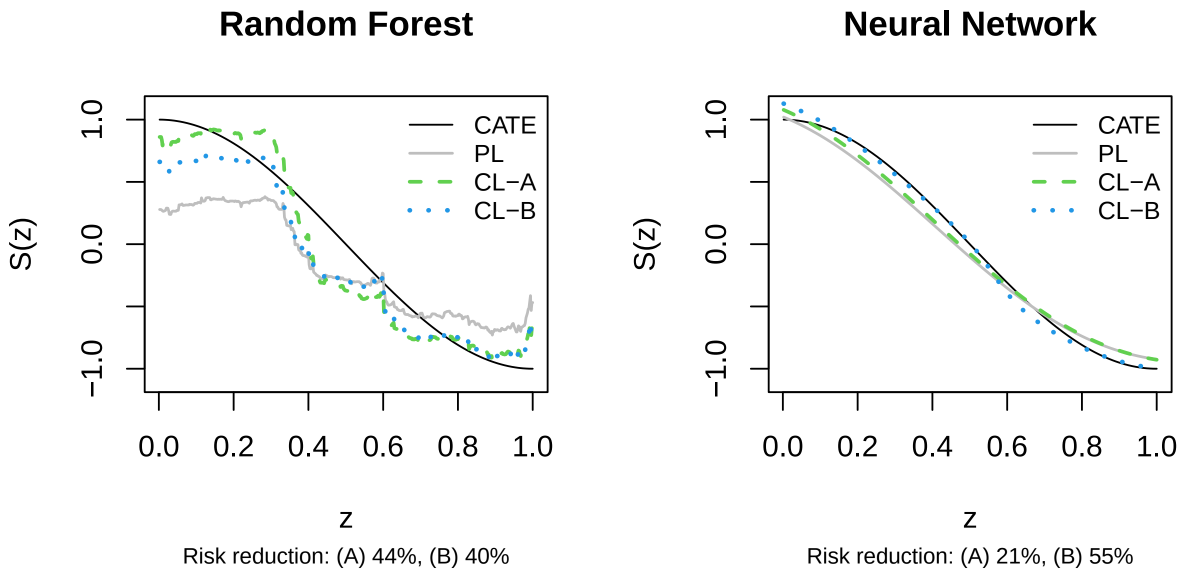


FIGURE 3. Comparison of Predictive Machine Learners vs Causal Learners Based on (A) and (B).

NOTES: The solid black curve is the CATE function $s_0(Z)$, and the solid grey (light) curve is the predictive learner $S(Z)$ (PL) obtained by random forest and neural network. The dashed green and dotted blue curves are estimators $S(Z)$ produced by the causal learners of CATE based on solving objectives (A) and (B) (CL-A and CL-B) . The underlying data is generated as $Y_i = b_0(Z_i) + s_0(Z_i)D_i + \xi_i$, where $\xi_i \sim N(0,1/4)$, $Z_i \sim U(0,1)$, D_i is Bernoulli with success probability $p(Z) = .1 \cdot 1\{Z < 1/2\} + .5 \cdot 1\{Z \geqslant 1/2\}$, $b_0(z) = z$, $s_0(z) = \cos(2\pi z)$, and $i = 1,..,500$.

Comment 5.1 (Connections to the Literature)**.** Residual learning like (A), but without weighting by $w(Z)$, appears in the debiased machine learning of Robinson's partially linear model in Chernozhukov et al. (2017).[23] In the nonparametric setting, Nie and Wager (2020) and Semenova et al. (2017) proposed the unweighted version of type (A) objection function around the same time as we did ours (all appeared in December of 2017 in ArXiv). Both of these papers target CATE learning in non-experimental settings. Their proposal does not use weighting by $w(Z)$ as ours and therefore does not provide the best approximation property to the CATE in population. However, it is easy to verify that their proposal provides the best approximation to CATE weighted by $p(Z)(1-p(Z))$:

$$\min_{s \in \mathscr{S}} \mathrm{E}[p(Z)(1-p(Z))(s_0(Z)-s(Z))^2].$$

[23] Also called "partialling out", residual, and orthogonal learning, building upon classical econometric ideas due to Frisch-Waugh-Lovell and Robinson (1988). Note Chernozhukov et al. (2017) presents other strategies as well, with residual learning being just one of them.

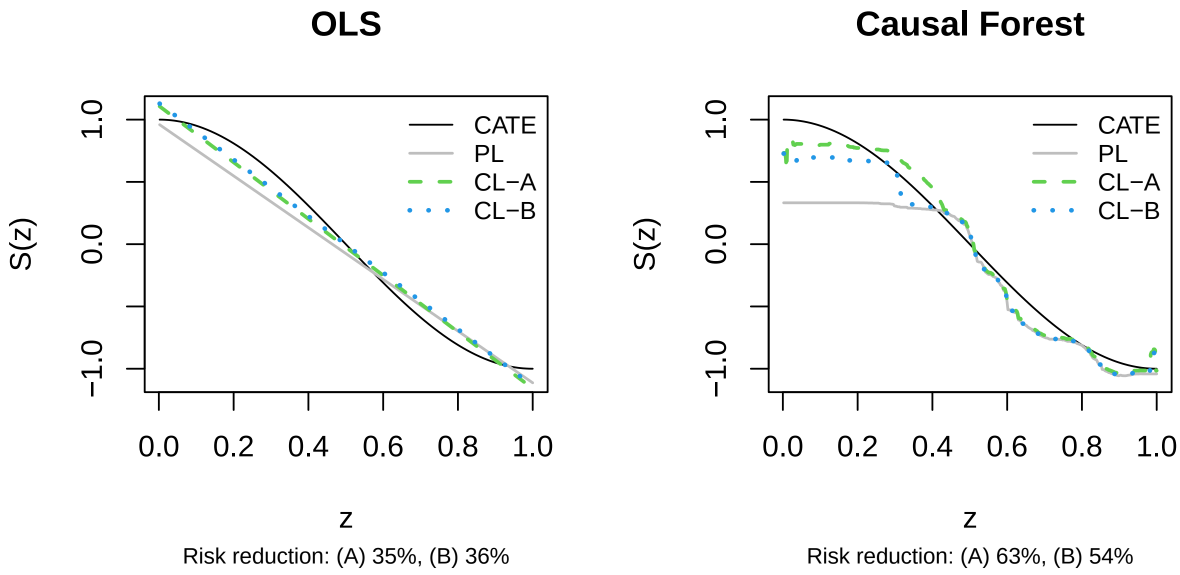


FIGURE 4. Comparison of OLS and Causal Forest as Predictive Learners vs OLS Causal Learners and Forest Causal Learners Based on (A) and (B).

NOTES: The solid black curve is the CATE function $s_0(Z)$, and the solid grey (light) curve is the predictive CATE learner $S(Z)$ (PL) obtained by OLS and Causal Random Forest. The dashed green and dotted blue curves are estimators $S(Z)$ produced by the causal learners of CATE based on solving objectives (A) and (B) (CL-A and CL-B) . The underlying data is generated in the same way as in fig. 3.

In contrast to our proposal, this weighting gives relatively less attention to units that are either more likely or less likely to be treated than units that are equally likely. While this property is not appealing in general, in pure RCTs with constant propensity score $p(Z)$, the objective function above reduces to the best approximation to CATE. Moreover, weighting by $w(Z)$ in (A) plays less important role when s_0 can be estimated consistently, as in Nie and Wager (2020), Semenova et al. (2017), and Foster and Syrgkanis (2019). We are interested, however, in the high-dimensional settings in which consistent learning of s_0 might not be possible. Therefore, below we provide formal estimation results for causal learners (A) and (B) under this "agnostic" setting. Our inference methods of Section 4 also apply to the causal learners (A) and (B) used as first-stage proxies.

5.2. **Learning Guarantees.** The learning guarantees of the causal learners (A) and (B) follow from the state-of-art statistical learning theory (Liang, Rakhlin, and Sridharan, 2015), in particular through the use of the expected off-set Rademacher complexity (ORC).

Liang et al. (2015) define the expected ORC of the function class $\mathscr{H}$ as:

$$\mathscr{R}^o(A,\mathscr{H},c) := \mathrm{E} \sup_{h\in\mathscr{H}} \frac{1}{|A|}\sum_{i\in A}\left[e_i h(Z_i) - c h(Z_i)^2\right],$$

where $\{e_i\}$ are i.i.d. Rademacher variables taking values -1 and 1 with probability $1/2$, that are generated independently of the data $\{Z_i\}_{i\in A}$, and $c > 0$ is a positive constant. ORC is a statistical measure of complexity that captures the ability of the functional class to fit i.i.d. Rademacher noise. The more complex the class is, the higher the ORC. As shown in Liang et al. (2015), the expected ORC naturally scales like $d/|A|$, where d is the effective dimension of the function class and $|A|$ is the sample size. For example, for linear classes, d is the actual dimension of the linear class; and for VC classes, d scales like the VC dimension. Liang et al. (2015) show that ORC is the sharpest characterization of complexity: in particular, the ORC is upper bounded by the standard critical radii of function classes defined in terms of local Rademacher complexity (e.g., Wainwright, 2019) or in terms of the standard uniform covering entropy (Dudley, 2000). This makes the ORC bounds readily available for all function classes used in modern ML.

The following result establishes formal learning guarantees for causal learners in randomized experiments under general settings that do not assume we can learn s_0 consistently.

Theorem 5.2 (Near-Oracle Guarantees for Causal Learners)**.** *Suppose that Y, the elements of $\mathscr{B}$ and $\mathscr{S}$, and $w(Z)$ are bounded in absolute values by K, and $\mathscr{B}$ and $\mathscr{S}$ are closed, convex, and symmetric sets. The estimator S obtained as a solution of either* (A) *or* (B) *is as good as using the best in class approximation, say $s_\bullet(Z)$, up to an error expressed in terms of ORC:*

$$0 \leqslant \mathrm{E}[s_\bullet(Z) - S(Z)]^2 \leqslant \underbrace{\mathrm{E}[s_0(Z) - S(Z)]^2 - \overbrace{\mathrm{E}[s_0(Z) - s_\bullet(Z)]^2}^{\text{oracle risk}}}_{\text{excess risk}} \leqslant C_K \mathscr{R}^o(A,\mathscr{H},c_K), \tag{5.3}$$

where C_K and c_K are positive constants that only depend on K, $\mathscr{H} := 4(w(Z)^2\mathscr{B} + Hw(Z)\mathscr{S})$ for type (A) loss, and $\mathscr{H} := 4(H\mathscr{B} + \mathscr{S})$ for type (B) loss.

The result shows that if the ORC of the functional parameter spaces is small, the excess risk of this estimator relative to the oracle approximation to the CATE is small. Note that the lower bound also bounds the distance of S to the oracle (best-in-class) $s_\bullet$ predictor of the CATE. Therefore, the bounds on the excess risk and distance readily follow from the existing characterization of the ORC. For example, if $\mathscr{H}$ has VC type covering entropy with VC index d, then the ORC is of order $d/|A|$.

Since the "base" functions B's play only a noise-reducing rule, we can always select $\mathscr{B}$ to be no more complex than $\mathscr{S}$. For example, we can use $\mathscr{B} \subseteq \mathscr{S}$ or pre-train B using a separate auxiliary sample, in which case $\mathscr{B}$ is a singleton.[24] In either case, learning the technical baseline function

[24] This also applies to cross-fitting, which is a better form of sample-splitting for practice.

does not affect the rate of learning the oracle prediction $s_{\bullet}$; and the ORC is determined solely by the complexity of $\mathscr{S}$.

Comment 5.2 (Extensions of Theorem 5.2)**.** The result of Theorem 5.2 follows from combining Theorem 3 of Liang et al. (2015) with Theorem 5.1. We assumed boundedness conditions to make the statement as simple as possible. Bounds on excess risk without the boundedness conditions follow from Theorem 4 in Liang, Rakhlin, and Sridharan (2015). If the class $\mathscr{S}$ is not convex, similar performance bound is attained by Audibert (2007)'s "star" modification of the optimizer S, by Theorems 3 and 4 by Liang et al. (2015). We refer to Liang et al. (2015) and Vijaykumar (2021) for detailed general discussion.

5.3. **Using Losses** (A) **and** (B) **for Choosing the Best ML Method.** The loss functions (A) and (B) can also be used to aggregate several learning methods using separate auxiliary subsets.[25] To fix ideas, suppose we have a set of methods giving scores S_k and B_k, $k = 1,...,K$, where K is small, obtained using a subset $A_1 \subset A$. Then, we can combine these scores into

$$S(Z) = \sum_{k=1}^{K} \lambda_k^S S_k(Z); \;\; B(Z) = \sum_{k=1}^{K} \lambda_k^B B_k(Z),$$

and then we can learn the weights λ^S and λ^B by optimizing the loss functions (A) or (B) evaluated on subset A_2, such that A_2 does not overlap with A_1, e.g. $A_2 = A \setminus A_1$.

