Conformal Inference of Counterfactuals and Individual Treatment Effects

Lihua Lei[1] and Emmanuel J. Candès[1,2]

[1]Department of Statistics, Stanford University, Stanford, CA 94305
[2]Department of Mathematics, Stanford University, Stanford, CA 94305

June 12, 2020

Abstract

Evaluating treatment effect heterogeneity widely informs treatment decision making. At the moment, much emphasis is placed on the estimation of the conditional average treatment effect via flexible machine learning algorithms. While these methods enjoy some theoretical appeal in terms of consistency and convergence rates, they generally perform poorly in terms of uncertainty quantification. This is troubling since assessing risk is crucial for reliable decision-making in sensitive and uncertain environments. In this work, we propose a conformal inference-based approach that can produce reliable interval estimates for counterfactuals and individual treatment effects under the potential outcome framework. For completely randomized or stratified randomized experiments with perfect compliance, the intervals have guaranteed average coverage in finite samples regardless of the unknown data generating mechanism. For randomized experiments with ignorable compliance and general observational studies obeying the strong ignorability assumption, the intervals satisfy a doubly robust property which states the following: the average coverage is approximately controlled if either the propensity score or the conditional quantiles of potential outcomes can be estimated accurately. Numerical studies on both synthetic and real datasets empirically demonstrate that existing methods suffer from a significant coverage deficit even in simple models. In contrast, our methods achieve the desired coverage with reasonably short intervals.

1 From Average Effects To Individual Effects

Estimating the average treatment effect (ATE) for a population of interest has been a main focus of a rich literature on causal inference, and the last decades have seen the development of extensive statistical theory addressing important issues such as identification, estimation, and uncertainty quantification [e.g. Rubin, 1974, Pearl, 1995]. That said, the average effect only provides a coarse summary of the distribution of a treatment effect, and may be insufficient, or even misleading, to validate an intervention. Imagine, for instance, that a drug cures 70% of the patients while it makes the symptoms worse for the remaing 30%. Clearly, the drug should not be approved in spite of the positive average effect. This is not an artificial example; trials typically do not come out black and white and examples of this kind are frequent. On May 31, 2018, the National Academy of Medicine (NAM) held a workshop to discuss approaches of examining individual treatment effects (ITE) to support individualized patient care. In conjunction with this event, NAM published a report highlighting the importance of ITE in medicine. We quote from this document:

> "The individuality of the patient should be at the core of every treatment decision. One-size-fits-all approaches to treating medical conditions are inadequate; instead, treatments should be tailored to individuals based on heterogeneity of clinical characteristics and their personal preferences."

Outside of medical science, treatment effect heterogeneity is also of great concern to political scientists [Imai and Strauss, 2011, Grimmer et al., 2017], psychologists [Bolger et al., 2019, Winkelbeiner et al., 2019], sociologists [Xie et al., 2012, Breen et al., 2015], economists [Florens et al., 2008, Djebbari and Smith, 2008], and education researchers [Morgan, 2001, Brand and Xie, 2010]. In short, there is a wide range of fields that would benefit from a better understanding of ITE.

To move past the average treatment effect as the object of inference, most existing works have targeted, instead, the *conditional average treatment effects* (CATE)—although we will introduce a formal definition later in Section 2, this is the expectation of the ITE conditional on the values of the covariates. While CATE naturally provide a richer summary than ATE, they surely still neglect the inherent variability in the response (the conditional variance, if you will) which might be crucial for decision-making; unless, of course, the covariates explain away most of the variation in ITE. As a consequence, two types of variability are of immediate concern: (1) the variability of the response around the regression function and, (2) the variability of CATE estimators due to finite samples. Getting both (1) and (2) under control requires having sufficiently many covariates to explain away a significant fraction of the variability in the response, and at the same time, a nearly perfect estimate of CATE for every value of the covariates. Neither of these seem to be realistic for everyday causal inference problems.

In areas like medical science or public policy, point estimates are insufficient to inform decisions due to the huge loss potentially incurred by wrong actions. A confidence interval, or at least a p-value, is required by the U.S. Food and Drug Administration to approve a drug, in order to guarantee sufficient evidence and confidence in favor of the drug. The issues are that despite the importance of uncertainty quantification, the reliability of modern machine learning methods is typically under-studied, and that theoretical guarantees are hard to come by. For instance, existing theory usually requires strong non-verifiable assumptions and asymptotic regimes one does not encounter in practice. This of course limits the applications of machine learning methods in sensitive causal inference problems. In addition, we will demonstrate later that the confidence intervals for CATE or ITE produced by frequently discussed methods (including Bayesian methods) typically have unsatisfactory or unacceptable coverage, even in very simple and smooth models with at most ten covariates.

In this work, we leverage ideas from conformal inference to construct valid confidence intervals for individual treatment effects under the potential outcome framework [Neyman, 1923/1990, Rubin, 1974]. In particular, we address two challenges:

(1) Construct intervals for ITE with reliable coverage for subjects in the study, for which one of the potential oucomes is missing;

(2) Construct intervals for ITE with reliable coverage for subjects not in the study, and for which both potential oucomes are missing.

To be sure, in the case where a method is able to estimate ITE accurately, there is no need to distinguish these two tasks. In practice, however, this is rarely possible due to insufficient sample sizes, model misspecification, and inherent variability. We will thus see that the second challenge is generally far more ambitious since two potential outcomes are never observed simultaneously so that one cannot model them jointly.

In brief, Section 3 shows how the first challenge reduces to counterfactual inference since one potential outcome is observed for each subject; here, an interval for the missing outcome can be shifted into an interval for the individual treatment effect by contrasting it with the observed outcome. This section also introduces methods with guaranteed coverage even under model misspecification. In particular, for completely randomized or stratified randomized trials with perfect compliance so that the propensity score is known but may not be constant, our method achieves coverage in finite samples *without any assumption other than that of operating on i.i.d. samples.* For general observational studies under the strong ignorability assumption [Rubin, 1974], or randomized experiments with ignorable compliance, our methods have guaranteed coverage provided that either the outcome model or the treatment model is accurately estimated. This is analogous to the *doubly robust* property applicable to average treatment effect [e.g. Robins et al., 1994, Kang and Schafer, 2007] which speaks to the consistency of point estimates.

Having addressed the first challenge, we will see in Section 4 that our methods can serve as a stepping stone for the second. A naive approach, here, would be to apply the counterfactual inference on both potential outcomes and contrast the two intervals to induce an interval for ITE. Such an approach would entirely decouple the pair of potential outcomes and would likely be overly conservative. Instead, we introduce a less conservative approach which can exploit any solution to the first task. It applies the counterfactual inference to generate intervals for ITE applicable to subjects in the dataset as an intermediate step, and trains a model to generalize these intervals to subjects not in the study.

Finally, our methods easily extend to a widely studied problem which goes by the name of generalizability, or transportability, or external validity, and which concerns settings in which there is a distributional shift between the target population and the study population [e.g. Stuart et al., 2011, Tipton, 2014], see Sections 2–4. Furthermore, as explained in Section 5, our methods also naturally adapt to other causal inference frameworks, such as causal diagrams [Pearl, 1995] and invariant prediction [Peters et al., 2016].

2 From Point Estimates To Interval Estimates

2.1 Problem setup

Throughout the paper we focus on the potential outcome framework [Neyman, 1923/1990, Rubin, 1974] with a binary treatment. Extensions to other causal inference frameworks are discussed in Section 5. Given n subjects, denote by $T_i \in \{0, 1\}$ the binary treatment indicator, by $(Y_i(1), Y_i(0))$ the pair of potential outcomes and by X_i the vector of other covariates. We assume that

$$(Y_i(1), Y_i(0), T_i, X_i) \overset{i.i.d.}{\sim} (Y(1), Y(0), T, X),$$

where $(Y(1), Y(0), T, X)$ denotes a generic random vector. Under the *stable unit treatment value assumption* (SUTVA) commonly assumed in the literature [Rubin, 1990], the observed dataset comprises triples $(Y_i^{\text{obs}}, T_i, X_i)$ where

$$Y_i^{\text{obs}} = \begin{cases} Y_i(1), & T_i = 1, \\ Y_i(0), & T_i = 0. \end{cases}$$

The individual treatment effect τ_i is defined as

$$\tau_i \triangleq Y_i(1) - Y_i(0). \tag{2.1}$$

By definition, only one potential outcome is observed for every unit while the other is missing. Therefore, the ITE are unobserved and have to be inferred. Throughout the paper we assume the strong ignorability:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X. \tag{2.2}$$

Strong ignorability rules out any source of unmeasured confounders, which affect both the treatment assignment and the potential outcomes. Under this assumption, the treatment assignment is purely randomized conditional on any covariate values. Although ignorability is a strong assumption, it is a widely used starting point to formulate statistical theory and methodology [Rubin, 1978, Rosenbaum and Rubin, 1983, Imbens and Rubin, 2015].

2.2 Traditional inferential targets

As mentioned earlier, existing methods mostly focus on CATE, defined as

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]. \tag{2.3}$$

The CATE function $\tau(\cdot)$ can be expressed as $\tau(x) = m_1(x) - m_0(x)$, where

$$m_1(x) = \mathbb{E}[Y(1) \mid X = x] \text{ and } m_0(x) = \mathbb{E}[Y(0) \mid X = x].$$

It is standard in the literature to impose modeling restrictions on $m_1(x)$ and $m_0(x)$ (or equivalently $\tau(x)$ and $m_0(x)$) and estimate these functions using parametric or nonparametric techniques. Under the strong ignorability assumption, a standard argument shows that

$$m_1(x) = \mathbb{E}[Y^{\text{obs}} \mid X = x, T = 1] \text{ and } m_0(x) = \mathbb{E}[Y^{\text{obs}} \mid X = x, T = 0].$$

Thus, the estimation problem reduces to that of estimating a regression function. However, drawing reliable confidence bands—which can be trusted in practice—from finite samples around estimates $\hat{m}_1(\cdot)$ and/or $\hat{m}_0(\cdot)$ is delicate, to say the least. As we will show later, many standard techniques, e.g.

normal approximation and resampling techniques, may significantly underestimate the variability of such estimates.

An alternative estimand is the conditional quantile treatment effect (CQTE), which happens to be far less investigated in the literature [e.g. Fort, 2016]. Instead of contrasting the mean functions $m_1(x)$ and $m_0(x)$, CQTE is defined as the difference between the β-th quantiles of the distributions of $Y(1)$ and $Y(0)$ conditional on $X = x$ for a given β of interest. CQTE is to be distinguished from the quantiles of $Y(1)-Y(0)$; they are not the same at all! It turns out that the conditional quantiles of ITE are unidentifiable in general since they involve the joint distribution of $(Y(0), Y(1))$ and that we can never observe joint outcomes. Furthermore, the difficulties associated with uncertainty quantification still persist.

2.3 Coverage of interval estimates

In this work, we take counterfactuals $(Y_i(1), Y_i(0))$'s and the ITE τ_i's as objects of inference, and attempt to construct predictive intervals covering these random variables. Taking the potential outcome $Y(1)$ as an example, we wish to construct intervals $\hat{C}_1(x)$, which depend on the location in covariate space, and obey

$$\mathbb{P}(Y(1) \notin \hat{C}_1(X)) \leq \alpha, \tag{2.4}$$

for a pre-specified level α. Similarly, we seek $\hat{C}_0(x)$ for $Y(0)$ and $\hat{C}_{\mathrm{ITE}}(x)$ for $Y(1) - Y(0)$ obeying marginal coverage in the same sense, i.e.

$$\mathbb{P}(Y(1) - Y(0) \notin \hat{C}_{\mathrm{ITE}}(X)) \leq \alpha. \tag{2.5}$$

If we had perfect knowledge of the the quantiles $q_\beta(x)$ of $Y(1)$ given $X = x$ for each $\beta \in (0, 1)$, then the oracle estimate,

$$C_1(x) = [q_{\alpha/2}(x), q_{1-\alpha/2}(x)],$$

would automatically satisfy (2.4). This is arguably the best confidence interval one could produce. In addition, coverage would hold conditionally on $X = x$. In reality, conditional quantiles may be hard to estimate due to the limited effective sample size and imperfect model knowledge. In this case, substituting the true quantiles in $C_1(x)$ with estimates may fail to yield valid coverage.

A typical objection to the criterion (2.4) is that it only controls coverage in an average (marginal) sense—just as the root mean squared error (RMSE) measures *average* performance. Admittedly, it does not say much about the validity of the predicted range for a patient with *this* x. Without modeling assumptions, it is known to be impossible to construct non-trivial predictive intervals with *guaranteed* conditional coverage [Barber et al., 2019b]. This does not mean that conditional coverage cannot be achieved in any particular application; in fact, we make conditional coverage a focus point of this work and demonstrate reasonable approximations.

2.4 General coverage criteria

In classical causal inference, it is often argued that the average treatment effect on the treated (ATT), defined as $\mathbb{E}[Y(1)-Y(0) \mid T = 1]$, is more meaningful than the average treatment effect (ATE), defined as $\mathbb{E}[Y(1)-Y(0)]$, because the treatment effect for treated units is more important than that for control units. Analogously, in such settings, the coverage of treated units is arguably more important than that of control units. The criterion (2.4) can be then modified as

$$\mathbb{P}(Y(t) \in \hat{C}_t(X) \mid T = 1) \geq 1 - \alpha, \quad (t = 0, 1). \tag{2.6}$$

Compared to (2.4), (2.6) substitutes the marginal distribution of X with the conditional distribution of X given $T = 1$ (recall that we operate under the strong ignorability assumption). Similarly we may consider the counterpart of the average treatment effect on the controls (ATC) by conditioning on $T = 0$. These considerations motivate the following general criterion:

$$\mathbb{P}_{(X,Y(t))\sim Q_X \times P_{Y(t)|X}}(Y(t) \in \hat{C}_t(X)) \geq 1 - \alpha, \quad (t = 0, 1). \tag{2.7}$$

For instance, (2.6) is a special case with $Q_X = P_{X|T=1}$. The general formulation (2.7) is useful when the study population differ from the target population. In this case, Q_X can be chosen to be the covariate distribution in the target population. Inference on ATE and CATE in this setting has been a subject of much recent research, and is known under the name of generalizability, or transportability, or external validity [e.g. Tipton, 2013, Pearl and Bareinboim, 2014].

3 From Observables To Counterfactuals

3.1 Counterfactuals and covariate shift

Counterfactual inference is both the ultimate goal in areas such as policy evaluation [e.g. Athey et al., 2017, Ben-Michael et al., 2018, Arkhangelsky et al., 2018] as well as the stepping stone for inferring ITE in general. We construct predictive intervals for $Y(1)$ and $Y(0)$ using the i.i.d. observations $(Y_i^{\mathrm{obs}}, T_i, X_i)$, and make no assumption other than the strong ignorability assumption. Therefore, only the samples in the treatment (resp. control) group are useful for constructing $\hat{C}_1(x)$ (resp. $\hat{C}_0(x)$).

Under the strong ignorability assumption, the joint distribution of (X, Y^{obs}) of the observed treated samples is given by

$$P_{X|T=1} \times P_{Y(1)|X}.$$

Once again, we want to achieve (2.7) under the target distribution $Q_X \times P_{Y(1)|X}$ on the basis of an i.i.d. sample drawn from $P_{X|T=1} \times P_{Y(1)|X}$. These two distributions share the same conditional distribution $P_{Y(1)|X}$ of the outcome but otherwise differ in the distribution of the covariates. Covariate shifts have been widely studied in the machine learning literature [e.g. Shimodaira, 2000], yet, the heavy focus there is on point estimates. For interval estimates, we shall rely on weighted conformal inference recently developed by Barber et al. [2019a].

3.2 Weighted conformal inference

Conformal inference was introduced by Vladimir Vovk and his collaborators [e.g. Vovk et al., 2005, Gammerman and Vovk, 2007, Shafer and Vovk, 2008, Vovk et al., 2009, Vovk, 2012, 2013, Balasubramanian et al., 2014, Vovk, 2015]. It later gained significant attention from the statistics community for regression problems [e.g. Lei et al., 2013, Lei and Wasserman, 2014, Lei et al., 2018, Barber et al., 2019b,c] and classification problems [e.g. Sadinle et al., 2019, Romano et al., 2020], spurring further developments. Given i.i.d. samples $(X_i, Y_i)_{i=1}^n$ drawn from a distribution $P_X \times P_{Y|X}$, conformal inference takes an arbitrary predictive model—such as an estimate of the conditional quantile—as input and calibrates it to produce a prediction set $\hat{C}(x)$ with guaranteed marginal coverage, such that

$$\mathbb{P}_{(X,Y)\sim P_X \times P_{Y|X}}(Y \in \hat{C}(X)) \geq 1 - \alpha. \tag{3.1}$$

For instance, with estimates $\hat{q}_{\alpha_{\text{lo}}}(x)$ and $\hat{q}_{\alpha_{\text{hi}}}(x)$ of the α_{lo}-th and α_{hi}-th conditional quantiles of $Y \mid X = x$, the conformal quantile regression (CQR) introduced in Romano et al. [2019] as a variant of the standard conformal inference, would produce an interval estimate of the form

$$\hat{C}(x) = [\hat{q}_{\alpha_{\text{lo}}}(x) - \eta, \hat{q}_{\alpha_{\text{hi}}}(x) + \eta]; \tag{3.2}$$

above, η is a data-driven constant computed in a particular way. In contrast to asymptopia—we refer here to classical asymptotic normality theory or asymptotic empricial process theory which characterizes the accuracy of certain resampling methods—CQR enjoys finite sample coverage guarantees without regard to the unknown joint distribution of X and Y.