Comment 5.3 (Learning Guarantee for Aggregation)**.** Let $\widehat{\lambda}^S$ and $\widehat{\lambda}^B$ denote the weights learned in this way. Then, another application of the results of Liang et al. (2015) for linear regression, under the assumption that $|Y|$, $|B|$, $|S|$, $|w(Z)|$ are all bounded by R, gives the excess risk bound:

$$\mathrm{E}\Big[s_0(Z) - \sum_{k=1}^{K} \widehat{\lambda}_k^S S_k(Z)\Big]^2 - \overbrace{\min_{\{\lambda_k^S\}_{k=1}^K} \mathrm{E}\Big[s_0(Z) - \sum_{k=1}^{K} \lambda_k^S S_k(Z)\Big]^2}^{\text{oracle risk}} \leqslant C_R K / |A_2|,$$

where C_R is some constant that depends on R and $|A_2|$ is the sample size used to perform the aggregation. Thus, if the right-hand side is small, the excess risk of this estimator relative to the oracle aggregation method is negligible. Since the oracle aggregation risk here is weakly smaller than the oracle risk of choosing the best prediction rule $\min_k \mathrm{E}[s_0(Z) - S_k(Z)]^2$, convex aggregation here is approximately better than choosing the best ML method.

Comment 5.4 (Large K)**.** The method above gives a small excess risk when $K/|A_2|$ is small; otherwise, the excess risk can be large. In the latter case, we can apply Lasso to select a sparse linear combination of rules, and the sharp bounds on excess risk follow from Example 4 in Koltchinskii et al. (2011). Finally, we may choose the "best" machine learning algorithm using objective

[25] This is in contrast to our main proposal, where we choose the best ML method based on goodness-of-fit measures in the second stage.

functions (A) and (B) evaluated on the data subset A_2. Results of Wegkamp et al. (2003) imply certain good guarantees for the ”best” approach, but sharp bounds on the excess risk that scale like $\log K/|A_2|$ only hold for the ”star” modification of the ”best” method (Audibert, 2007).

6. Application: Where are nudges for immunization the most effective?

We apply our methods to an RCT in India that was conducted to improve immunization and provide detailed implementation algorithms. Our main specification reports median intervals (MI) from causal learners via Boosting as described in Algorithm 6.2. We also estimate results using predictive ML methods and report them in the Online Appendix C as they perform worse than the causal learners. Finally, inferential results are robust to using prediction intervals for the majority values (MP, reported in Online Appendix D). For the sample splitting, we allocate 1/3 of the sample to the main sample.[26]

Immunization is widely believed to be one of the most cost-effective ways to save children’s lives. Much progress has been made in increasing immunization coverage since the 1990s. For example, according to the World Health Organization (WHO), global measles deaths have decreased by 73% from 536,000 estimated deaths in 2000 to 142,000 in 2018. In the last few years, however, global vaccination coverage has remained stuck at around 85% (until the COVID-19 epidemics, when they plummeted). In 2018, 19.7 million children under the age of one year did not receive basic vaccines. Around 60% of these children lived in ten countries: Angola, Brazil, the Democratic Republic of the Congo, Ethiopia, India, Indonesia, Nigeria, Pakistan, the Philippines, and Vietnam. The WHO estimates that immunization saves 2-3 million deaths every year and that an additional 1.5 million deaths could be averted every year if global vaccination coverage improves (this is comparable to 689,000 deaths from COVID-19 between January and August 2020).[27]

While most of the early efforts have been devoted to building an immunization infrastructure and ensuring that immunization is available close to people’s homes, there is a growing recognition that it is important to also address the demand for immunization. Part of the low demand reflects deep-seated mistrust, but in many cases, parents seem to be perfectly willing to immunize their children. For example, in our data for Haryana, India, among the sample’s older siblings who should all have completed their immunization course, 99% had received polio drops, and about 90% had an immunization card. 90% of the parents claimed to believe immunization is beneficial, and 3% claimed to believe it is harmful. However, only 37% of the older children had completed the course and received the measles vaccine, according to their parents (which is likely to be an overestimate), and only 19.4% had done so before the fifteen month of life, when it is supposed

[26] We find similar results using 1/2 splits. These results are available upon request.

[27] See WHO “10 facts on immunization”, `https://www.who.int/features/factfiles/immunization/facts/en/index1.html`

to be done between the 10th and the 12th month. It seems that parents lose steam over the course of the immunization sequence, and nudges could be helpful to boost demand. Indeed, recent literature cited in the introduction suggests that "nudges," such as small incentives, leveraging the social network, SMS, etc., may have a large effect on those services.

In 2017, Esther Duflo, one of the authors of this paper, led a team that conducted a large-scale experiment with the government of Haryana in North India to test various strategies to increase the takeup of immunization services. The government health system rolled out an e-health platform designed by a research team and programmed by an MIT group (SANA health), in which nurses collected data on which child was given which shot at each immunization camp. The platform was implemented in over 2,000 villages in seven districts and provides excellent administrative data on immunization coverage.[28] From the individual data, we constructed the monthly sum of the number of children eligible for the program (i.e., age 12 months or younger at their first vaccines) who received each particular immunization at a program location. These children were aged between 0 and 15 months. This paper focuses on the number of children who received the measles shot, as it is the last vaccine in the sequence and thus a reliable marker for full immunization.

Before the launch of the interventions, survey data were collected in 912 of those villages using a sample of 15 households with children aged 1-3 per village. The baseline data covers demographic and socio-economic variables and the immunization history of these children, who were too old to be included in the intervention. In these 912 villages, three different interventions (and their variants) were cross-randomized at the village level:

(1) Small incentives for immunization: parents/caregivers receive mobile phone credit upon bringing children for vaccinations.
(2) Immunization ambassador intervention: information about immunization camps was diffused through key members of a social network.
(3) Reminders: a fraction of parents/caregivers who had come at least one time received SMS reminders for pending vaccinations of the children.

For each of these interventions, there were several possible variants: incentives were either low or high and either flat or increasing with each shot; the immunization ambassadors were either randomly selected or chosen to be information hubs, using the "gossip" methodology developed by Banerjee et al. (2021), a trusted person, or both; and reminders were sent to either 33% or 66% of the people concerned. Moreover, each intervention was cross-cut, generating 75 possible treatment combinations.

Banerjee et al. (2021) developed and implemented a two-step methodology to identify the most cost-effective and the most effective policy to increase the number of children completing the full course of immunization at the village level and estimate its effects (correcting for bias due to

[28] Banerjee et al. (2019) discuss validation data from random checks conducted by independent surveyors.

TABLE 1. Selected Descriptive Statistics of Villages

	All	Treated	Control
Outcome Variables *(Village-Month Level)*			
Number of children who completed the immunization schedule	8.234	10.071	7.304
Baseline Covariates–Demographic Variables *(Village Level)*			
Household financial status (on 1-10 scale)	3.479	3.17	3.627
Fraction Scheduled Caste-Scheduled Tribe (SC/ST)	0.191	0.199	0.188
Fraction Other Backward Caste (OBC)	0.222	0.207	0.23
Fraction Hindu	0.911	0.851	0.939
Fraction Muslim	0.059	0.109	0.035
Fraction Christian	0.001	0.003	0
Fraction Literate	0.797	0.786	0.802
Fraction Single	0.053	0.052	0.053
Fraction Married (living with spouse)	0.517	0.499	0.526
Fraction Married (not living with spouse)	0.003	0.003	0.003
Fraction Divorced or Separated	0.002	0.005	0
Fraction Widow or Widower	0.04	0.037	0.041
Fraction who received Nursery level educ. or less	0.152	0.154	0.151
Fraction who received Class 4 level educ.	0.081	0.08	0.082
Fraction who received Class 9 educ.	0.157	0.162	0.154
Fraction who received Class 12 educ.	0.246	0.223	0.257
Fraction who received Graduate or Other Diploma	0.085	0.078	0.088
Baseline Covariates–Immunization History of Older Cohort *(Village Level)*			
Number of vaccines administered to pregnant mother	2.276	2.211	2.307
Number of vaccines administered to child since birth	4.485	4.398	4.527
Fraction of children who received polio drops	0.999	1	0.999
Number of polio drops administered to child	2.982	2.985	2.98
Fraction of children who received an immunization card	0.913	0.871	0.933
Fraction of kids who received Measles vaccine by 15 months of age	0.194	0.175	0.203
Fraction of kids who received Measles vaccine at credible locations	0.386	0.368	0.395
Number of Observations			
Villages	103	25	78
Village-Months	844	204	640

the fact that the policy is found to be the best). First, they used a specific version of LASSO to determine which policies are irrelevant and which policy variants can be pooled together. Second, they obtained consistent estimates of this restricted set of pooled policies using post-LASSO (Chernozhukov et al., 2015). They found that the most cost-effective policy (and the only one to

TABLE 2. Comparison of Causal ML Methods: Immunization Incentives

	Elastic Net	Boosting	Neural Network	Random Forest
Best BLP (Λ)	67.750	32.900	53.420	25.200
	[51.491, 82.368]	[23.246, 44.665]	[42.516, 67.647]	[18.328, 34.705]
Best GATES ($\bar{\Lambda}$)	8.254	5.104	6.001	4.492
	[7.329, 9.314]	[4.27, 6.079]	[5.087, 6.888]	[3.339, 5.507]

Notes: Medians over 250 splits. Note that we used Neural Network Causal Boosting for all methods, using Algorithm 6.2. The brackets report interquartile ranges for goodness-of-fit statistics.

reduce the cost of each immunization compared to the status quo) is to combine information hub ambassadors (trusted or not) and SMS reminders. But the policy that increases immunization the most is the combination of information-hub ambassador, the presence of reminders, and increasing incentives (regardless of levels). This is also the most expensive package, so the government was interested in prioritizing villages: where should they scale up the full package? This is an excellent application of this methodology because there was no strong prior.

We compare 25 treated villages where this particular policy bundle was implemented with 78 control villages that received neither sloped incentives and social network intervention nor reminder. Our data constitute an approximately balanced monthly panel of the 103 treated and control villages for 12 months (the duration of the intervention). The outcome variable, Y, is the number of children 15 months or younger in a given month in a given village who receive the measles shot. The treatment variable, D, is an indicator of the household being in a village that receives the policy. The covariates, Z, include 36 baseline village-level characteristics such as religion, caste, financial status, marriage and family status, education, and baseline immunization. The propensity score is constant.

Table 1 shows sample averages in the control and treated groups for some of the variables used in the analysis weighted by village population, as the rest of the analysis. Treatment and control villages have similar baseline characteristics (in particular, the immunization status of the older cohort was similar). The combined treatment was very effective on average. During the course of the intervention, on average, 7.30 children per month aged 15 months or less got the measles shot that completes the immunization sequence in control villages, and 10.08 did so in treatment villages. This is a raw difference of 2.77 or 38% of the control mean. Note that while these effects are not insignificant, we are far from reaching full immunization: The baseline survey suggests that about 38% of children aged 1-3 had received the measles shot at baseline, and 19.4% had received it before they turned 15 months. These estimates imply that the fraction getting their measles shot before 15 months would only go up to $26.7\%(19.4+0.38*19.4)$.

TABLE 3. BLP of Immunization Incentives Using Causal Proxies

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.814	1.047	2.441	0.899
(1.087,4.506)	(0.826,1.262)	(0.846,3.979)	(0.685,1.107)
[0.004]	[0.000]	[0.004]	[0.000]

Notes: Medians over 250 splits. Median Confidence Intervals ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternative in brackets.

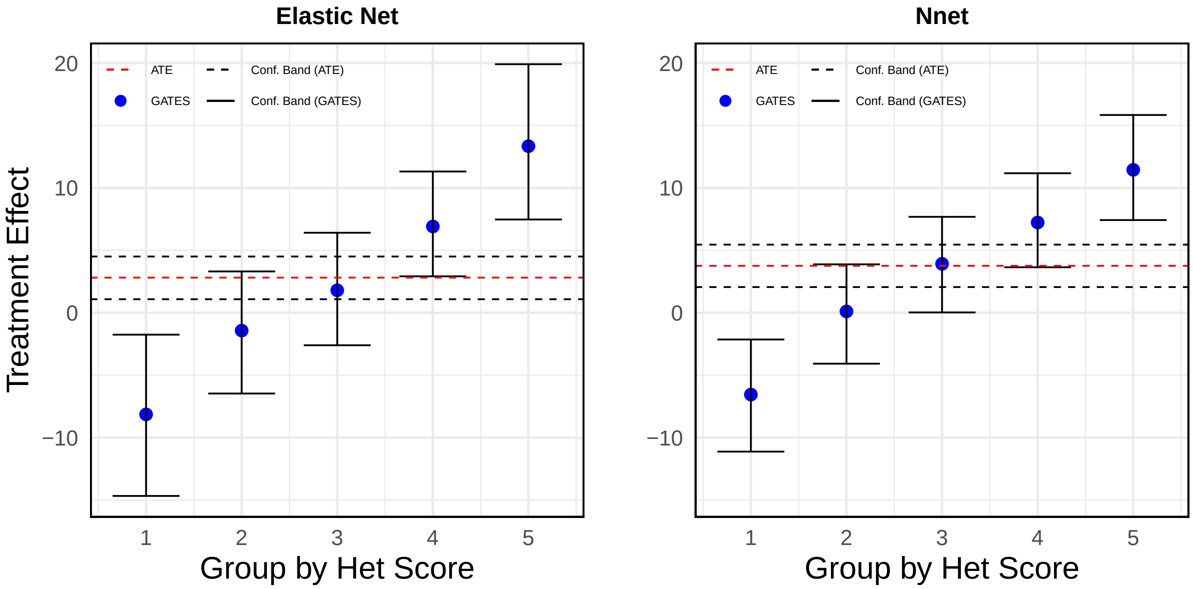
The implementation details for the heterogeneity analysis follow Algorithm 6.1 below, with three characteristics due to the design: we weight village-level estimations by village population, include district–time fixed effects, and cluster standard errors at the village level. Table 2 compares the four ML methods for producing proxy predictors $S(Z_i)$ using the criteria in (3.11) and (3.12). We find that Elastic Net and Neural Network outperform the other methods, with Elastic Net beating Neural Network by a smaller margin than the other methods. Accordingly, we shall focus on these two methods for the rest of the analysis.