Because our inference problem involves a potential covariate shift between the target distribution and the sampling distribution, we need to adjust the criterion (3.1) into

$$\mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}(Y \in \hat{C}(X)) \geq 1 - \alpha. \tag{3.3}$$

To address such a situation, Barber et al. [2019a] introduced a weighted variant of conformal inference achieving (3.3). This holds with the proviso that the the likelihood ratio $w(x) = dQ_X(x)/dP_X(x)$ is known, although the algorithm is shown to perform well when it is only estimated. When applied to CQR, the weighted interval estimate is still of the form (3.2), the difference being that the algorithm computing η incorporates information about the likelihood ratio $w(x)$. Algorithm 1 sketches split-CQR, a type of weighted conformal inference we shall use in this work.[1] In passing, it is worth mentioning that in this algorithm, η remains invariant if $w(x)$ is rescaled to become $c \cdot w(x)$ in which c is an arbitrary positive constant.

[1] Barber et al. [2019a] introduced weighted conformal inference and applied to conditional mean estimates [Lei et al., 2018, e.g.]. The extension to other conformal inference techniques such as CQR is straightforward.

Algorithm 1 Weighted split-CQR

```
Input: level $\alpha$, data $\mathcal{Z} = (X_i, Y_i)_{i \in \mathcal{I}}$, testing point $x$, function $\hat{q}_\beta(x; \mathcal{D})$ to fit $\beta$-th conditional
    quantile and function $\hat{w}(x; \mathcal{D})$ to fit the weight function at $x$ using $\mathcal{D}$ as data
Procedure:
    Randomly split $\mathcal{Z}$ into a training fold $\mathcal{Z}_{\mathrm{tr}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\mathrm{tr}}}$ and a calibration fold $\mathcal{Z}_{\mathrm{ca}} \triangleq (X_i, Y_i)_{i \in \mathcal{I}_{\mathrm{ca}}}$
    For each $i \in \mathcal{I}_{\mathrm{ca}}$, compute the non-conformity score $V_i = \max\{\hat{q}_{\alpha_{\mathrm{lo}}}(X_i; \mathcal{Z}_{\mathrm{tr}}) - Y_i, Y_i - \hat{q}_{\alpha_{\mathrm{hi}}}(X_i; \mathcal{Z}_{\mathrm{tr}})\}$
    For each $i \in \mathcal{I}_{\mathrm{ca}}$, compute the weight $W_i = \hat{w}(X_i; \mathcal{Z}_{\mathrm{tr}})$
    Compute the normalized weights $\hat{p}_i = \frac{W_i}{\sum_{i \in \mathcal{I}_{\mathrm{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\mathrm{tr}})}$ and $\hat{p}_\infty = \frac{\hat{w}(x; \mathcal{Z}_{\mathrm{tr}})}{\sum_{i \in \mathcal{I}_{\mathrm{ca}}} W_i + \hat{w}(x; \mathcal{Z}_{\mathrm{tr}})}$
    Compute $\eta$ as the $(1-\alpha)$-th quantile of the discrete distribution $\sum_{i \in \mathcal{I}_{\mathrm{ca}}} \hat{p}_i \delta_{V_i} + \hat{p}_\infty \delta_\infty$
Output: $\hat{C}(x) = [\hat{q}_{\alpha_{\mathrm{lo}}}(x; \mathcal{Z}_{\mathrm{tr}}) - \eta, \hat{q}_{\alpha_{\mathrm{hi}}}(x; \mathcal{Z}_{\mathrm{tr}}) + \eta]$
```

If the likelihood ratio $w(x)$ is known and finite almost surely, Barber et al. [2019a] prove that the interval $\hat{C}(x)$ from Algorithm 1 achieves (3.3). Moreover, in a companion paper [Lei and Candès, 2020], we show that the inequality (3.3) is almost an equality if the non-conformity scores have no ties and the covariate shift has a bounded χ^r-divergence defined as $\chi^r(Q_X \parallel P_X) = \int (dQ_X/dP_X)^r \, dP_X$.

Proposition 1. *Consider Algorithm 1 and assume* $(X_i, Y_i) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$ *and* $\hat{w}(x) = w(x)$.

(1) If $w(X) < \infty$ *almost surely, then Corollary 1 of Barber et al. [2019a] asserts that* (3.3) *holds.*

(2) If the non-conformity scores $\{V_i : i \in \mathcal{I}_{\mathrm{ca}}\}$ *have no ties almost surely and* $\chi^r(Q_X \parallel P_X) < \infty$ *for some* $r \geq 2$, *then*

$$1 - \alpha \leq \mathbb{P}_{(X,Y) \sim Q_X \times P_{Y|X}}(Y \in \hat{C}(X)) \leq 1 - \alpha + c\, n^{1/r - 1}.$$

Above, c *is a positive constant that only depends on* $\chi^r(Q_X \parallel P_X)$ *and* r.

Note that the proposition holds uniformly over all conditional distributions $P_{Y|X}$ and all procedures used to fit conditional quantiles. Because we express the dependence on P_X and Q_X only through their divergence, an immediate consequence is that when the likelihood ratio is bounded, we can take $r = \infty$ and the upper bound matches the rate for unweighted conformal inference. This essentially implies that weighted split-CQR has almost exact coverage when the calibration fold is large.

3.3 The role of the propensity score

Propensity scores were introduced by Rosenbaum and Rubin [1983] for the analysis of observational studies. Given a binary treatment, the propensity score $e(x)$ is defined as the probability of getting treated given the covariate value, i.e.

$$e(x) = \mathbb{P}(T = 1 \mid X = x).$$

With the full knowledge of $e(x)$, one can identify ATE/ATT/ATC using inverse propensity weighting (IPW) [Imbens and Rubin, 2015] without any assumption on potential outcomes other than some mild moment conditions. This holds with the proviso that certain overlap/positivity conditions hold.

Specifically, if the overlap condition $0 < e(X) < 1$ holds almost surely, ATE can be identified as follows:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[w_1(X)Y^{\text{obs}}I(T=1) - w_0(X)Y^{\text{obs}}I(T=0)\right], \tag{3.4}$$

where $w_1(x) = 1/e(x)$ and $w_0(x) = 1/(1-e(x))$. This holds under moment conditions on $1/e(X)$, $1/(1-e(X))$, $Y(1)$ and $Y(0)$. The overlap condition guarantees that $w_1(x)$ and $w_0(x)$ stay finite, and is necessary for identification through (3.4). Indeed, if $e(x) = 0$ for a subset of covariate values with non-zero probability, there will be no treated unit on that strata and hence the strata-wise ATE can never be identified without modeling assumptions on the potential outcomes. The overlap condition is as fundamental as the strong ignorability assumption in observational studies. We refer the readers to D'Amour et al. [2017] for an extensive discussion. Similarly, under the weaker overlap condition $e(X) < 1$ a.s., ATT can be identified through (3.4) with $w_1(x) = 1/\mathbb{P}(T=1)$ and $w_0(x) = e(x)/(1-e(x))\mathbb{P}(T=1)$ [Hirano et al., 2003].

The weights in IPW estimators are essentially calibrating the observed covariate distribution to the target one. This is similar in spirit to weighted conformal inference. For ATE-type conformal inference on $Y(1)$, a simple application of Bayes formula implies that

$$w_1(x) = \frac{dP_X(x)}{dP_{X|T=1}(x)} = \frac{\mathbb{P}(T=1)}{e(x)}.$$

Now recall that weighted conformal inference is invariant to rescaling of the likelihood ratio $w_1(x) \propto 1/e(x)$, so that we can simply ignore the numerator, in which case everything reduces to the inverse propensity score. Similarly, for inference on $Y(0)$, $w_0(x)$ can be chosen as $1/(1-e(x))$. The concurrence with IPW-type estimators is here unsurprising since the reweighting scheme is motivated by the covariate shift due to treatment selection in both settings.

For ATT-type conformal inference on $Y(1)$, $Q_X = P_{X|T=1}$. Thus, no weighting is needed and unweighted conformal inference may be applied. By contrast, the inference on $Y(0)$ still requires reweighting because $Q = P_{X|T=1} \neq P_{X|T=0}$. Applying Bayes formula,

$$w_0(x) = \frac{\mathbb{P}(T=0)}{\mathbb{P}(T=1)} \frac{e(x)}{1-e(x)}.$$

We can therefore choose $w_0(x) = e(x)/(1-e(x))$ per our previous discussion. This happens to coincide with the weights used by the IPW estimator for ATT.