Table 3 presents the results of the BLP of CATE on the ML proxies. We report estimates of the coefficients β_1 and β_2, which correspond to the ATE and heterogeneity loading (HET) parameters in the BLP. The ATE estimates in Columns 1 and 3 indicate that the package treatment increases the number of immunized children by 2.81 based on elastic net estimates and by 2.44 based on neural network estimates. Reassuringly, these estimates are on either side of the raw difference in means (2.77). Focusing on the HET estimates, we find strong heterogeneity in treatment effects, as indicated by the statistically significant estimates. Moreover, the estimates are close to 1, suggesting that the ML proxies are good predictors of the CATE.

Next, we estimate the GATES by quintiles of the ML proxies. Figure 5 presents the estimated GATES coefficients $\gamma_1 - \gamma_5$ along with joint confidence bands and the ATE estimates. In Table 4 we present the result from the hypothesis test that the difference of the ATE for the most and least affected groups is statistically significant. We find that this difference is 21.60 and 18.13 based on elastic net and neural network methods, respectively, and is statistically significant. Given that the ATE estimates in the whole population are about 2.5, these results suggest a large and potentially policy-relevant heterogeneity.

The analysis so far reveals very large heterogeneity, with two striking results. First, the results are very large for the most affected villages. In these villages, an average of 13.23 extra children eligible for baseline incentives get the measles vaccines every month (starting from a mean of 2.19

FIGURE 5. GATES of Immunization Incentives



Notes: GATES of Immunization Incentives, based upon Causal Learners. Median point estimates and Median confidence interval ($\alpha = .05$) in parenthesis, over 250 splits.

TABLE 4. GATES of 20% Most and Least Affected Groups

	Elastic Net			Nnet		
	20% Most (G_5)	20% Least (G_1)	Difference	20% Most (G_5)	20% Least (G_1)	Difference
GATE	13.230	-8.000	21.60	11.210	-6.551	18.13
$\gamma_k := \widehat{\mathrm{E}}[s_0(Z) \mid G_k]$	(8.219,18.67)	(-13.41,-2.574)	(13.70,29.74)	(7.721,14.47)	(-10.37,-2.786)	(12.84,23.52)
	[0.000]	[0.009]	[0.000]	[0.000]	[0.002]	[0.000]
Control Mean	2.19	12.68	-10.56	1.19	10.32	-9.17
$:= \widehat{\mathrm{E}}[b_0(Z) \mid G_k]$	(1.27,3.06)	(11.73,13.59)	(-11.84,-9.38)	(0.44,1.87)	(9.65,11.02)	(-10.17,-8.14)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]

Notes: Medians over 250 splits. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternative in brackets.

in the elastic net estimation). Second, the impact is *negative* and significant in the least affected villages (an average decline of 8.00 immunization per month, starting from 12.68 in the elastic net

TABLE 5. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Number of vaccines to pregnant mother	2.187 (2.115,2.259) -	2.277 (2.212,2.342) -	-0.081 (-0.180,0.015) [0.190]	2.174 (2.111,2.234) -	2.285 (2.224,2.345) -	-0.112 (-0.202,-0.028) [0.019]
Number of vaccines to child since birth	4.077 (3.858,4.304) -	4.639 (4.444,4.859) -	-0.562 (-0.863,-0.260) [0.001]	4.264 (4.091,4.434) -	4.734 (4.549,4.900) -	-0.490 (-0.739,-0.250) [0.000]
Fraction of children received polio drops	0.998 (0.995,1.001) -	1.000 (0.997,1.003) -	-0.002 (-0.006,0.002) [0.683]	1.000 (1.000,1.000) -	1.000 (1.000,1.000) -	0.000 (0.000,0.000) [0.943]
Number of polio drops to child	2.955 (2.935,2.974) -	2.993 (2.976,3.010) -	-0.037 (-0.063,-0.010) [0.013]	2.965 (2.953,2.977) -	2.998 (2.985,3.010) -	-0.032 (-0.049,-0.016) [0.000]
Fraction of children received immunization card	0.803 (0.754,0.851) -	0.926 (0.882,0.969) -	-0.121 (-0.187,-0.054) [0.001]	0.908 (0.881,0.932) -	0.927 (0.898,0.959) -	-0.027 (-0.059,0.007) [0.217]
Fraction of children received Measles vaccine by 15 months of age	0.133 (0.097,0.169) -	0.243 (0.209,0.276) -	-0.106 (-0.153,-0.056) [0.000]	0.126 (0.095,0.159) -	0.260 (0.228,0.291) -	-0.131 (-0.176,-0.085) [0.000]
Fraction of children received Measles vaccine at credible locations	0.293 (0.246,0.338) -	0.399 (0.358,0.444) -	-0.110 (-0.174,-0.045) [0.002]	0.289 (0.246,0.331) -	0.433 (0.391,0.475) -	-0.142 (-0.206,-0.084) [0.000]

Notes: Medians over 250 splits. Median confidence interval ($\alpha = .05$) in parenthesis. P-values for the hypothesis that the parameter is equal to zero against the two-sided alternatives in brackets.

estimation). It looks like in some contexts, the combined package of small incentives, reminders, and persuasion by members of the social network put people off immunization.

Given these large differences, it is important to determine whether this heterogeneity seems to be associated with pre-existing characteristics. To answer this question, we ask what variables are associated with the heterogeneity detected in BLP and GATES via CLAN. Table 5 reports the CLAN estimates for a selected set of covariates and Tables 1–2 in Online Appendix B for the rest of covariates. Regardless of the method used, the estimated differences in the means of most and least affected groups for the number of vaccines to child since birth, number of polio drops to child, the fraction of children receiving measles vaccines by 15 months of age, and fraction of children receiving measles vaccine at credible locations, are negative and statistically significant. Those are various measures of pretreatment immunization levels, all survey-based, that have nothing to do with our measure of impact. These results suggest that the villages with low levels of pretreatment immunization are the most affected by the incentives. These are, in fact, the only variables that consistently pop up from the CLAN. Thus, in this instance, the policy preferred ex-ante by the government (since it is equality-enhancing) also happens to be the most effective.

While the heterogeneity associated with the baseline immunization rates cannot be causally interpreted (it could always be proxying for other things), it still sheds light on the negative effect we find for the least affected group. Note that this effect is *not* mechanical. Even in the least affected villages, there was a good number of children who did not receive the measles shot, and they were not close to reaching full immunization, where they could not have experienced an increase. It may be that it would have been difficult to vaccinate 13.23 extra children every month, but there was scope to experience an increase in immunization. We had no prior on that the effect would be larger in the villages with the lowest immunization rate. On the contrary, immunization rates could have been low precisely because parents were more doubtful about immunization. For example, immunization is particularly low in Muslim-majority villages, which is believed to reflect their lack of trust in the health system. There were, therefore, reasons to be genuinely uncertain about where immunization would have had the largest effect.

One possible interpretation of the negative impact in some villages is that villagers were intrinsically motivated to get immunized. The nudging with small incentives and mild social pressure may have backfired, crowding out intrinsic motivation without providing a strong enough extrinsic motivation to act as in Gneezy and Rustichini (2000). A point estimate of 13.23 extra immunization per month in the most affected group might seem high: a multiplication by 6.0 of the baseline level (based on the elastic net specification). This increase in immunizations is not inconsistent with the literature: in a set of villages with a very low immunization rate in Rajasthan, Banerjee et al. (2010) find that small incentives increase immunization from 18% to 39% (relative to a treatment that just improves infrastructures, and 6% relative to the control group) in a low immunization region (in the entire sample, not in the places where it is most effective), which was also a very large increase. Given the restrictions imposed on the data set (only children 1 year or less at their first immunization were included), the data cover children who were 15 months or younger when getting the measles shot. Among the older cohort in the most affected group, only 12.7% of children were vaccinated before 15 months. Taking this as a benchmark for the control group, the estimate would still imply that only 64% of the treatment group was immunized before 15 months: a big improvement but not implausible.

Our last exercise is to compute the cost-effectiveness of the program in various groups. To do so, we compute in each village the average number of immunizations delivered per dollar spent in a month in each group. The dollar spent is the fixed cost to run an immunization program per month (nurse salaries, administrative overheads, etc.) plus the marginal cost of each vaccine multiplied by the number of vaccines administered (incentives distributed to local health workers, vaccines doses, syringes, etc.) in both treatment and control villages, plus the extra cost of running each particular treatment (the cost of the tablets used for recording in all the treatment villages, the cost of contacting and enrolling the ambassadors, and the cost of the incentives). We then estimate the cost-effectiveness in each GATES group as $\mathrm{E}[X(1)-X(0) \mid G_k]$, where X is the immunizations per

TABLE 6. Cost Effectiveness in GATE quintiles

	Elastic Net			Nnet		
	Mean in Treatment	Mean in Control	Difference	Mean in Treatment	Mean in Control	Difference
	$(\widehat{\mathrm{E}}[X \mid D=1, G_k])$	$(\widehat{\mathrm{E}}[X \mid D=0, G_k])$		$(\widehat{\mathrm{E}}[X \mid D=1, G_k])$	$(\widehat{\mathrm{E}}[X \mid D=0, G_k])$	
Imm. per dollar (G_1)	0.034	0.047	-0.013	0.033	0.047	-0.014
	(0.030,0.037)	(0.045,0.048)	(-0.017,-0.009)	(0.029,0.036)	(0.045,0.049)	(-0.018,-0.010)
	-	-	[0.000]	-	-	[0.000]
Imm. per dollar (G_2)	0.031	0.044	-0.013	0.035	0.044	-0.009
	(0.027,0.036)	(0.042,0.046)	(-0.018,-0.008)	(0.031,0.039)	(0.042,0.046)	(-0.014,-0.005)
	-	-	[0.000]	-	-	[0.000]
Imm. per dollar (G_3)	0.037	0.043	-0.007	0.037	0.043	-0.006
	(0.033,0.041)	(0.041,0.046)	(-0.011,-0.002)	(0.034,0.041)	(0.041,0.045)	(-0.010,-0.001)
	-	-	[0.015]	-	-	[0.022]
Imm. per dollar (G_4)	0.039	0.039	-0.001	0.038	0.041	-0.004
	(0.036,0.042)	(0.036,0.042)	(-0.005,0.004)	(0.034,0.041)	(0.038,0.044)	(-0.008,0.000)
	-	-	[1.000]	-	-	[0.163]
Imm. per dollar (G_5)	0.036	0.035	0.001	0.035	0.034	0.001
	(0.031,0.041)	(0.030,0.040)	(-0.006,0.008)	(0.031,0.040)	(0.029,0.040)	(-0.006,0.008)
	-	-	[1.000]	-	-	[1.000]

Notes: Medians over 250 splits. Median confidence interval $(\alpha = .05)$ in parenthesis. P-values for the hypothesis that the parameter is equal to zero against two-sided alternative in brackets.

dollar. $\mathrm{E}[X(1)-X(0) \mid G_k] = \mathrm{E}[X \mid D=1, G_k] - \mathrm{E}[X \mid D=0, G_k]$ by the randomization assumption, and we can estimate each of $\mathrm{E}[X \mid D=1, G_k]$ and $\mathrm{E}[X \mid D=0, G_k]$ analogously to CLAN, that is by taking sample averages within treatment groups for each sample split and aggregating over sample splits.

The results are presented in Table 6. They highlight the crucial importance of treatment effect heterogeneity for policy decisions in this context. Overall, as shown in Banerjee et al. (2021) the treatment is not cost-effective compared to the control (the immunization per dollar spent goes down). This analysis reveals that this is driven (not surprisingly) by negative impacts on cost-effectiveness in the groups where it is least effective. However, in the fourth and fifth quintile of cost-effectiveness, we cannot reject that the immunization per dollar spent is the same in the control group and in the treatment group, despite the added marginal cost of the incentives and the vaccines: this is because the fixed cost of running the program is now spread over a larger number of immunizations.

We also performed the following additional analysis. Since the main result in Banerjee et al. (2021) is that the most cost-effective option on average is the combination of SMS plus Information hubs, an alternative policy question may therefore be whether there are places where it may be more cost-effective to add the incentives to this cheaper treatment. We replicated the heterogeneity analysis comparing these two treatments and looked at the cost-effectiveness of GATES in these two options. There is also considerable heterogeneity in this comparison (see Online Appendix Figure 1). The results for cost-effectiveness are shown in Table 3 in the Online Appendix. There, again, we find that in the two quintiles where adding incentives is most effective, it would need to be cost-effective, even compared to an alternative status quo of just having SMS and information hubs.

6.1. **Implementation.** We describe two general algorithms and provide some specific implementation details for the empirical example.