In general, if Q_X is the covariate distribution in another population, as in the context of generalizability/transportability/external validity, then

$$w_1(x) = \frac{dQ_X(x)}{dP_{X|T=1}(x)} = \frac{dQ_X(x)}{dP_X(x)} \frac{\mathbb{P}(T=1)}{e(x)}.$$

In this case, we can choose $w_1(x) = (dQ_X/dP_X)(x)/e(x)$ to be the inverse propensity tilted likelihood ratio. Similarly $w_0(x) = (dQ_X/dP_X)(x)/(1-e(x))$. All these weight functions are displayed in Table 1. In sum, weighted conformal inference depends on propensity scores in the same way IPW estimation of average causal effects depends on these same scores.

Inferential type	ATE	ATT	ATC	General
$w_1(x)$	$1/e(x)$	1	$(1-e(x))/e(x)$	$(dQ/dP)(x)/e(x)$
$w_0(x)$	$1/(1-e(x))$	$e(x)/(1-e(x))$	1	$(dQ/dP)(x)/(1-e(x))$

Table 1: Summary of weight functions for different inferential targets

3.4 Conformalized counterfactual inference is exact for randomized trials

For randomized trials with perfect compliance, the strong ignorability assumption is satisfied by randomization and the propensity score is known since it is designed by researchers. In completely randomized experiments, $e(\cdot)$ is a constant mapping, in which case weighting is not required for either $Y(1)$ nor $Y(0)$. For general stratified experiments such as blocking experiments, $e(x)$ could vary with some of the covariates (e.g. age, gender), and one would apply weighted conformal inference by using the weight functions from Table 1. Either way, weighted conformal inference achieves coverage in finite samples (Proposition 1) even if our conditional quantile estimates are completely off. Moreover, taking $Y(1)$ and the ATE-type coverage as an example, $\chi^r(Q_X \parallel P_X) \le \mathbb{E}[1/e(X)^r]$. By Proposition 1, our method has almost exact coverage when the calibration fold is large and $\mathbb{E}[1/e(X)^2] < \infty$. This overlap condition is weaker than the stricter condition $0 < a < e(X) < b < 1$ typically assumed in the literature [e.g. Chen et al., 2008, Hirshberg and Wager, 2017, Ma and Wang, 2019, Hong et al., 2020].

We close this section with a last important bibliographical comment. As we were putting the finishing touch on this paper, we became aware of the independent work by Kivaranovic et al. [2020] applying unweighted standard conformal inference to construct counterfactual intervals for completely randomized experiments. Clearly, our two papers have a similar aim. That said, and as mentioned in their Section 2.1, the approach in Kivaranovic et al. [2020] cannot handle stratified experiments even if the propensity score is known. Hence, the scopes of the two papers are very different.

3.5 Conformalized counterfactual inference is doubly robust

For observational studies or randomized trials with imperfect compliance, the propensity score is unknown and needs to be estimated. Let $\hat{e}(x)$ denote the estimate of $e(x)$. In this subsection, we will see that the coverage of weighted split-CQR is approximately guaranteed if either $\hat{e}(x) \approx e(x)$ or $\hat{q}_\beta(x) \approx q_\beta(x)$ with $\beta \in \{\alpha_{\mathrm{lo}}, \alpha_{\mathrm{hi}}\}$. Before stating a rigorous result of this fact, we first provide an intuitive justification. On the one hand, if $\hat{e}(x) \approx e(x)$, our method approximates the oracle version of weighted split-CQR with the true weights and the intervals should, therefore, approximately achieve the desired coverage even if $\hat{q}_\beta(x)$ drastically deviates from the true conditional quantiles. On the other hand, if $\hat{q}_\beta(x) \approx q_\beta(x)$, where $q_\beta(x)$ is the β-th quantile of $Y(1)$ (or $Y(0)$) given $X = x$, then

$$V_i \approx \max\{q_{\alpha_{\mathrm{lo}}}(X_i) - Y_i(1), Y_i(1) - q_{\alpha_{\mathrm{hi}}}(X_i)\}.$$

As a result,

$$\mathbb{P}(V_i \le 0 \mid X_i) \approx \mathbb{P}(Y_i(1) \in [q_{\alpha_{\mathrm{lo}}}(X_i), q_{\alpha_{\mathrm{hi}}}(X_i)] \mid X_i) = \alpha_{\mathrm{hi}} - \alpha_{\mathrm{lo}}.$$

If $\alpha_{\text{hi}} - \alpha_{\text{lo}} = 1 - \alpha$, then 0 is approximately the $(1-\alpha)$-th quantile of the V_i's. In Algorithm 1 we have that η is the $(1-\alpha)$-th quantile of the random distribution $\sum_{i \in \mathcal{I}_{\text{ca}}} \hat{p}_i \delta_{V_i} + \hat{p}_\infty \delta_\infty$. Denote by G the cumulative distribution function (cdf) of this random distribution. Then

$$G(0) \approx \mathbb{E}[G(0) \mid X_i] = \sum_{i \in \mathcal{I}_{\text{ca}}} \hat{p}_i \mathbb{P}(V_i \le 0 \mid X_i) \approx \sum_{i \in \mathcal{I}_{\text{ca}}} \hat{p}_i (1-\alpha) \approx 1 - \alpha.$$

This says that 0 is just about the $(1-\alpha)$-th quantile of G. This implies that $\eta \approx 0$ and thus $\hat{C}(x) \approx [q_{\alpha_{\text{lo}}}(x), q_{\alpha_{\text{hi}}}(x)]$. By definition, the coverage in this case is approximately $\alpha_{\text{hi}} - \alpha_{\text{lo}} = 1 - \alpha$.

The following theorem, whose proof is in Appendix B, formalizes the above heuristics.

Theorem 1. *Let $N = |\mathcal{Z}_{\text{tr}}|$ and $n = |\mathcal{Z}_{\text{ca}}|$. Further, let $\hat{q}_{\beta,N}(x) = \hat{q}_{\beta,N}(x; \mathcal{Z}_{\text{tr}})$ be an estimate of the β-th conditional quantile $q_\beta(x)$ of $Y(1)$ given $X = x$, $\hat{e}_N(x) = \hat{e}_N(x; \mathcal{Z}_{\text{tr}})$ be an estimate of $e(x)$, and $\hat{C}_{N,n}(x)$ be the resulting interval from Algorithm 1. Assume that $\mathbb{E}[1/\hat{e}_N(X) \mid \mathcal{Z}_{\text{tr}}] < \infty$ and $\mathbb{E}[1/e(X)] < \infty$. Assume that one of the following holds:*

***A**1* $\lim_{N \to \infty} \mathbb{E}\left|\frac{1}{\hat{e}_N(X)} - \frac{1}{e(X)}\right| = 0$;

***A**2* *(1)* $\alpha_{\text{hi}} - \alpha_{\text{lo}} = 1 - \alpha$,

(2) there exists $r, b_1, b_2 > 0$ such that $\mathbb{P}(Y(1) = y \mid X = x) \in [b_1, b_2]$ uniformly over all (x, y) with $y \in [q_{\alpha_{\text{lo}}}(x) - r, q_{\alpha_{\text{lo}}}(x) + r] \cup [q_{\alpha_{\text{hi}}}(x) - r, q_{\alpha_{\text{hi}}}(x) + r]$,

(3) there exists $\delta > 0$ such that

$$\limsup_{N \to \infty} \mathbb{E}\left[\frac{1}{\hat{e}_N(X)^{1+\delta}}\right] < \infty, \quad \lim_{N \to \infty} \mathbb{E}\left[\frac{H_N(X)}{\hat{e}_N(X)}\right] = \lim_{N \to \infty} \mathbb{E}\left[\frac{H_N(X)}{e(X)}\right] = 0,$$

where

$$H_N(x) = \max\{|\hat{q}_{\alpha_{\text{lo}},N}(x) - q_{\alpha_{\text{lo}}}(x)|, |\hat{q}_{\alpha_{\text{hi}},N}(x) - q_{\alpha_{\text{hi}}}(x)|\}.$$

Then

$$\lim_{N,n \to \infty} \mathbb{P}_{(X,Y(1)) \sim P_X \times P_{Y(1)|X}}(Y(1) \in \hat{C}_{N,n}(X)) \ge 1 - \alpha. \tag{3.5}$$

*Furthermore, if **A**2 holds, then for any $\epsilon > 0$,*

$$\lim_{N,n \to \infty} \mathbb{P}_{X \sim P_X}\Big(\mathbb{P}(Y(1) \in \hat{C}_{N,n}(X) \mid X) \le 1 - \alpha - \epsilon\Big) = 0. \tag{3.6}$$

Theorem 1 is a special case of Theorem 3 in Appendix A on the double robustness of general weighted split-CQR. With the latter, it is a simple exercise to extend Theorem 1 to other types of coverage and to $Y(0)$ by consulting Table 1.

Property (3.5) is analogous to the doubly robust point estimation of ATE [e.g. Robins et al., 1994, Kang and Schafer, 2007], which yields consistent estimators if either the propensity score or the conditional mean of potential outcomes are consistent. Nonetheless, we emphasize that our double robustness is not the same since consistency of point estimates and coverage of interval estimates are different concepts.[2]

[2] To the best of our knowledge, no analogue of the double robustness exists for CATE. Double robustness is usually formulated in terms of optimal risk rather than consistency [e.g. Kennedy, 2020].

Property (3.6) implies that weighted split-CQR has approximately guaranteed conditional coverage if the conditional quantiles are estimated accurately. This is of course sufficient but not necessary. In practice, we may work with less accurate estimates and shall nevertheless empirically demonstrate the robustness of weighted split-CQR in terms of conditional coverage.

3.6 Numerical experiments

In this subsection we demonstrate the performance of our methods via simulation studies. In particular, we consider a variant of the example in Wager and Athey [2018]:

- The covariate vector $X = (X_1, \ldots, X_d)^T$ is such that $X_j = \Phi(X_j')$, where Φ denotes the cdf of the standard normal distribution and $(X_1', \ldots, X_d')$ is an equicorrelated multivariate Gaussian vector with mean zero and $\mathrm{Var}(X_j') = 1$, $\mathrm{Cov}(X_j', X_{j'}') = \rho$ for $j \neq j'$. When $\rho = 0$, X is uniformly distributed on the unit cube. When $\rho > 0$, the variables are positively correlated.

- The baseline potential outcome is such that $Y(0) \equiv 0$. This simplifies the problem into a pure counterfactual inference problem.

- The potential outcome $Y(1)$ is generated as follows:

$$\mathbb{E}[Y(1) \mid X] = f(X_1)f(X_2), \quad f(x) = \frac{2}{(1 + \exp(-12(x - 0.5)))},$$

  which is the same as in Wager and Athey [2018], and

$$Y(1) = \mathbb{E}[Y(1) \mid X] + \sigma(X)\epsilon, \quad \epsilon \sim N(0, 1);$$

  the homoscedastic case $\sigma^2(x) \equiv \sigma^2$ is considered in Wager and Athey [2018];

- The propensity score $e(x)$ is set as in Wager and Athey [2018]:

$$e(x) = \frac{1}{4}\left(1 + \beta_{2,4}(x_1)\right),$$

  where $\beta_{2,4}$ is the cdf of the beta distribution with shape parameters $(2, 4)$. This ensures that $e(x) \in [0.25, 0.5]$, thereby providing sufficient overlap.

In our experiments, we will consider $8 = 2 \times 2 \times 2$ scenarios: low ($d = 10$) and high ($d = 100$) dimensions, uncorrelated ($\rho = 0$) and correlated ($\rho = 0.9$) covariates, and homoscedastic ($\sigma^2(x) \equiv 1$) and heteroscedastic ($\sigma^2(x) = -\log(1 - x_1)$) errors.

We present comparisons with three competing methods offering qualitatively different approaches to uncertainty quantification as well as well-written `R` packages:

- Causal Forest [Wager and Athey, 2018] uses the infinitesimal jackknife [Efron, 2014, Wager et al., 2014] to estimate the variance of CATE estimators. The authors established asymptotically valid coverage of CATE under regularity assumptions. This work does not discuss ITE. The method is implemented in the `grf` package [Tibshirani et al., 2019].

	CF	X-learner	BART	CQR
CATE	✓	✓	✓	✗
ITE	✗	✗	✓	✓

(a) Guarantees of coverage in theory

	CF	X-learner	BART	CQR
CATE	✗	✗	✗	✓
ITE	✗	✗	✗	✓

(b) Guarantees of coverage in the simulation study

Table 2: Summary of coverage guarantees in theory (left) and in our simulation study (right).

- X-learner [Künzel et al., 2019] uses the bootstrap to estimate the variance of CATE estimators. As with Causal Forest, the authors did not develop tools to cover ITE. The method is implemented in the `causalToolbox` package [Knzel et al., 2020].

- Bayesian Additive Regression Trees (BART) were initially developed as a flexible general-purpose Bayesian machine learning algorithm [Chipman et al., 2010]. They were later applied to causal inference [e.g. Hill, 2011, Green and Kern, 2012, Hahn et al., 2020] and found to outperform other methods both in terms of accuracy and coverage [Dorie, 2017, Dorie et al., 2019]. The method constructs Bayesian credible intervals to cover CATE. By replacing credible intervals with predictive intervals, the method can be adapted to cover ITE. We use the functions `calc_credible_intervals` and `calc_prediction_intervals` from the `bartMachine` package [Kapelner and Bleich, 2016] to compute both intervals.

For weighted split-CQR, we estimate the propensity score via the gradient boosting algorithm [Friedman, 2001] by using the `gbm` package [Greenwell et al., 2019]. We further estimate the conditional quantiles in three different ways: (1) via quantile random forest [Athey et al., 2019] by using the `grf` package [Tibshirani et al., 2019], (2) via quantile gradient boosting by using the `gbm` package [Greenwell et al., 2019], and (3) via the predictive posterior quantiles from BART by using the `bartMachine` package [Kapelner and Bleich, 2016]. For all conditional quantile estimators, we set $\alpha_{\mathrm{lo}} = \alpha/2, \alpha_{\mathrm{hi}} = 1 - \alpha/2$. Lastly, we use 75% data as the training fold, as suggested by Sesia and Candès [2020]. Our method is implemented in `R` `cfcausal` package, available at https://github.com/lihualei71/cfcausal. Code to replicate all the results from the paper is available at https://github.com/lihualei71/cfcausalPaper.

Each time, we generate 100 independent datasets with sample size $n = 1000$. In each run, we also generate $n_{\mathrm{test}} = 10000$ extra independent data points, and construct 95% confidence intervals for each of them via the aforementioned methods. We then estimate the empirical marginal coverage of CATE and ITE as $(1/n_{\mathrm{test}})\sum_{i=1}^{n_{\mathrm{test}}} I(\tau(X_i) \in \hat{C}_{\mathrm{ITE}}(X_i))$ and $(1/n_{\mathrm{test}})\sum_{i=1}^{n_{\mathrm{test}}} I(Y_i(1) - Y_i(0) \in \hat{C}_{\mathrm{ITE}}(X_i))$, respectively, where $\hat{C}_{\mathrm{ITE}}(X_i) = \hat{C}_1(X_i)$ in this case. Note that our metric is not asking for coverage at every point x. Insteads, it demands coverage in an average sense. Therefore, a reliable method should, at the very least, have coverage close to or above 0.95. To be sure, a method with invalid marginal coverage certainly cannot have valid conditional coverage. As summarized in Table 2a, note, and this is important, that Causal Forest and X-learner are only guaranteed to cover CATE, whereas weighted split-CQR is only guaranteed to cover ITE. Lastly, BART has guarantees to cover both CATE and ITE by employing two types of intervals.

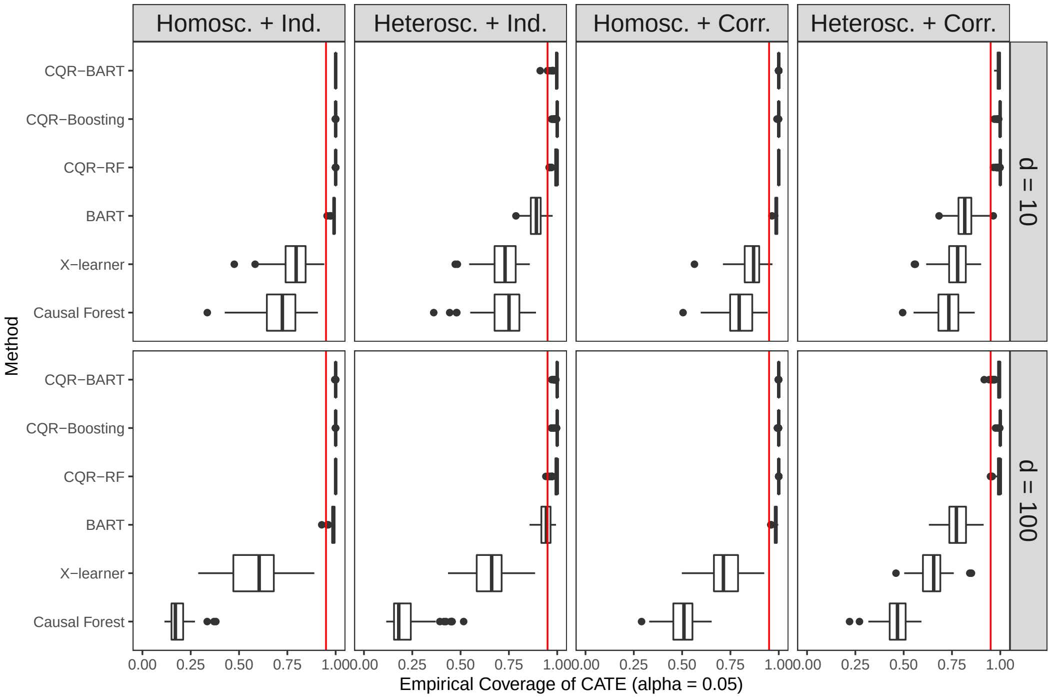

Figure 1: Empirical 95% coverage of CATE. Each panel corresponds to one of the eight scenarios. “Homosc.” and “Heterosc.” are short for “homoscedastic” and “heteroscedastic errors”; “Ind.” and “Corr.” are short for “independent” and “correlated covariates”.