Algorithm 6.1 (**Inference Algorithm**). *The inputs are given by the data* $\{(Y_i, D_i, Z_i, p(Z_i)\}$ *on units* $i \in [N] = \{1, ..., N\}$. *Fix the number of splits* N_S *and the significance level* α, *e.g.* $N_S = 250$ *and* $\alpha = 0.05$. *Fix a set of ML or Causal ML methods.*

1. Generate N_S *random splits of* $[N]$ *into the main sample, M, and the auxiliary sample, A. Over each split apply the following steps:*

- a. *Using A, train each ML method and output predictions B and S for M.*
- b. *Optionally, choose the best or aggregate ML methods using the results of Section 5.*
- c. *Estimate the BLP parameters via WR BLP* (3.3) *or HT BLP* (3.5) *in M.*
- d. *Estimate the GATES parameters by WR GATEs* (3.7) *or HT GATEs* (3.9) *in M.*
- e. *Estimate the CLAN parameters by taking averages (3.10) in M.*
- f. *Compute the goodness of fit measures via (3.13) in M.*

2. If the winning ML methods were not chosen in Step 1b, median-aggregate the goodness-of-fit measures and choose the best ML methods.

3. Compute and report the quantile-aggregated point estimates, p-values, and confidence intervals of Section 4. If Step 2 is used, compute and report the union of these statistics for all winners.

Comment 6.1. (Choices) We choose N_S sufficiently large to get enough representative values of the estimates of the target parameter values. In our experience, 250 splits are more than sufficient to obtain stable results in the sense that the point and interval estimates are not sensitive to increasing the number of splits in the empirical application. We followed a version of the algorithm with Step 2 and without Step 1a. Note that it is also possible to choose the best methods using a hold out sample using either the loss functions of Section 5 or the goodness-of-fit measures of Section 3. More research is needed to determine better practice for choosing the best ML methods.

We implemented our causal learners via a boosting approach that looks for relatively simple deviations from the initial predictive learner to improve CATE predictions. The reason is that this approach performed better in our simulation experiments than directly solving the objective functions (A) and (B) over large parameter spaces. We also observed that this boosting implementation performed better in the empirical example. The performance improvement occurs because the objective functions (A) or (B) tend to be much noisier than the objective functions in predictive learning and harder to tune. The following algorithm summarizes the implementation.

Algorithm 6.2 (Causal Learner via Boosting). *The inputs are given by data $\{Y_i, D_i, Z_i, p(Z_i)\}$ on units $i \in A \subset \{1,...,N\}$. Fix a predictive learner. Fix deviation (or boosting) parameter spaces: $\mathscr{B}_\Delta$ and $\mathscr{S}_\Delta$ that contain functions $z \mapsto b_\Delta(z)$ and $z \mapsto s_\Delta(z)$, mapping the support of Z to the real line.*

1. *Train the predictive learner on the input, and output the base proxy function $z \mapsto B_A(z)$ and CATE proxy function $z \mapsto S_A(z)$.*
2. *Solve the objective function (A) or (B) for the parameter sets $\mathscr{B} = \{B_A + b_\Delta, b_\Delta \in \mathscr{B}_\Delta\}$ and $\mathscr{S} = \{S_A + s_\Delta, s_\Delta \in \mathscr{S}_\Delta\}$, and update $z \mapsto S_A(z)$ and $z \mapsto B_A(z)$ to be the solution.*
3. *Optionally, iterate on step 2 a few times.*

Ideally, step 1 of the algorithm should make use of cross-fitting, but we present a simplified version for clarity.

Comment 6.2 (Choices for Computational Experiments)**.** We use simple deviation parameter spaces with low complexity (small ORC). For example, in the computational experiments reported in Figures 3 and 4, we made the following choices: for Causal Neural Network (NN) Learner, we used shallow, regularized NN as deviation spaces when the predictive learner was NN; for Forest Causal Learner, we used the shallow forest as the deviation space when the predictive learner was Random Forest; for Causal OLS Learner, we used linear deviation space, when the predictive learner was OLS. Finally, when we used Causal Forest as the predictive learner, we used the shallow forest as the deviation space to obtain the Forest Causal Learner.[29] We have also experimented with hybrid versions, for example, using one type of predictive learner and a different type of booster (for example, RF plus NN as a causal boost or Elastic Net plus NN as a casual boost).

Comment 6.3 (Choices for Empirical Example)**.** We used simple, regularized neural networks as deviation spaces in all results. The number of neurons was kept less than or equal to 10 and was chosen based on cross-validating the objective function of type (A) over auxiliary data subsamples. We used this choice regardless of the predictive ML as the starter. Causal Learners constructed in this way improved the performance of each predictive ML method, raising the goodness of fit metrics by 5-10% in relative terms. However, they did not have any qualitative impact on empirical results (we report the results for predictive learners in Online Appendix C).

[29] As this may sound confusing, we note that that Forest Causal Learners (FCL) differ from Causal Random Forests (CRF) in stratified experiments as FCL are based on weighted residualization whereas CRF are based on unweighted residualization of Nie and Wager (2020). Therefore we use a slightly different name "Forest Causal Learner" rather than "Causal Forest" to distinguish the proposal.

Comment 6.4 (Predictive ML Methods)**.** We considered four ML methods to estimate the proxy predictors: elastic net, boosted trees, neural network with feature extraction, and random forest. The ML methods are implemented in R using the package caret (Kuhn, 2008). The names of the elastic net, boosted tree, neural network with feature extraction, and random forest methods in caret are glmnet, gbm, pcaNNet and rf, respectively. For each split of the data, we choose the tuning parameters separately for $B(z)$ and $S(z)$ based on mean squared error estimates of repeated 2-fold cross-validation, except for random forest, for which we use the default tuning parameters to reduce the computational time.[30] In tuning and training the ML methods we use only the auxiliary sample. In all the methods we rescale the outcomes and covariates to be between 0 and 1 before training.

7. Conclusion and Extensions

We propose to focus inference on key features of heterogeneous effects in randomized experiments, and develop the corresponding methods. These key features include best linear predictors of the effects and average effects sorted by groups, as well as average characteristics of most and least affected units. Our approach is valid in high dimensional settings, where the effects are estimated by machine learning methods. The main advantage of our approach is its agnostic nature; it avoids making strong assumptions. Estimation and inference relies on data splitting, where the latter allows us to avoid overfitting and all kinds of non-regularities. Our inference aggregates the results across many splits, reducing the replication risks, and could be of independent interest. An empirical application illustrates the practical use of the approach.

Our hope is that applied researchers use the method to discover whether there is heterogeneity in their data in a disciplined way. A researcher might be concerned about the application of our method due to the possible power loss induced by sample splitting. This power loss is the price to pay when the researcher is not certain or willing to fully specify the form of the heterogeneity prior to conducting the experiment. Thus, if the researcher has a well-defined pre-analysis plan that spells out a small number of heterogeneity groups in advance, then there is no need of splitting the sample.[31] However, this situation is not common. In general, the researchers might not be able to fully specify the form of the heterogeneity due to lack of information, economic theory, or

[30] We have the following tuning parameters for each method: Elastic Net: alpha (Mixing Percentage), lambda (Regularization Parameter), Boosted trees: n.trees (Number of Boosting Iterations), interaction.depth (Max Tree Depth), shrinkage (Shrinkage), n.minobsinnode (Min. Terminal Node Size), size (Number of Hidden Units) , decay (Weight Decay), mtry (Number of Randomly Selected Predictors).

[31] More generally, the plan needs to specify a parametric form for the heterogeneity as a low dimensional function of pre-specified covariates (e.g., Chernozhukov et al., 2015). In this case, ML tools can still be used to efficiently estimate the CATEs in the presence of control variables but are not required to detect heterogeneity (Belloni et al., 2017; Chernozhukov et al., 2017).

willingness to take a stand at the early stages of the analysis. They might also face data limitations that preclude the availability of the desired covariates. Here we recommend the use of our method to avoid overfitting and p-hacking, and impose discipline to the heterogeneity analysis at the cost of some power loss due to sample splitting.[32] If discovering and exploiting heterogeneity in treatment effect is a key goal of the research, the researcher should indeed plan for larger sample sizes (relative to just testing whether the treatment has an effect), but the required sample size remains within the realm of what is feasible in the field. In many applications we are aware of, there was apparent heterogeneity according to some covariates of interest, but the disciplined ML heterogeneity exercise found no systematic difference. This could be because this heterogeneity was a fluke, or because the method does not have the power to detect it in a small sample. In any case, what this experience suggests is that one should not rely on ex-post heterogenous effects in such cases.

The application to immunization in India is of substantive interest. Our findings suggest that a combination of small incentives, relay by information hub, and SMS reminders can have very large effect on vaccine take up in some villages where immunization was low at baseline, and even be more cost-effective than the status quo, but can also backfire in other places. This suggests that these type of strategy need to be piloted in the relevant context before being rolled out, and that heterogeneity needs to be taken into account.

Extensions to Unbiased Signal Framework. Our inference approach generalizes to any problem of the following sort, studied in Semenova and Chernozhukov (2021) using more conventional inference approaches. Suppose we can construct an *unbiased signal* $\tilde{Y}$ such that

$$\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z),$$

where $s_0(Z)$ is now a generic target function. Let $S(Z)$ denote an ML proxy for $s_0(Z)$. In experimental settings the unbiased signals arise from multiplying an outcome with a Riesz representer for the effect of interest, as we explain below.

Then, using previous arguments, we immediately can generate the following conclusions:

(1) The projection of $\tilde{Y}$ on the ML proxy $S(Z)$ identifies the BLP of $s_0(Z)$ on $S(Z)$.
(2) The grouped average of the target (GAT) $\mathrm{E}[s_0(Z) \mid G_k]$ is identified by $\mathrm{E}[\tilde{Y} \mid G_k]$.
(3) Using ML tools we can train proxy predictors $S(Z)$ to predict $\tilde{Y}$ in auxiliary samples.
(4) We can post-process $S(Z)$ in the main sample, by estimating the BLP and GATs.

[32] It is not clear whether this loss is real though, as we are not aware of any alternative method that avoids sample splitting and that works at the same level of agnosticism as ours. In a previous version of the paper we provided a numerical example using a simple parametric model where standard methods without sample splitting are available. We find that the extent of the power loss for not using the parametric form of the heterogeneity roughly corresponds to reducing the sample size by half in a test for the presence of heterogeneity, although the exact comparison depends on features of the data generating process.

(5) We can perform split-sample robust inference on functionals of the BLP and GATs.

Example 1 (Forecasting or Predicting Regression Functions using ML proxies)**.** This is the most common type of the problem arising in forecasting. Here the target is the best predictor of Y using Z, namely $s_0(Z) = \mathrm{E}[Y \mid Z]$, and $\tilde{Y} = Y$ trivially serves as the unbiased signal. The interesting part here is the use of the inference tools developed in this paper for constructing confidence intervals for the predicted values produced by the estimated BLP of $s_0(Z)$ using $S(Z)$.

Example 2 (Predicting Structural Derivatives using ML proxies)**.** Suppose we are interested in predicting the conditional average partial derivative $s_0(z) = \mathrm{E}[g'(D,Z) \mid Z = z]$, where $g'(d,z) = \partial g(d,z)/\partial x$ and $g(d,z) = \mathrm{E}[Y \mid D = d, Z = z]$. In the context of demand analysis, Y is the log of individual demand, D is the log-price of a product, and Z includes prices of other products and individual characteristics. Then, the unbiased signal is given by $\tilde{Y} = -Y[\partial \log p(D \mid Z)/\partial d]$, where $p(\cdot \mid \cdot)$ is the conditional density function of D given Z, which is known if D is generated experimentally conditional on Z. That is, using the integration by parts formula, $\mathrm{E}[\tilde{Y} \mid Z] = s_0(Z)$ under mild conditions on the density.

Example 3 (Other Causal Objects)**.** Chernozhukov et al. (2018) presented a number of other examples where a causal parameter of interest $s_0(Z)$ is expressed as a linear functional of the regression function $g(D,Z) = \mathrm{E}[Y \mid D,Z]$, that is, $s_0(Z) = \mathrm{E}[m(Y,D,Z,g) \mid Z]$, for some moment function m that is linear in g; this includes the examples above for instance. Then, under mild conditions, we can construct an unbiased signal

$$\tilde{Y} = Y\alpha(D,Z), \tag{7.1}$$

where $\alpha(D,Z)$ is the Riesz Representer, such that $\mathrm{E}[Y\alpha(D,Z) \mid Z] = s_0(Z)$. For instance, in CATE, the representer $\alpha(D,Z)$ is the HT transform H; in Example 2, $\alpha(D,Z) = [\partial \log p(X \mid Z)/\partial x]$; and in Example 1, the representer $\alpha(D,Z)$ is just 1. In addition to these examples, other examples that fall in this framework include causal effects from transporting covariates and from distributional shift in covariates induced by policies; see Chernozhukov et al. (2018) for more details. In experimental settings, $\alpha(D,Z)$ will typically be known.

The noise reduction strategies, like the ones we used in the context of H-transformed outcomes, can be useful in these cases as well. For this purpose we can use terms of the form $\{\alpha(D,Z) - \mathrm{E}[\alpha(D,Z) \mid Z]\}B(Z)$ for denoising where $\alpha(D,Z)$ now plays the same role as H before.

References

Alberto Abadie. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19, 2005.

Alberto Abadie, Matthew M Chingos, and Martin R West. Endogenous stratification in randomized experiments. Technical report, National Bureau of Economic Research, 2017.

Anish Agarwal, Abdullah Alomar, Romain Cosson, Devavrat Shah, and Dennis Shen. Synthetic interventions, 2020.