Figure 1 presents CATE coverage results for this simple example; recall that the model smoothly depends only upon two variables out of ten or a hundred. Causal Forest and X-learner have poor coverage in all scenarios. Their performance degrades even further in the higher dimensional setting $d = 100$. Clearly, we must be far from the asymptotic setting considered in the literature. BART has better coverage than Causal Forest and X-learner. In the first three columns, we can see that the BART credible intervals cover CATE. In the last column where the covariates are correlated and the errors are heteroscedastic, BART has poor coverage, especially in high dimensions. Although our method is not guaranteed to cover CATE, we see that it achieves coverage in all scenarios, although it may be conservative. Intuitively, this happens because weighted split-CQR is designed to cover ITE and that in reasonable models, prediction intervals are wider and often include confidence intervals for the mean. In sum, CQR is the only method achieving valid coverage in all scenarios (contrast this with the theoretical predictions from Table 2a).

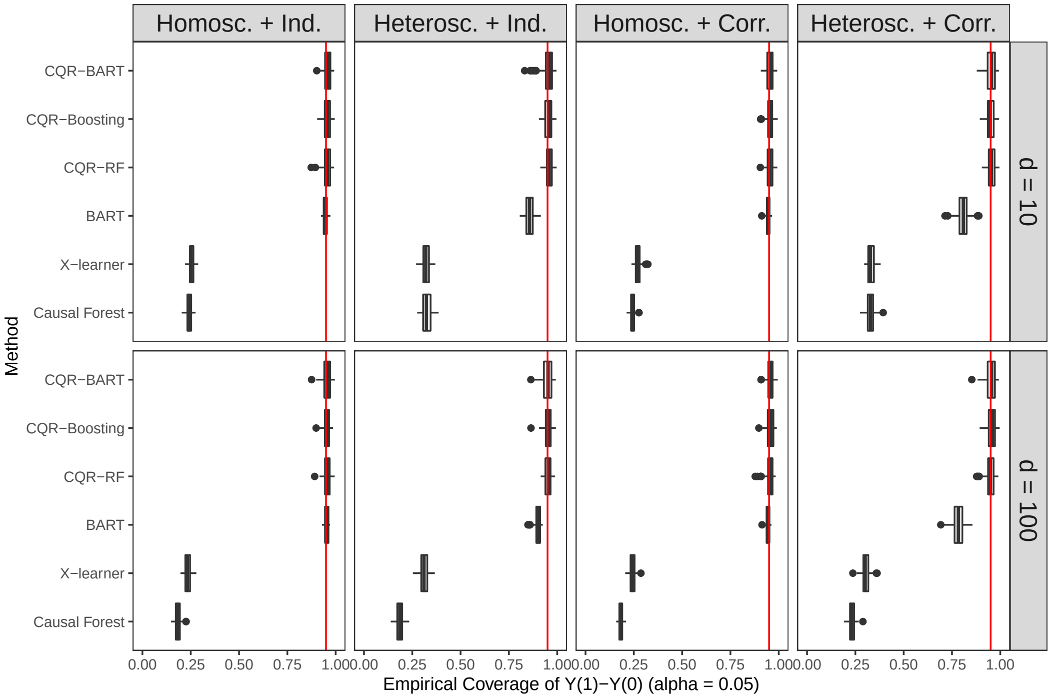

Figure 2: Empirical 95% coverage of ITE. Everything else is as in Figure 1.

Figure 2 presents ITE coverage, which is the real subject of this paper. Causal Forest and X-learner are not designed to cover ITE and it is therefore unsurprising that each method has low coverage. The predictive intervals of BART have perfect coverage with homoscedastic errors in both low and high dimensions. However, in heteroscedastic cases, BART has unsatisfactory coverage especially when the covariates are correlated. Finally, our method achieves almost exact coverage regardless of the learning procedures, regardless of whether the variables are correlated or not, the 'noise' is homoscedastic or not, and the dimension is low or high.

Next, we present interval lengths in Figure 3. Causal Forest and X-learner have short intervals and we have seen that this is because they are poorly calibrated. In homoscedastic settings where BART has valid coverage, BART also has the shortest intervals. This confirms the good performance of BART observed in the literature. Notably, weighted split-CQR, with BART as the learner of conditional quantiles, produces intervals that are almost as narrow as those produced with BART. As explained in Section 3.5, the correction $\eta \approx 0$ since BART fits the quantiles very well. As a result, $\hat{C}(x) \approx [\hat{q}_{\alpha_{\mathrm{lo}}}(x), \hat{q}_{\alpha_{\mathrm{hi}}}(x)]$. We would like to observe that a major source of power loss is the data splitting step since weighted split-CQR only uses 75% data to train BART. However, as shown in our companion paper [Lei and Candès, 2020], this can be mitigated via more sophisticated conformal inference methods such as cross-validation+ [Barber et al., 2019c]. In heteroscedastic settings, BART has shorter intervals because it has poor coverage as shown in Figure 2. We also see that weighted split-CQR has much larger variability in interval lengths when using BART as the learner. This is due to the fact that BART fails to estimate the conditional quantiles well and thus yields a noisy conformity correction η.

To evaluate the tightness of these intervals, we compute the average length of oracle intervals formed by the true 0.025-th and 0.975-th conditional quantiles. In this case, the errors are normally distributed and thus the expected length is $(2 \times 1.96)\mathbb{E}[\sigma(X)]$. This is equal to 3.92 in both cases because $\int_0^1 1dz = \int_0^1(-\log(1-z))dz = 1$. In all cases, we observe that the interval lengths of weighted split-CQR with gradient boosting and random forest are reasonably short. In homoscedastic cases, weighted split-CQR with BART almost achieves the oracle length, despite having an expected sample size for inferring $Y(1)$ equal to $n\mathbb{E}[e(X)] = (5/12)n \approx 417$.

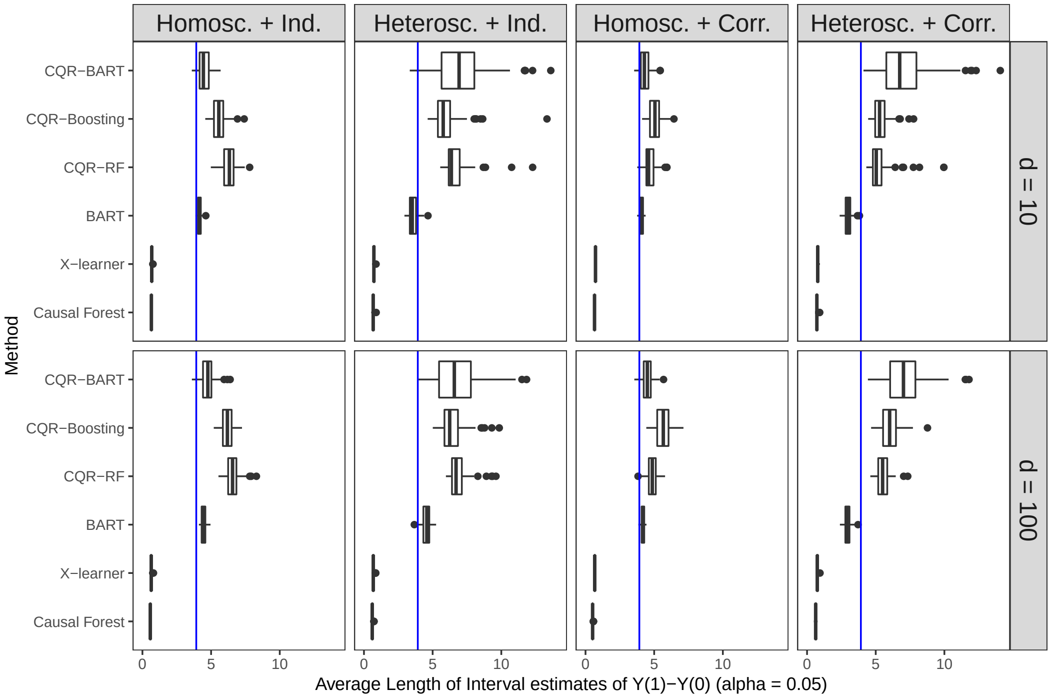

Figure 3: Lengths of interval estimates for ITE. The blue vertical line corresponds to the average length of oracle intervals. Everything else is as in Figure 1.