Vivi Alatas, Arun G Chandrasekhar, Markus Mobius, Benjamin A Olken, and Cindy Paladines. When celebrities speak: A nationwide twitter experiment promoting vaccination in indonesia. Technical report, National Bureau of Economic Research, 2019.

Joshua D. Angrist and Jorn-Steffan Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

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. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20, 2007.

Keith Ball. Volume ratios and a reverse isoperimetric inequality. *Journal of the London Mathematical Society*, 2(2):351–359, 1991.

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Suresh Dalpath, John Floretta, Matthew Jackson, Harini Kannan, Anna Schrimpf, and Mahesh Shrestha. Leveraging the social network amplifies the effectiveness of interventions to stimulate take up of immunization. 2019.

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Suresh Dalpath, John Floretta, Matthew Jackson, Loza Francine, Harini Kannan, and Anna Schrimpf. Inference on winners. Technical Report 28726, NBER Working Paper, 2021.

Abhijit Vinayak Banerjee, Esther Duflo, Rachel Glennerster, and Dhruva Kothari. Improving immunisation coverage in rural india: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *BMJ*, 340, 2010. ISSN 0959-8138. doi: 10.1136/bmj.c2220.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability, 2022.

G. Barnard. Discussion of "Cross-validatory choice and assessment of statistical predictions" by Stone. *Journal of the Royal Statistical Society. Series B (Methodological)*, page 133?135, 1974.

Diego G Bassani, Paul Arora, Kerri Wazny, Michelle F Gaffey, Lindsey Lenters, and Zulfiqar A Bhutta. Financial incentives and coverage of child health interventions: a systematic review and meta-analysis. *BMC Public Health*, 13(S3):S30, 2013.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Review of Economic Studies*, 81:608–650, 2014.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post selection inference for lad regression models. *arXiv preprint arXiv:1304.0282*, 2013.

Vidmantas Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. Debiased machine learning without sample-splitting for stable estimators. *arXiv preprint arXiv:2206.01825*, 2022.

V. Chernozhukov, I. Fernández-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009. ISSN 0006-3444. doi: 10.1093/biomet/asp030.

V. Chernozhukov, I. Fernandez-Val, and Y. Luo. The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *ArXiv e-prints*, December 2015.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with very many controls and instruments. *American Economic Review: Papers and Proceedings*, 105(5):486–490, 2015.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

Victor Chernozhukov, Whitney Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. A simple and general debiased machine learning theorem with finite sample guarantees. *arXiv preprint arXiv:2105.15197*, 2021a.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021b.

DR Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975.

Bruno Crepon, Esther Duflo, Huillery Elisa, William Pariente, Juliette Seban, and Paul-Armand Veillon. Cream skimming and the comparison between social interventions evidence from entrepreneurship programs for at-risk youth in france. 2021. Mimeo.

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.

Jonathan MV Davis and Sara B Heller. Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *Review of Economics and Statistics*, 102(4):664–677, 2020.

Tatyana Deryugina, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif. The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219, 2019.

Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *arXiv preprint arXiv:1606.03940*, 2016.

Cyrus J DiCiccio, Thomas J DiCiccio, and Joseph P Romano. Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865, 2020.

Gretchen J Domek, Ingrid L Contreras-Roldan, Sean T O'Leary, Sheana Bull, Anna Furniss, Allison Kempe, and Edwin J Asturias. Sms text message reminders to improve infant vaccination coverage in guatemala: a pilot randomized controlled trial. *Vaccine*, 34(21):2437–2443, 2016.

R. Dudley. *Uniform Cental Limit Theorems*. Cambridge Studies in advanced mathematics, 2000.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962, 2007.

Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327, 2022.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

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

David A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008. ISSN 0196-8858.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

A Ronald Gallant and Halbert White. *A unified theory of estimation and inference for nonlinear dynamic models*. Blackwell, 1988.

Christopher Genovese and Larry Wasserman. Adaptive confidence bands. *The Annals of Statistics*, pages 875–905, 2008.

Dustin G Gibson, Benard Ochieng, E Wangeci Kagucia, Joyce Were, Kyla Hayford, Lawrence H Moulton, Orin S Levine, Frank Odhiambo, Katherine L O'Brien, and Daniel R Feikin. Mobile phone-delivered reminders and incentives to improve childhood immunisation coverage and timeliness in kenya (m-simu): a cluster randomised controlled trial. *The Lancet Global Health*, 5(4):e428–e438, 2017.

Evarist Giné and Richard Nickl. Confidence bands in density estimation. *Ann. Statist.*, 38(2): 1122–1170, 2010. ISSN 0090-5364.

Uri Gneezy and Aldo Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Christian Hansen, Damian Kozbur, and Sanjog Misra. Targeted undersmoothing. *arXiv preprint arXiv:1706.07328*, 2017.

John A Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64(328):1303–1317, 1969.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity*, volume 143 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015. ISBN 978-1-4987-1216-3. The lasso and generalizations.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003. ISSN 0012-9682. doi: 10.1111/1468-0262.00442. URL `http://dx.doi.org/10.1111/1468-0262.00442`.

Wassily Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192, 1952.

Reshmaan Hussam, Natalia Rigol, and Benjamin N Roth. Targeting high ability entrepreneurs using community information: Mechanism design in the field. *American Economic Review*, 112 (3):861–98, 2022.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

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

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning—with applications in R*. Springer Texts in Statistics. Springer, New York, 2021. Second edition.

Mira Johri, Myriam Cielo Pérez, Catherine Arsenault, Jitendar K Sharma, Nitika Pant Pai, Smriti Pahwa, and Marie-Pierre Sylvestre. Strategies to increase the demand for childhood vaccination in low-and middle-income countries: a systematic review and meta-analysis. *Bulletin of the World Health Organization*, 93:339–346, 2015.

Anne Karing. Social signaling and childhood immunization: A field experiment in sierra leone. *University of California, Berkeley Working Paper*, 2018.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Leslie Kish and Martin Richard Frankel. Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–37, 1974.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Max Kuhn. Caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Mark G Low et al. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Regularised orthogonal machine learning for nonlinear semiparametric models. *The Econometrics Journal*, 25(1):233–255, 2022.

J Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes (masters thesis); justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci*, 5:463–472, 1923.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 09 2020.

Angela Oyo-Ita, Charles S Wiysonge, Chioma Oringanje, Chukwuemeka E Nwachukwu, Olabisi Oduwole, and Martin M Meremikwu. Interventions for improving coverage of childhood immunisation in low-and middle-income countries. *Cochrane Database of Systematic Reviews*, (7), 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Martin Raič. A multivariate berry–esseen theorem with explicit constants. *Bernoulli*, 25(4A): 2824–2853, 2019.

Annette K Regan, Lauren Bloomfield, Ian Peters, and Paul V Effler. Randomized controlled trial of text message reminders for increasing influenza vaccination. *The Annals of Family Medicine*, 15(6):507–514, 2017.

Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*, 2016.

P. M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 0012-9682. doi: 10.2307/1912705.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 08 2021.

Vira Semenova, Matt Goldman, Victor Chernozhukov, and Matt Taddy. Estimation and inference on heterogeneous treatment effects in high-dimensional dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. ISSN 0090-5364.

Md Jasim Uddin, Md Shamsuzzaman, Lily Horng, Alain Labrique, Lavanya Vasudevan, Kelsey Zeller, Mridul Chowdhury, Charles P Larson, David Bishai, and Nurul Alam. Use of mobile phones for improving vaccination coverage among children living in rural hard-to-reach areas and urban streets of bangladesh. *Vaccine*, 34(2):276–283, 2016.

Unicef. 20 million children miss out on lifesaving measles, diphtheria and tetanus vaccines in 2018. *https://www.unicef.org/eca/press-releases/20-million-children-miss-out-lifesaving-measles-diphtheria-and-tetanus-vaccines-2018*, 2019.

Suhas Vijaykumar. Localization, convexity, and star aggregation. *Advances in Neural Information Processing Systems*, 34:4570–4581, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113 (45):12673–12678, 2016.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Hotenzia Wakadha, Subhash Chandir, Elijah Victor Were, Alan Rubin, David Obor, Orin S Levine, Dustin G Gibson, Frank Odhiambo, Kayla F Laserson, and Daniel R Feikin. The feasibility of using mobile-phone based sms reminders and conditional cash transfers to improve timely immunization in rural kenya. *Vaccine*, 31(6):987–993, 2013.

Larry Wasserman. Machine learning overview. In *Becker-Friedman Institute, Conference on ML in Economics*, 2016.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Marten Wegkamp et al. Model selection in nonparametric regression. *The Annals of Statistics*, 31 (1):252–273, 2003.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

Michael Zimmert and Michael Lechner. Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*, 2019.

Appendix A. Proofs of Section 3

Proof of Theorem 3.1. The subset of the normal equations, which correspond to $\alpha := (\alpha_1, \alpha_2)'$, are $\mathrm{E}[w(Z)(Y - \alpha_0' X_1 - \alpha' X_2)X_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $X_2 = X_2(Z,D) = [D - p(Z), (D - p(Z)(S - \mathrm{E}S)]'$, $X_1 = X_1(Z)$, and the law of iterated expectations, we notice that:

$$\mathrm{E}[w(Z)b_0(Z)X_2] = \mathrm{E}[w(Z)b_0(Z)\underbrace{\mathrm{E}[X_2(Z,D) \mid Z]}_{=0}] = 0,$$
$$\mathrm{E}[w(Z)UX_2] = \mathrm{E}[w(Z)\underbrace{\mathrm{E}[U \mid Z,D]}_{0}X_2(Z,D)] = 0,$$
$$\mathrm{E}[w(Z)X_1X_2] = \mathrm{E}[w(Z)X_1(Z)\underbrace{\mathrm{E}[X_2(Z,D) \mid Z]}_{=0}] = 0.$$

Hence the normal equations simplify to: $\mathrm{E}[w(Z)(s_0(Z)D - \alpha' X_2)X_2] = 0$. Since

$$\mathrm{E}[\{D - p(Z)\}\{D - p(Z)\} \mid Z] = p(Z)(1 - p(Z)) = w^{-1}(Z),$$

and $S = S(Z)$, the components of X_2 are orthogonal by the law of iterated expectations:

$$\mathrm{E}[w(Z)(D - p(Z))(D - p(Z))(S - \mathrm{E}S)] = \mathrm{E}(S - \mathrm{E}S) = 0.$$

Hence the normal equations above further simplify to

$$\mathrm{E}[w(Z)\{s_0(Z)D - \alpha_1(D - p(Z))\}(D - p(Z))] = 0,$$
$$\mathrm{E}[w(Z)\{s_0(Z)D - \alpha_2(D - p(Z))(S - \mathrm{E}S)\}(D - p(Z))(S - \mathrm{E}S)] = 0.$$

Solving these equations and using the law of iterated expectations, we obtain

$$\begin{aligned}
\alpha_1 &= \frac{\mathrm{E}[w(Z)\{s_0(Z)D(D - p(Z))]\}}{\mathrm{E}[w(Z)(D - p(Z))^2]} = \frac{\mathrm{E}[w(Z)s_0(Z)w^{-1}(Z)]}{\mathrm{E}[w(Z)w^{-1}(Z)]} = \mathrm{E}s_0(Z),\\
\alpha_2 &= \frac{\mathrm{E}[w(Z)\{s_0(Z)D(D - p(Z))(S - \mathrm{E}S)\}]}{\mathrm{E}[w(Z)(D - p(Z))^2(S - \mathrm{E}S)^2]}\\
&= \frac{\mathrm{E}[w(Z)s_0(Z)w^{-1}(Z)(S - \mathrm{E}S)]}{\mathrm{E}[w(Z)w^{-1}(Z)(S - \mathrm{E}S)^2]} = \frac{\mathrm{Cov}(s_0(Z), S)}{\mathrm{Var}(S)}.
\end{aligned}$$

The conclusion follows by noting that these coefficients also solve the normal equations

$$\mathrm{E}\{[s_0(Z) - \alpha_1 - \alpha_2(S - \mathrm{E}S)][1, (S - \mathrm{E}S)]'\} = 0,$$

which characterize the optimum in the problem of best linear approximation/prediction of $s_0(Z)$ using S. ■

Proof of Theorem 3.2. The subset of the normal equations, which correspond to $\mu := (\mu_1, \mu_2)'$, are $\mathrm{E}[(YH - \mu_0' X_1 H - \mu' \tilde{X}_2)\tilde{X}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{X}_2 = \tilde{X}_2(Z) = [1, (S(Z) - \mathrm{E}S(Z))]'$, $X_1 = X_1(Z)$, and the law of iterated expectations, we notice that:

$$\mathrm{E}[b_0(Z)H\tilde{X}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H(D,Z) \mid Z]}_{=0}\tilde{X}_2(Z)] = 0,$$
$$\mathrm{E}[UH\tilde{X}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z, D]}_{0}H(D,Z)\tilde{X}_2(Z)] = 0,$$
$$\mathrm{E}[X_1(Z)H\tilde{X}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H(D,Z) \mid Z]}_{=0}\tilde{X}_2(Z)] = 0.$$

Hence the normal equations simplify to: $\mathrm{E}[(s_0(Z)DH - \mu'\tilde{X}_2)\tilde{X}_2] = 0$. Since 1 and $S - \mathrm{E}S$ are orthogonal, the normal equations above further simplify to

$$\mathrm{E}\{s_0(Z)DH - \mu_1\} = 0, \ \ \mathrm{E}[\{s_0(Z)DH - \mu_2(S - \mathrm{E}S)\}(S - \mathrm{E}S)] = 0.$$

Using that $\mathrm{E}[DH \mid Z] = [p(Z)(1 - p(Z))]/[p(Z)(1 - p(Z))] = 1$, $S = S(Z)$, and the law of iterated expectations, the equations simplify to

$$\mathrm{E}\{s_0(Z) - \mu_1\} = 0, \ \ \mathrm{E}[\{s_0(Z) - \mu_2(S - \mathrm{E}S)\}(S - \mathrm{E}S)] = 0.$$

These are normal equations that characterize the optimum in the problem of best linear approximation/prediction of $s_0(Z)$ using S. Solving these equations gives the expressions for β_1 and β_2 stated in Definition 3.1. ■

Proof of Theorem 3.3. The proof is similar to the proof of Theorem 3.1- 3.2. Moreover, since the proofs for the two strategies are similar, we will only demonstrate the proof for the second strategy.

The subset of the normal equations, which correspond to $\mu := (\mu_k)_{k=1}^K$, are given by $\mathrm{E}[(YH - \mu_0' X_1 H - \mu'\tilde{W}_2)\tilde{W}_2] = 0$. Substituting $Y = b_0(Z) + s_0(Z)D + U$, and using the definition $\tilde{W}_2 = \tilde{W}_2(Z) = [1(G_k)_{k=1}^K]'$, $X_1 = X_1(Z)$, and the law of iterated expectations, we notice that:

$$\mathrm{E}[b_0(Z)H\tilde{W}_2(Z)] = \mathrm{E}[b_0(Z)\underbrace{\mathrm{E}[H(D,Z) \mid Z]}_{=0}\tilde{W}_2(Z)] = 0,$$
$$\mathrm{E}[UH\tilde{W}_2(Z)] = \mathrm{E}[\underbrace{\mathrm{E}[U \mid Z, D]}_{0}H(D,Z)\tilde{W}_2(Z)] = 0,$$
$$\mathrm{E}[X_1 H\tilde{W}_2(Z)] = \mathrm{E}[X_1(Z)\underbrace{\mathrm{E}[H(D,Z) \mid Z]}_{=0}\tilde{W}_2(Z)] = 0.$$

Hence the normal equations simplify to: $\mathrm{E}[\{s_0(Z)DH - \mu'\tilde{W}_2\}\tilde{W}_2] = 0$. Since components of $\tilde{W}_2 = \tilde{W}_2(Z) = [1(G_k)_{k=1}^K]'$ are orthogonal, the normal equations above further simplify to $\mathrm{E}[\{s_0(Z)DH - \mu_k 1(G_k)\}1(G_k)] = 0$. Using that $\mathrm{E}[DH \mid Z] = 1$, $S = S(Z)$, and the law of iterated expectations, the equations simplify to

$$\mathrm{E}[\{s_0(Z) - \mu_k 1(G_k)\}1(G_k)] = 0 \iff \ \mu_k = \mathrm{E}s_0(Z)1(G_k)/\mathrm{E}[1(G_k)] = \mathrm{E}[s_0(Z) \mid G_k].$$

The asserted result follows. ■

Appendix B. Supporting Results and Proofs of Section 4

Proof of Lemma 4.1. To show (4.2) note that,

$$\mathrm{E}|\widehat{\theta}-\theta'| = \mathrm{EE}[|\widehat{\theta}-\theta'| \mid \text{Data}] \leqslant \mathrm{EE}[|\widehat{\theta}_A-\theta'| \mid \text{Data}] \leqslant \mathrm{E}|\widehat{\theta}_A-\theta'|,$$

where the inequality follows from (any) median minizing average absolute loss and its equivariance property. The equalities hold by the law of iterated expectation. The claim (4.3) follows in the same way.

To show (4.4), let $U^* = \{U^*_a\}_{a\in\mathscr{A}}$ and $L^* = \{L^*_a\}_{a\in\mathscr{A}}$ denoted non-decreasing monotone rearrangements of $\{U_a\}_{a\in\mathscr{A}}$ and $L = \{L_a\}_{a\in\mathscr{A}}$. Then

$$|U-L| \leqslant \|U^*-L^*\|_\infty \leqslant \|U-L\|_\infty,$$

where the second inequality follows from the rearrangement having contractive property in the max distance. ■

Proof of Theorem 4.1. We demonstrate the result for p^+. The proofs for other p-values follow similarly. We use $M_{\mathscr{A}}[\cdot]$ as short hand for $\mathrm{M}[\cdot|\text{Data}]$, with overlined and underlined versions defined similarly.

To show claim (ii) we note that for $z = \Phi^{-1}(1-\alpha)$ and using that $\Phi(z) = 1-\alpha$:

$$\begin{aligned}
\mathrm{P}\Big(M_{\mathscr{A}}[1-\Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A-\theta_0))] < \alpha\Big) &= \mathrm{P}\Big(M_{\mathscr{A}}[\Phi(-\widehat{\sigma}_A^{-1}(\widehat{\theta}_A-\theta_0))] < \alpha-1\Big) \\
&= \mathrm{P}\Big(\, M_{\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)] < -z\Big) \\
&\leqslant \mathrm{P}\Big(\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A) < -z\Big) + \gamma''_N \\
&\leqslant \Phi(-z) + \gamma'_N + \gamma''_N = \alpha + \gamma'_N + \gamma''_N,
\end{aligned}$$

where the first inequality uses the concentration of median assumption, and the last inequality follows from the approximate normality assumption (4.5).

To show claim (i), we note that

$$\begin{aligned}
\mathrm{P}\Big(M_{\mathscr{A}}[1-\Phi(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A-\theta_0))]<\alpha\Big) &= \mathrm{P}\Big(\,M_{\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)]<-z\Big)\\
&\leqslant \mathrm{P}\left(\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}}1\{\,\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)]<-z\}\geqslant 1/2\right)\\
&\leqslant 2\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}}1\{\,\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)]<-z\}\right]\\
&= 2\mathrm{P}\{\,\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)]<-z\}\\
&\leqslant 2\Phi(-z)+2\gamma_N'=2\alpha+2\gamma_N',
\end{aligned}$$

where the first equality reused the previous calculation, the first inequality holds by definition of the numerical median, the second inequality holds by Markov inequality, and the equality that follows holds by

$$\begin{aligned}
2\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}}1\{\,\widehat{\sigma}_a^{-1}(\theta_a-\widehat{\theta}_a)]<-z\}\right] &= 2\mathrm{EP}\Big(\,\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)<-z\mid \mathrm{Data}\Big)\\
&= 2\mathrm{P}\{\,\widehat{\sigma}_A^{-1}(\theta_A-\widehat{\theta}_A)]<-z\},
\end{aligned}$$

and the last inequality follows from the approximate normality assumption (4.5).

Proof of Theorem 4.2. Define $\mathscr{D}=\{\theta_a,[L_a,U_a]:a\in\mathscr{A}\}$, and let $A\sim U(\mathscr{A})$ given $\mathscr{D}$. Then,

$$\begin{aligned}
\mathrm{P}(\theta_A<L) &= \mathrm{EP}(\theta_A<L\mid\mathscr{D})=\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{\theta_a<L\}\right]\\
&= \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{\theta_a<L,\,L_a<L\}+\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{\theta_a<L,\,L_a\geqslant L\}\right]\\
&\leqslant \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{L_a<L\}\right]+\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{\theta_a<L_a\}\right]\\
&\leqslant \mathrm{E}(\beta)+\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{\theta_a<L_a\}\right]\\
&\leqslant \beta+\mathrm{EP}(\theta_A<L_A\mid\mathscr{D})\\
&\leqslant \beta+\mathrm{P}\{\theta_A<L_A\}\leqslant\beta+\alpha/2+o(1),
\end{aligned}$$

where the first equality holds by the law of iterated expectations; the second by the fact that, given $\mathscr{D}$, L is fixed but $A\sim U(\mathscr{A})$; the second inequality holds by definition of L:

$$\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{L_a<L\}\leqslant\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}}1\{L_a<\overline{Q}_\beta[L_A\mid\mathrm{Data}]\}\leqslant\beta,$$

by the definition of the upper quantile; and the third by the same argument as the second equality, the penultimate inequality holds by the law of iterated expectations, and the last inequality holds by assumption (4.8). We conclude similarly

$$\mathrm{P}(\theta_A > U) \leqslant \beta + \mathrm{P}\{\theta_A > U_A\} \leqslant \beta + \alpha/2 + o(1).$$

The asserted result follows. ■

B.1. **Proof of Theorem 4.3.** In the proof let $z = \Phi^{-1}(1-\alpha/2)$, and use $M_{\mathscr{A}}[\cdot]$ and $Q_{\mathscr{A}}[\cdot]$ as short hand for $\mathrm{M}[\cdot|\text{Data}]$ and $\mathrm{Q}[\cdot|\text{Data}]$, respectively, with overlined and underlined versions defined similarly.

We first note

$$\begin{aligned} &\mathrm{P}(\theta^* \notin [L,U]) = \mathrm{P}(\theta^* > M_{\mathscr{A}}(\widehat{\theta}_A + z\widehat{\sigma}_A)) + \mathrm{P}(\theta^* < M_{\mathscr{A}}(\widehat{\theta}_A - z\widehat{\sigma}_A)) \\ &= \ \mathrm{P}(0 > M_{\mathscr{A}}(\widehat{\theta}_A - \theta^* + z\widehat{\sigma}_A)) + \mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)). \end{aligned}$$

To show part (i) with $\beta = 1/2$,

$$\begin{aligned} \mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)) &\leqslant \mathrm{P}\left(\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right) \geqslant 1/2\right) \\ &\leqslant 2\mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right)\right] \\ &= 2\mathrm{E}\mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) > z \mid \text{Data}\right) \\ &= 2\mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta^*) > z\right) \\ &\leqslant 2\mathrm{P}\left(\widehat{\sigma}_A^{-1}(\widehat{\theta}_A - \theta_A) > z - r_N\right) + 2\gamma_N''' \\ &\leqslant 2(1 - \Phi(z - r_N)) + 2\gamma_N' + 2\gamma_N''' \\ &\leqslant 2\alpha/2 + 2r_N/\sqrt{2\pi} + 2\gamma_N' + 2\gamma_N''', \end{aligned}$$

where the first inequality follows from the definition of the numerical median, the second from the Markov inequality; the first equality holds by $A \sim U(\mathscr{A})$ given Data, the second by the law of iterated expectations; the third inequality holds by the concentration condition (R3) and the union bound, the penultimate inequality holds by the approximation normality conditions (R1), and the last from the properties of Φ.

We derive similarly that

$$\mathrm{P}(0 > M_{\mathscr{A}}(\widehat{\theta}_A - \theta^* + z\widehat{\sigma}_A)) \leqslant 2\alpha/2 + 2r_N/\sqrt{2\pi} + 2\gamma_N' + 2\gamma_N'''.$$

Therefore the part (i) holds for the term

$$o(1) := 4r_N/\sqrt{2\pi} + 4(\gamma_N' + \gamma_N''').$$

To show part (ii), from the analysis of part (i),

$$\mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)) \leqslant \mathrm{P}\left(\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right) \geqslant 1/2\right).$$

Then we bound

$$T = \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta^*) > z\right) \leqslant T_1 + T_2$$

$$T_1 := \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\widehat{\theta}_a - \theta_a) > z - r_N\right), \quad T_2 := \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} 1\left(\widehat{\sigma}_a^{-1}(\theta_a - \theta^*) > r_N\right).$$

By the union bound

$$\mathrm{P}(T \geqslant 1/2) \leqslant \mathrm{P}\left(T_1 > 1/2 - 2\sqrt{\gamma_N'''}\right) + \mathrm{P}\left(T_2 \geqslant \sqrt{\gamma_N'''}\right).$$

Then for $\beta_N = 1/2 - 2\sqrt{\gamma_N'''}$,

$$\begin{aligned}
\mathrm{P}(T_1 > \beta_N) &\leqslant \mathrm{P}\left(\overline{\mathrm{Q}}_{\beta_N,\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A)] < -z + r_N\right) \\
&\leqslant \mathrm{P}\left(\mathrm{Q}_{\beta_N,\mathscr{A}}[\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A)] < -z + r_N\right) \\
&\leqslant \mathrm{P}\left(\widehat{\sigma}_A^{-1}(\theta_A - \widehat{\theta}_A) < -z + r_N\right) + \gamma_N'' \\
&\leqslant \Phi(-z + r_N) + \gamma_N' + \gamma_N'' \\
&\leqslant \alpha/2 + r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'',
\end{aligned}$$

where first inequality holds by the definition of the numerical quantile, the third by the concentration of medians assumption (R2), and the fourth by the approximate normality (R1).