Finally, we turn to investigating the conditional coverage of all these methods. Recall that in the heteroscedastic setting, $\sigma^2(x) \to \infty$ as $x_1 \to 1$. Therefore, we expect the conditional coverage for instances with larger values of x_1 to be lower. Figure 4 displays the estimated conditional coverage of ITE as a function of the percentiles of $\sigma^2(x)$ when $d = 10$. Specifically, we stratify $\sigma^2(X_i)$'s on the 10000 testing points into 10 folds based on their $10\%, 20\%, \ldots, 90\%$ percentiles and estimate the coverage within each interval. It is clear that Causal Forest, X-learner and BART all have decreasing conditional coverage as $\sigma^2(x)$ increases. Although BART has much better marginal coverage than the other two methods, the poor conditional coverage near the right end point is worrisome. In contrast, weighted split-CQR with quantile random forest or quantile gradient boosting maintains conditional coverage. While the weighted split-CQR with BART does not perform as well as the other two variants, it improves upon BART implying that the calibration is also helpful in securing conditional coverage. In Appendix C, we show the same plots for $d = 100$ as well as plots of conditional coverage stratified by the CATE function $\tau(\cdot)$. They all exhibit similar patterns as shown in Figure 4.

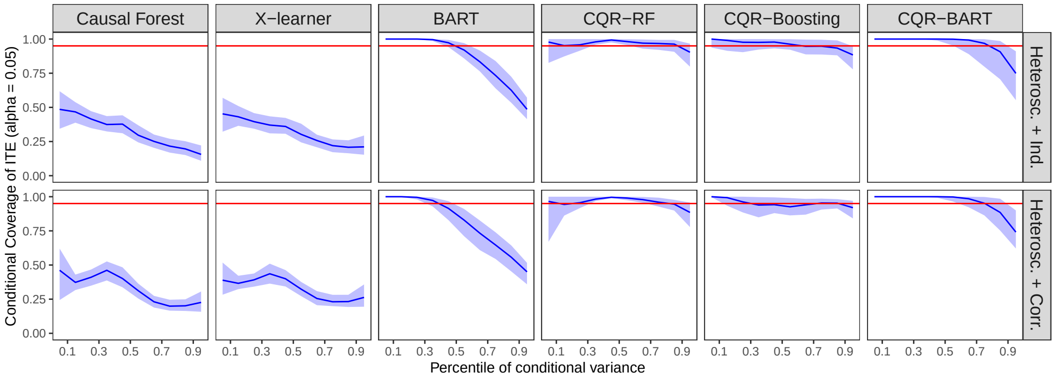

Figure 4: Estimated conditional coverage of ITE as a function of the conditional variance. Here, the dimension $d = 10$ and the coverage is set to 95%. The blue curves correspond to the median and the blue confidence bands are the 5% and 95% quantiles of these estimates across 100 replicates.

4 From Counterfactuals To Treatment Effects

We finally turn our attention to intervals for ITE for subjects not in the study, and for which both potential oucomes are missing.

4.1 A naive approach

Consider an arbitrary testing point x. Having constructed tools for counterfactual inference producing interval estimates of a given potential outcome, we can construct a pair of confidence intervals at level $1-\alpha/2$, namely, $[\hat{Y}^L(1;x), \hat{Y}^R(1;x)]$ for $Y(1)$ and $[\hat{Y}^L(0;x), \hat{Y}^R(0;x)]$ for $Y(0)$. By contrasting these two intervals, we can obtain an interval for ITE as follows:

$$\hat{C}_{\mathrm{ITE}}(x) = [\hat{Y}^L(1;x) - \hat{Y}^R(0;x), \hat{Y}^R(1;x) - \hat{Y}^L(0;x)].$$

If the counterfactual intervals have guaranteed coverage, then $\hat{C}_{\mathrm{ITE}}(x)$ also covers ITE at the level $1-\alpha$ in the sense of (2.5). This means that we can use counterfactual intervals produced by weighted split-CQR or by BART whenever they are suitable.

4.2 A nested approach

Despite its simplicity, the naive approach entirely decouples two potential outcomes and thus can be conservative. In this subsection, we propose a nested approach to mitigate this. The nested procedure starts by splitting the data into two folds. On the first fold we train $\hat{C}_1(x)$ and $\hat{C}_0(x)$ by applying counterfactual inference. On the second fold, for each unit i, we compute $\hat{C}_0(X_i)$ if $T_i = 1$ and compute $\hat{C}_1(X_i)$ if $T_i = 0$. This induces an interval $\hat{C}_{\mathrm{ITE}}(x; t, y^{\mathrm{obs}})$ for ITE defined as

$$\hat{C}_{\mathrm{ITE}}(x; t, y^{\mathrm{obs}}) = \begin{cases} y^{\mathrm{obs}} - \hat{C}_0(x), & t = 1, \\ \hat{C}_1(x) - y^{\mathrm{obs}}, & t = 0. \end{cases}$$

Unit	$Y_i(1)$	$Y_i(0)$	Y_i^{obs}
Treatment Group			
1	$Y_1(1)$	×	$Y_1(1)$
2	$Y_2(1)$	×	$Y_2(1)$
3	$Y_3(1)$	×	$Y_3(1)$
4	$Y_4(1)$	×	$Y_4(1)$
5	$Y_5(1)$	×	$Y_5(1)$
Control Group			
6	×	$Y_6(0)$	$Y_6(0)$
7	×	$Y_7(0)$	$Y_7(0)$
8	×	$Y_8(0)$	$Y_8(0)$
9	×	$Y_9(0)$	$Y_9(0)$
10	×	$Y_{10}(0)$	$Y_{10}(0)$

Unit	$\hat{Y}_i(1)$	$\hat{Y}_i(0)$	$\hat{C}_i$
Treatment Group			
11	$Y_{11}(1)$	$[\hat{Y}_{11}^L(0), \hat{Y}_{11}^R(0)]$	$[Y_{11}(1) - \hat{Y}_{11}^R(0), Y_{11}(1) - \hat{Y}_{11}^L(0)]$
12	$Y_{12}(1)$	$[\hat{Y}_{12}^L(0), \hat{Y}_{12}^R(0)]$	$[Y_{12}(1) - \hat{Y}_{12}^R(0), Y_{12}(1) - \hat{Y}_{12}^L(0)]$
13	$Y_{13}(1)$	$[\hat{Y}_{13}^L(0), \hat{Y}_{13}^R(0)]$	$[Y_{13}(1) - \hat{Y}_{13}^R(0), Y_{13}(1) - \hat{Y}_{13}^L(0)]$
14	$Y_{14}(1)$	$[\hat{Y}_{14}^L(0), \hat{Y}_{14}^R(0)]$	$[Y_{14}(1) - \hat{Y}_{14}^R(0), Y_{14}(1) - \hat{Y}_{14}^L(0)]$
15	$Y_{15}(1)$	$[\hat{Y}_{15}^L(0), \hat{Y}_{15}^R(0)]$	$[Y_{15}(1) - \hat{Y}_{15}^R(0), Y_{15}(1) - \hat{Y}_{15}^L(0)]$
Control Group			
16	$[\hat{Y}_{16}^L(1), \hat{Y}_{16}^R(1)]$	$Y_{16}(0)$	$[\hat{Y}_{16}^L(1) - Y_{16}(0), \hat{Y}_{16}^R(1) - Y_{16}(0)]$
17	$[\hat{Y}_{17}^L(1), \hat{Y}_{17}^R(1)]$	$Y_{17}(0)$	$[\hat{Y}_{17}^L(1) - Y_{17}(0), \hat{Y}_{17}^R(1) - Y_{17}(0)]$
18	$[\hat{Y}_{18}^L(1), \hat{Y}_{18}^R(1)]$	$Y_{18}(0)$	$[\hat{Y}_{18}^L(1) - Y_{18}(0), \hat{Y}_{18}^R(1) - Y_{18}(0)]$
19	$[\hat{Y}_{19}^L(1), \hat{Y}_{19}^R(1)]$	$Y_{19}(0)$	$[\hat{Y}_{19}^L(1) - Y_{19}(0), \hat{Y}_{19}^R(1) - Y_{19}(0)]$
20	$[\hat{Y}_{20}^L(1), \hat{Y}_{20}^R(1)]$	$Y_{20}(0)$	$[\hat{Y}_{20}^L(1) - Y_{20}(0), \hat{Y}_{20}^R(1) - Y_{20}(0)]$

Table 3: Sketch of the two folds in the nested approach: the first fold (left) is used to construct counterfactual intervals; (the second fold (right) is used to infer ITE.