Also, by Markov inequality

$$\mathrm{P}\left(T_2 \geqslant \sqrt{\gamma_N'''}\right) \leqslant \mathrm{E}T_2/\sqrt{\gamma_N'''} = \mathrm{P}\left(\sigma_A^{-1}|\theta_A - \theta^*| > r_N\right)/\sqrt{\gamma_N'''} \leqslant \gamma_N'''/\sqrt{\gamma_N'''},$$

where we are using (R3) and the relation

$$\begin{aligned}
\mathrm{E}T_2 &= \mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} 1\left(\sigma_a^{-1}|\theta_a - \theta^*| > r_N\right)\right] \\
&= \mathrm{EP}(\sigma_A^{-1}|\theta_A - \theta^*| > r_N \mid \mathrm{Data}) = \mathrm{P}\left(\sigma_A^{-1}|\theta_A - \theta^*| > r_N\right),
\end{aligned}$$

using our formalism that $A \sim U(\mathscr{A})$ independently of Data.

Collecting terms conclude

$$\mathrm{P}(0 < M_{\mathscr{A}}(\widehat{\theta} - \theta^* - z\widehat{\sigma}_A)) \leqslant \alpha/2 + r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'' + \sqrt{\gamma_N'''}.$$

We derive similarly that

$$\mathrm{P}(0 > M_{\mathscr{A}}(\widehat{\theta}_A - \theta^* + z\widehat{\sigma}_A)) \leqslant \alpha/2 + r_N/\sqrt{2\pi} + \gamma_N' + \gamma_N'' + \sqrt{\gamma_N'''}.$$

Therefore the part (ii) holds for the term

$$o(1) = 2\left(r_N/\sqrt{2\pi} + \gamma'_N + \gamma''_N + \sqrt{\gamma'''_N}\right).$$

To show claim (iii) note that by construction $L \leqslant \widehat{\theta} \leqslant U$ and the coverage event $L \leqslant \theta^* \leqslant U$ implies that $|\widehat{\theta} - \theta^*| \leqslant U - L$. ■

B.2. **Concentration of Estimands Around Their Median.** The purpose of this section is to demonstrate that the concentration assumptions made in the inference section are plausible. To show this we focus on the BLP parameter

$$\theta_A = \frac{\mathrm{Cov}_Z(s_0(Z), S_A(Z)}{\mathrm{Var}_Z S_A(Z))},$$

where A is a uniform variable on $\mathscr{A}$, and the variance and covariance are taken with respect to the marginal distribution of Z. We want to show the concentration of this parameter around

$$\theta^* = \mathrm{Med}[\theta_A \mid \mathrm{Data}].$$

We show the difference can be bounded using measures of estimation and algorithmic stabilities; we derive inspiration from Chernozhukov et al. (2021b) and Chen et al. (2022)).

In what follows, we assume the same set-up as in the main text, in particular the exchangeability.

Estimation Stability or Pseudo-Consistency. Statistical learning theory, for example, results in Section 5, provides bounds on estimation errors of the form:

$$\mathrm{E}\mathrm{E}_Z(S_A(Z) - s_\bullet(Z))^2 = \mathrm{E}(S_A(Z) - s_\bullet(Z))^2 \leqslant R^2_{|A|},$$

where $s_\bullet$ is a fixed "pseudo-true" function that does not depend on A, and this function does not have to be the CATE s_0 in the misspecified case. Here E_Z denotes the expectation taken with respect to the marginal distribution of Z. For example, in Section 5, $s_\bullet$ minimizes the mean square approximation error

$$\min_{s \in \mathscr{S}} \mathrm{E}[s_0(Z) - s(Z)]^2 = \mathrm{E}[s_0(Z) - s_\bullet(Z)]^2,$$

but $s_\bullet$ above does not to be defined in this way.

Define the BLP parameter corresponding to $s_\bullet$ as:

$$\theta_\bullet = \frac{\mathrm{Cov}_Z(s_0(Z), s_\bullet(Z))}{\mathrm{Var}_Z s_\bullet(Z))}.$$

This is a fixed estimand.

If $R_{|A|} \to 0$ as $|A| \to \infty$, then S_A converges to the pseudo-true value $s_\bullet$. We call this property "pseudo"-consistency. The lemma shows that in this case, the random estimand θ_A approaches $\theta_\bullet$ at the rate $R_{|A|}$.

Lemma B.1 (Concentration from "Pseudo"-Consistency)**.** *Assume that $S_a \in \mathscr{S}$ for all $a \in \mathscr{A}$ and $s_\bullet \in \mathscr{S}$, that the elements of $\mathscr{S}$ and s_0 are all bounded above by a finite constant K, and that* $\operatorname{Var}_Z S(Z)$ *is bounded below by a positive constant $k > 0$ for all $S \in \mathscr{S}$. Then*

$$\mathrm{E}|\theta_A - \theta_\bullet| \leqslant C_{K,k}[R_{|A|} \wedge 1]$$

where $C_{K,k}$ is a numeric constant that only depends on K and k.

Concentration under Algorithmic Stability. On the other hand, algorithmic or statistical influence analysis often implies that

$$\mathrm{E}\mathrm{E}_Z(S_A(Z) - S_{A'}(Z))^2 \leqslant R'^2_{|A|},$$

where A and A' are independent uniform variables on $\mathscr{A}$. To explain the notion, let M and M' be the complements of A and A' relative to $\{1,...,N\}$. The symmetric difference between A and A' is $M \cap M'$. If the latter set is small in cardinality relative to the cardinality of A, then we would expect S_A and $S_{A'}$ not to differ if the machine producing S's is a smooth function of data. The definition above provides one way to measure this stability. We provide further discussion below.

By triangle inequality, the algorithmic stability can be bounded by estimation stability:

$$\sqrt{\mathrm{E}\mathrm{E}_Z(S_A(Z) - S_{A'}(Z))^2} \leqslant 2\sqrt{\mathrm{E}\mathrm{E}_Z(S_A(Z) - S_\bullet(Z))^2}$$

Therefore algorithmic stability is more general.

Lemma B.2 (Concentration from Algorithmic Stability)**.** *Suppose the assumptions of the previous lemma hold. Then if $R'_{|A|} \to 0$ as $|A| \to \infty$, then*

$$\mathrm{E}|\theta_A - \theta_{A'}| \leqslant C_{K,k}[R'_{|A|} \wedge 1],$$

where $C_{K,k}$ is a numeric constant that only depends on K and k.

Putting it Together: Concentration Around Median. The following result shows that the desired concentration condition holds if either estimation stability or algorithmic stability is strong enough.

Lemma B.3 (Stability of Median Target from Estimation or Algorithmic Stability)**.** *Suppose the assumptions of the previous lemma hold. Then*

$$\mathrm{E}|\theta_A - \theta^*| \leqslant \mathrm{E}|\theta_A - \theta_{A'}| \wedge \mathrm{E}|\theta_A - \theta_\bullet|.$$

Therefore, if

$$\sqrt{n} C_{K,k}[R'_{|A|} \wedge R_{|A|}] \leqslant \delta_N \tag{B.1}$$

for $\delta_N \to 0$ as $N \to \infty$, then

$$\mathrm{P}(\sqrt{n}|\widehat{\theta}_A - \theta^*| > \sqrt{\delta_N}) \leqslant \sqrt{\delta_N}.$$

The latter implies the condition we want provided $\widehat{\sigma}_A/\sqrt{n}+\sqrt{n}/\widehat{\sigma}_A = O_P(1)$.

We conclude here with some comparisons of the two notions of stability. Estimation stability readily follows from the available statistical learning theory. In particular $R^2_{|A|}$ scales like $d/|A|$ where d is the intrinsic dimension of the function class $\mathscr{S}$, as we discussed in Section 5. Therefore, n needs to be much smaller than $d(N-n)$ to satisfy the last condition of the last lemma.

Algorithmic stability does not require estimation stability, even though the latter property seems quite mild. On the other hand, its characterizations are not well-studied and are much less available. See Chernozhukov et al. (2021b) for analysis of constrained Lasso and Ridge that is applicable here; see also Chen et al. (2022) for leave-one-out stability analysis for bagged estimators over the subsamples (this analysis requires extension to the present framework).

It is useful to give a simple example to compare the two measures of stability. If $S_A(Z)$'s are generated by linear least squares $Z'\widehat{\beta}_A$ with $d = \dim(Z)$, then we have a crude upper bound on the algorithmic stability bound R'^2_A scaling like $nd/(N-n)^2$. This is generally smaller than R_A scaling like $d/(N-n)$. It implies a weaker same qualitative requirement on n: n needs to be smaller than $\sqrt{d}(N-n)$ to satisfy the condition (B.1) of Lemma B.3.

Proof of Lemmas B.1- B.3. To show Lemma B.1, it is convenient to define $f^o(Z) = f(Z) - \mathrm{E}_Z f(Z)$. Then, using the boundedness assumption we have

$$\begin{aligned}
|\mathrm{Cov}_Z(s_0(Z),S_A(Z)) - \mathrm{Cov}_Z(s_0(Z),s_\bullet(Z))| &= |\mathrm{E}_Z[s_0^o(Z)S_A(Z)] - \mathrm{E}[s_0^o(Z)s_\bullet(Z)]| \\
&\leqslant K\mathrm{E}_Z|S_A(Z)-s_\bullet(Z))|,
\end{aligned}$$

$$\begin{aligned}
|\mathrm{Var}_Z(S_A(Z)) - \mathrm{Var}_Z(s_\bullet(Z))| &= |\mathrm{E}_Z(S_A^o(Z))^2 - \mathrm{E}_Z(s_\bullet^o(Z))^2| \\
&\leqslant \mathrm{E}_Z|S_A^o(Z)+s_\bullet^o(Z))||S_A^o(Z)-s_\bullet^o(Z))| \\
&\leqslant 2K\mathrm{E}_Z|S_A^o(Z)-s_\bullet^o(Z))|.
\end{aligned}$$

Then using elementary inequalities and boundedness assumptions conclude

$$|\theta_A - \theta_\bullet| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}_Z|S_A(Z) - s_\bullet(Z)|$$

Taking expectation over A,

$$\mathrm{E}|\theta_A - \theta_\bullet| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}\mathrm{E}_Z|S_A(Z) - s_\bullet(Z))| \leqslant C_{K,k}R_{|A|},$$

where the last inequality follows from the norm inequality.

Lemma B.2 follows analogously, replacing $s_\bullet$ with $S_{A'}$ to obtain

$$|\theta_A - \theta_{A'}| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}_Z|S_A(Z) - S_{A'}(Z)|$$

Taking expectation over (A,A'), we obtain:

$$\mathrm{E}|\theta_A - \theta_{A'}| \leqslant (k^{-1}K + 2k^{-2}K^2)\mathrm{E}\mathrm{E}_Z|S_A(Z) - S_{A'}(Z))| \leqslant C_{K,k}R'_{|A|},$$

where the last inequality follows from the norm inequality.

To show Lemma B.3, we note that

$$\begin{aligned} \mathrm{E}|\theta_A - \theta^*| &= \mathrm{E}\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} |\theta_a - \theta^*| \mid \mathrm{Data}\right] \\ &\leqslant \mathrm{E}\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} |\theta_a - \theta_\bullet| \mid \mathrm{Data}\right] = \mathrm{E}|\theta_A - \theta_\bullet|, \end{aligned}$$

where the first property holds by the law of iterated expectation and by $A \sim U(\mathscr{A})$ independently of the Data, the inequality holds by definition of θ^* as the median of the sample $\{\theta_a : a \in \mathscr{A}\}$, and the last equality holds by iterating expectations again.

Similarly, we note

$$\begin{aligned} \mathrm{E}|\theta_A - \theta^*| &= \mathrm{E}\mathrm{E}\left[\frac{1}{\mathscr{A}}\sum_{a\in\mathscr{A}} |\theta_a - \theta^*| \mid \mathrm{Data}\right] \\ &= \mathrm{E}\frac{1}{|\mathscr{A}|}\sum_{a'\in\mathscr{A}} \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} |\theta_a - \theta^*| \mid \mathrm{Data}\right] \\ &\leqslant \mathrm{E}\frac{1}{|\mathscr{A}|}\sum_{a'\in\mathscr{A}} \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} |\theta_a - \theta_{a'}| \mid \mathrm{Data}\right] \\ &= \mathrm{E}\left[\frac{1}{|\mathscr{A}|}\sum_{a'\in\mathscr{A}} \frac{1}{|\mathscr{A}|}\sum_{a\in\mathscr{A}} |\theta_a - \theta_{a'}| \mid \mathrm{Data}\right] = \mathrm{E}|\theta_A - \theta_{A'}|, \end{aligned}$$

where the first property holds by the law of iterated expectation and by $A \sim U(\mathscr{A})$ independently of the Data, the inequality holds by definition of θ^* as the median of $\{\theta_A : a \in \mathscr{A}\}$, the last equality holds by iterating expectations and independence of A and A'.

Finally, the second claim of the Lemma follows by the Markov inequality. ■

APPENDIX C. PROOFS OF SECTION 5

Proof of Theorem 5.1. For the objective (B), write $Y = b_0(Z) + Ds_0(Z) + \varepsilon$, where $\mathrm{E}[\varepsilon \mid D, Z] = 0$. Then

$$YH = \{Hb_0(Z) + (HD-1)s_0(Z)\} + s_0(Z) + \varepsilon H,$$

where the first term can be expressed as:

$$\{Hb_0(Z) + (HD-1)s_0(Z)\} = H(b_0(Z) + (1-p(Z))s_0(Z)) = H\bar{b}_0(Z).$$

So that we can decompose:

$$YH - b(Z)H - s(Z) = \{H(\bar{b}_0(Z) - b(Z))\} + \{s_0(Z) - s(Z)\} + \varepsilon H.$$

Then the result follows taking the square and expectation, by using (i) orthogonality of the three terms in the decomposition above:

$$\mathrm{E}[\varepsilon H^2(\bar{b}_0(Z)-b(Z))]=0,\ \ \mathrm{E}[\varepsilon H(s_0(Z)-s(Z))]=0,\ \ \mathrm{E}[H(\bar{b}_0(Z)-b(Z))(s_0(Z)-s(Z)]=0,$$

where the last relation follows from $\mathrm{E}[H \mid Z]=0$, and (ii) also noting that $\mathrm{E}[H^2|Z]=w(Z)$.

For the objective (A), write similarly,

$$Y-b(Z)-(D-p(Z))s(Z)=[\tilde{b}_0(Z)-b(Z)]+[D-p(z)](s_0(Z)-s(Z))+\varepsilon,$$

and then conclude that the three terms are orthogonal to each other. The result follows by completing the square and taking expectation, where we also observe that

$$\mathrm{E}w(Z)(D-p(Z))^2(s_0(Z)-s(Z))^2=\mathrm{E}(s_0(Z)-s(Z))^2,$$

since $\mathrm{E}[w(Z)(D-p(Z))^2 \mid Z]=1$. ■

Proof of Theorem 5.2. We demonstrate the result for type B loss; the demonstration for type A follows similarly. Application of Theorem 3 of Liang et al. (2015) gives the following bound on the excess risk $\mathscr{R}$ of the estimator (B,S):

$$0 \leqslant \mathscr{R}:=\mathrm{E}[YH-B(Z)H-S(Z)]^2-\mathrm{E}[YH-b_\bullet(Z)H-s_\bullet(Z)]^2 \leqslant C_K\mathscr{R}^o(A,\mathscr{H},c_K),$$

where $(b_\bullet,s_\bullet)$ minimize $\mathrm{E}[YH-b_\bullet(Z)H-s_\bullet(Z)]^2$ over $b\in\mathscr{B}$ and $s\in\mathscr{S}$, and C_K and c_K are positive constants that only depend on K, and $\mathscr{H}:=4(H\mathscr{B}+\mathscr{S})$. Theorem 5.1 then implies that

$$\begin{aligned}\mathscr{R} &= \mathrm{E}[s_0(Z)-S(Z)]^2-\mathrm{E}[s_0(Z)-s_\bullet(Z)]^2\\ &+ \mathrm{E}[w(Z)(\bar{b}_0(Z)-B(Z))]^2-\mathrm{E}[w(Z)(\bar{b}_0(Z)-b_\bullet(Z))]^2,\end{aligned}$$

where the second term is non-negative. Therefore,

$$\mathrm{E}[s_0(Z)-S(Z)]^2-\mathrm{E}[s_0(Z)-s_\bullet(Z)]^2 \leqslant C_K\mathscr{R}^o(A,\mathscr{H},c_K).$$

The lower bound

$$\mathrm{E}[s_0(Z)-S(Z)]^2-\mathrm{E}[s_0(Z)-s_\bullet(Z)]^2 \geqslant \mathrm{E}[S(Z)-s_\bullet(Z)]^2$$

follows from Pythagorian inequality for obtuse triangles and the fact that $s_\bullet$ minimizes $\mathrm{E}[s_0(Z)-S(Z)]^2$ over the convex set $\mathscr{S}$. ■

Appendix D. Gaussian Approximation for Split-Sample Least Squares Uniformly Over Convex Sets and in P.

We present a set-up that covers not only the split-sample least square estimators of the main text, but also other potential cases of interest. Let W denote a generic data vector. All the linear regressions or mean estimators used on the main sample M could be viewed as ordinary least squares with a suitable definition of W.

Throughout we assume that $\{(W_i)\}_{i=1}^N$ are i.i.d. copies of vector W that has law P. We abbreviate $(\mathscr{D}_A, \mathscr{D}_M) := (\mathrm{Data}_A, \mathrm{Data}_M)$. There is a learning algorithm that inputs $\mathscr{D}_A$ and outputs a map $f(\cdot; \mathscr{D}_A)$, which maps the support of W to $\mathbb{R}^{d+1}$ for a fixed d. This map defines the split-specific outcome and regressors:

$$(Y_{A,i}, X_{A,i}) = f(W_i; \mathscr{D}_A), \quad i \in M.$$

Let $\widehat{\beta}_A$ be a solution to $\mathbb{E}_{N,M}[X_{A,i}\widehat{\varepsilon}_{A,i}] = 0$ for $\widehat{\varepsilon}_{A,i} = Y_{A,i} - X'_{A,i}\widehat{\beta}_A$. Let $\widehat{V}_A$ denote the Eicker-Huber-White sandwich

$$\widehat{V}_A := (\mathbb{E}_{N,M} X_{A,i}X'_{A,i})^{-1}\mathbb{E}_{N,M}\widehat{\varepsilon}^2_{A,i}X_{A,i}X'_{A,i}(\mathbb{E}_{N,M}X_{A,i}X'_{A,i})^{-1},$$

whenever it exists.

Fix some positive finite constants c and C. Let β_A denote a solution to $\mathrm{E}[X_A\varepsilon_A] = 0$, for $\varepsilon_A = Y_A - X'_A\beta_A$, if it exists. And let

$$V_A := (\mathrm{E}_P[X_AX'_A \mid \mathscr{D}_A])^{-1}\mathrm{E}_P[\varepsilon^2_A X_AX'_A \mid \mathscr{D}_A](\mathrm{E}_P[X_AX'_A \mid \mathscr{D}_A])^{-1},$$

if it exists. Let $\mathscr{E}_{A,N}$ be the event that

$$\mathrm{E}_P|Y_A|^{4+\delta} + \mathrm{E}_P[\|X_A\|^{4+\delta} \mid \mathscr{D}_A] \leqslant C, \quad \min_{\|a\|=1} \mathrm{E}_P[(a'X_A)^2 \mid \mathscr{D}_A] > c.$$

On this event β_A and ε_A are well defined. Let $\mathscr{E}'_{A,N} \subset \mathscr{E}_{A,N}$ be the event such that

$$\min_{\|a\|=1} \mathrm{E}_P[(\varepsilon_A a'X_A)^2 \mid \mathscr{D}_A] > c.$$

On this event V_N is well-defined. Let $CS(\mathbb{R}^d)$ denote the collection of the convex sets in $\mathbb{R}^d$.

We observe that, by the i.i.d. sampling and $A \sim U(\mathscr{A})$ independently of Data, $(\mathscr{D}_A, \mathscr{D}_M)$ has the same distribution as $(\mathscr{D}_a, \mathscr{D}_m)$, for a fixed partition (a, m). This is an exchangeability property. Therefore, we can fix (A, M) to be a fixed partition $\{a, m\}$ in what follows. Moreover $(X_{a,i}, Y_{a,i})_{i=1}^{N-m}$ are i.i.d. conditional on $\mathscr{D}_a$. These observations simplify the verification of the following result.

Lemma D.1 (Gaussian Approximation)**.** *Using the setup above, let γ_N be a sequence of positive constants tending to zero. Suppose that for all $P \in \mathscr{P}$, we have $\mathrm{P}_P(\mathscr{E}'_{N,A}) \geqslant 1 - \gamma_N$. Then, uniformly in $P \in \mathscr{P}$, as $(n, N) \to \infty$:*

$$\sup_{R \in CS(\mathbb{R}^d)} \left| \mathrm{P}_P[\widehat{V}_A^{-1/2}(\widehat{\beta}_A - \beta_A) \in R \mid \mathscr{D}_A] - \mathrm{P}(N(0, I_d) \in R) \right| \xrightarrow{\mathrm{P}_P} 0,$$

$$\sup_{R \in CS(\mathbb{R}^d)} \left| \mathrm{P}_P[\widehat{V}_A^{-1/2}(\widehat{\beta}_A - \beta_A) \in R] - \mathrm{P}(N(0, I_d) \in R) \right| \longrightarrow 0,$$

and the same results hold with $\widehat{V}_A$ replaced by V_A; moreover, $\widehat{V}_N V_N^{-1} \to_{\mathrm{P}_P} I$ both conditional on $\mathscr{D}_N$ and unconditionally.

Proof of Lemma D.1. It suffices to demonstrate the argument for an arbitrary sequence $\{P_N\}$ in $\mathscr{P}$. Let

$$\widehat{t}_A := \widehat{V}_A^{-1/2}(\widehat{\beta}_A - \beta_A), \; t_A := V_A^{-1/2}(\widehat{\beta}_A^o - \beta_A), \; \widehat{\beta}_A^o := [\mathrm{E} X_A X_A']^{-1} \mathbb{E}_{N,M} X_A Y_A.$$

Consider the event $\mathscr{E}_{N,A}'' \subseteq \mathscr{E}_{N,A}'$ such that:

$$\mathscr{E}_{N,A}'' = \left\{ (\widehat{t}_A, t_A, \widehat{V}_N) \text{ exist and } \|\widehat{t}_A - t_A\| + \|\widehat{V}_N - V_N\| \leqslant r_N \right\}.$$

It follows from the standard arguments for asymptotic theory for least squares under i.i.d. sampling of data arrays $(Y_{A,i}, X_{A,i})_{i=1}^N$, e.g. Gallant and White (1988), that there exists a sequence of positive constants $\{r_N, \delta_N\} \searrow 0$ such that $\mathrm{P}\left(\mathscr{E}_{N,A}'' \mid \mathscr{D}_A\right) \geqslant 1 - \delta_N$ on the event $\mathscr{E}_{N,A}'$. Therefore by the union bound

$$\mathrm{P}\left(\mathscr{E}_{N,A}''\right) \geqslant 1 - \delta_N - \gamma_N, \tag{D.1}$$

for γ_N defined in the statement of the lemma. For $r > 0$ let $R^r = \{x \in \mathbb{R}^d : d(x, R) \geqslant r\}$ and $R^{-r} = \{x \in R : d(x, \mathbb{R}^d \setminus R) \geqslant r\}$, where $d(x, R) := \min_{x' \in R} \|x' - x\|$. Note that R^{-r} can be an empty set. Then, on the event $\mathscr{E}_{N,A}''$,

$$\begin{aligned} \mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) &\geqslant \mathrm{P}\left(t_A \in R^{-r_N} \mid \mathscr{D}_A\right) \\ &\geqslant \mathrm{P}(N(0, I_d) \in R^{-r_N}) - B_N d^{1/4}/\sqrt{n}, \\ &\geqslant \mathrm{P}(N(0, I_d) \in R) - 4 d^{1/4} r_N - B_N d^{1/4}/\sqrt{n}, \end{aligned}$$

where $B_N = C' \mathrm{E}[\|V_N^{-1/2} X_A \varepsilon_A\|^3 \mid \mathscr{D}_A]$, where C' is a numerical constant. The second inequality follows by the Bentkus bounds (Bentkus, 2003; Raič, 2019), which extend the Berry-Essen bounds to the multidimensional case, and the last inequality follows from the Ball's reverse isoperimetric inequality of the standard Gaussian vector (Ball, 1991). It follows similarly that

$$\begin{aligned} \mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) &\leqslant \mathrm{P}\left(t_A \in R^{r_N} \mid \mathscr{D}_A\right) \\ &\leqslant \mathrm{P}(N(0, I_d) \in R^{r_N}) + B_N/\sqrt{n}, \\ &\leqslant \mathrm{P}(N(0, I_d) \in R) + 4 d^{1/4} r_N + B_N d^{1/4}/\sqrt{n}. \end{aligned}$$

Since R above is arbitrary convex subset of $\mathbb{R}^d$, we have that on the event $\mathscr{E}''_{N,A}$:

$$\begin{aligned}
&\sup_{R \in CS(\mathbb{R}^d)} \left|\mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) - \mathrm{P}(N(0,I_d) \in R)\right| \\
&\leqslant \sup_{R \in CS(\mathbb{R}^d)} \left|\mathrm{P}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) - \mathrm{P}(N(0,I_d) \in R)\right| \\
&\leqslant 4d^{1/4} r_N + d^{1/4} B_N/\sqrt{n}.
\end{aligned}$$

Using Holder inequalities, we can check that $B_N \leqslant B$ on the event $\mathscr{E}''_N$, for some constant B that depends only on (c, C, d, δ). The first claim follows combining this inequality with (D.1).

To show that second claim note that

$$\begin{aligned}
&\sup_{R \in CS(\mathbb{R}^d)} \left|\mathrm{EP}\left(\widehat{t}_A \in R \mid \mathscr{D}_A\right) - \mathrm{P}(N(0,I_d) \in R)\right| \\
&\leqslant 4d^{1/4} r_N + d^{1/4} B_N/\sqrt{n} + (1 - \mathrm{P}(\mathscr{E}''_{N,A})) \leqslant 4d^{1/4} B_N/\sqrt{n} + \gamma_N + \delta_N.
\end{aligned}$$

Finally, $\|\widehat{V}_N V_N^{-1} - I\| \leqslant \|V_N^{-1}\| r_N \leqslant c r_N$ conditional on $\mathscr{D}_A$ and on the event $\mathscr{E}''_N$. The conditional convergence claim follows from (D.1). It then follows that $\mathrm{EP}(\|\widehat{V}_N V_N^{-1} - I\| \leqslant c r_N \mid \mathscr{D}_A) \geqslant \mathrm{P}(\mathscr{E}_{N,A}) \geqslant 1 - \delta_N - \gamma_N$. ■