Put $\hat{C}_i = \hat{C}_{\text{ITE}}(X_i; T_i, Y_i^{\text{obs}})$ as in Table 3, which illustrates the procedure. For any unit i in the second fold,

$$\mathbb{P}(Y_i(1) - Y_i(0) \in \hat{C}_i) = \mathbb{P}(T_i = 1)\mathbb{P}(Y_i(0) \in \hat{C}_0(X_i) \mid T_i = 1) + \mathbb{P}(T_i = 0)\mathbb{P}(Y_i(1) \in \hat{C}_1(X_i) \mid T_i = 0).$$

This gives that if

$$\mathbb{P}(Y_i(0) \in \hat{C}_0(X_i) \mid T_i = 1) \geq 1 - \alpha, \quad \mathbb{P}(Y_i(1) \in \hat{C}_1(X_i) \mid T_i = 0) \geq 1 - \alpha, \tag{4.1}$$

then

$$\mathbb{P}(Y_i(1) - Y_i(0) \in \hat{C}_i) \geq 1 - \alpha. \tag{4.2}$$

For randomized experiments with known propensity score $e(x)$, (4.1) is satisfied if we use $w_0(x) = e(x)/(1 - e(x))$ for $\hat{C}_0(x)$ and use $w_1(x) = (1 - e(x))/e(x)$ for $\hat{C}_1(x)$. Please note that we do not need to split α for $\hat{C}_0(x)$ and $\hat{C}_1(x)$. For observational studies, we can substitute $e(x)$ with $\hat{e}(x)$. By Theorem 1, (4.1) holds approximately if either the propensity scores or the conditional quantiles are estimated well.

The nested procedure creates an i.i.d. dataset $(X_i, \hat{C}_i)$, conditional on the first fold, such that $\hat{C}_i = \hat{C}_{\text{ITE}}(X_i; T_i, Y_i^{\text{obs}})$ covers the ITE with probability at least $1 - \alpha$. Therefore, the intervals $\hat{C}_i$ serve as "surrogate intervals" for ITE. If we can fit a model of $\hat{C}_i$ on X_i, denoted by $\tilde{C}_{\text{ITE}}(x)$, then the intervals can be generalized to subjects with both potential outcomes missing. If $\tilde{C}_{\text{ITE}}(X_i)$ does not shrink $\hat{C}_i$ drastically, it is likely that $\mathbb{P}(Y(1) - Y(0) \in \tilde{C}_{\text{ITE}}(X_i))$ shall be close to or above $1 - \alpha$.

It may be useful to think of the nested method as follows: instead of estimating the unobserved uncertainty, the nested method is fitting an observed "uncertainty measurement" $\hat{C}_i$. This is arguably simpler.

4.3 An inexact and an exact method under the nested framework

The function $\tilde{C}_{\text{ITE}}(x)$ can be obtained by training a model of the left- and right-end point of $\hat{C}_i$ on X_i separately using generic machine learning methods. Note that $\tilde{C}_{\text{ITE}}$ does not have the same theoretical guarantee as those offered by conformalized counterfactual inference since the fit may not be controlled.

For this reason, we refer to it as an "inexact method". Nonetheless, as will be shown later, the inexact method achieves, in our empirical examples, the target coverage with drastically shorter intervals than the naive approach.

In extremely sensitive settings where the validity of predictions is of serious concern, we may still need a method with a theoretical guarantee of coverage. Here, we propose a secondary conformal inference procedure on the induced dataset $(X_i, \hat{C}_i)$. Given a generic observation (X, T, Y^{obs}), let $C = \hat{C}_{\text{ITE}}(X, T, Y^{\text{obs}})$ denote the induced interval. The second procedure—the "exact method"—is based on the simple observation that if one can find an interval expansion function $\hat{\mathcal{C}}(\cdot)$ that maps a covariate value to an interval such that

$$\mathbb{P}(C \subset \hat{\mathcal{C}}_{\text{ITE}}(X)) \geq 1 - \gamma, \tag{4.3}$$

then by (4.2),

$$\mathbb{P}(Y(1) - Y(0) \notin \hat{\mathcal{C}}_{\text{ITE}}(X)) \leq \mathbb{P}(Y(1) - Y(0) \notin C) + \mathbb{P}(C \not\subset \hat{\mathcal{C}}_{\text{ITE}}(X)) \leq \alpha + \gamma.$$

```
Algorithm 2 (Unweighted) conformal inference for interval outcomes
Input: level $\gamma$, data $\mathcal{Z} = (X_i, C_i)_{i \in \mathcal{I}}$ where $C_i = [C_i^L, C_i^R]$, testing point $x$,
       functions $\hat{m}^L(x; \mathcal{D}), \hat{m}^R(x; \mathcal{D})$ to fit the conditional mean/median of $C^L, C^R$
Procedure:
  1: Randomly split $\mathcal{Z}$ into a training fold $\mathcal{Z}_{\text{tr}} \triangleq (X_i, C_i)_{i \in \mathcal{I}_{\text{tr}}}$ and a calibration fold $\mathcal{Z}_{\text{ca}} \triangleq (X_i, C_i)_{i \in \mathcal{I}_{\text{ca}}}$
  2: For each $i \in \mathcal{I}_{\text{ca}}$, compute non-conformity score $V_i = \max\{\hat{m}^L(X_i; \mathcal{Z}_{\text{tr}}) - C_i^L, C_i^R - \hat{m}^R(X_i; \mathcal{Z}_{\text{tr}})\}$
  3: Compute $\eta$ as the $(1-\gamma)(1 + 1/|\mathcal{Z}_{\text{ca}}|)$ quantile of the empirical distribution of $\{V_i : i \in \mathcal{I}_{\text{ca}}\}$
Output: $\hat{\mathcal{C}}(x) = [\hat{m}^L(x; \mathcal{Z}_{\text{tr}}) - \eta, \hat{m}^R(x; \mathcal{Z}_{\text{tr}}) + \eta]$
```

Denote by C^L and C^R the left- and the right-end point of C, respectively. To achieve (4.3), we need to find a lower confidence bound for C^L and an upper confidence bound for C^R. A naive approach is to apply standard (unweighted) one-sided conformal inference on C^L and C^R with level $\gamma/2$ separately. Such a crude Bonferroni correction may be conservative in practice. To overcome this, we propose a conformal inference procedure that jointly calibrates C^L and C^R. The coverage guarantee can be proved using the standard argument [e.g. Lei et al., 2018, Romano et al., 2019] and we include the proof in Appendix B for completeness.

Theorem 2. *Consider Algorithm 2 and assume* $(X_i, C_i) \overset{i.i.d.}{\sim} (X, C)$. *Then*

$$\mathbb{P}(C \subset \hat{\mathcal{C}}(X)) \geq 1 - \gamma.$$

With all this in place, both the inexact and the exact methods are stated in Algorithm 3.

4.4 Empirical performance

To evaluate the performance of our methods, we design numerical experiments on the data analyzed in the 2018 Atlantic Causal Inference Conference workshop on heterogeneous treatment effects [Car-

Algorithm 3 Nested approach for interval estimates of ITE

```
Input: level $\alpha$, level $\gamma$ (only for exact version), data $\mathcal{Z} = (X_i, Y_i, T_i)_{i=1}^n$, testing point $x$
Step I: data splitting
  Split the data into two folds $\mathcal{Z}_1$ and $\mathcal{Z}_2$
  Estimate propensity score $\hat{e}(x)$ on $\mathcal{Z}_1$
Step II: counterfactual inference on $\mathcal{Z}_2$
  for $i$ in $\mathcal{Z}_2$ with $T_i = 1$ do
      Compute $[\hat{Y}_i^L(0), \hat{Y}_i^R(0)]$ in Algorithm 1 on $\mathcal{Z}_1$ with level $\alpha$ and $w_0(x) = \hat{e}(x)/(1 - \hat{e}(x))$
      Compute $\hat{C}_i = [Y_i(1) - \hat{Y}_i^R(0), Y_i(1) - \hat{Y}_i^L(0)]$
  end for
  for $i$ in $\mathcal{Z}_2$ with $T_i = 0$ do
      Compute $[\hat{Y}_i^L(1), \hat{Y}_i^R(1)]$ in Algorithm 1 on $\mathcal{Z}_1$ with level $\alpha$ and $w_1(x) = (1 - \hat{e}(x))/\hat{e}(x)$
      Compute $\hat{C}_i = [\hat{Y}_i^L(1) - Y_i(0), \hat{Y}_i^R(1) - Y_i(0)]$
  end for
Step III: Interval of ITE on the testing point
  (Exact version) Apply Algorithm 2 on $(X_i, \hat{C}_i)_{i \in \mathcal{Z}_2}$ with level $\gamma$, yielding an interval $\hat{\mathcal{C}}_{\mathrm{ITE}}(x)$
  (Inexact version) Fit the conditional mean/median $\hat{m}^L(x)$ of $\hat{C}^L$ and $\hat{m}^R(x)$ of $\hat{C}^R$ separately and
  construct $\hat{\mathcal{C}}_{\mathrm{ITE}}(x) = [\hat{m}^L(x), \hat{m}^R(x)]$
Output: $\hat{\mathcal{C}}_{\mathbf{ITE}}(x)$
```

valho et al., 2019]. The workshop organizers generated a synthetic dataset based on the National Study of Learning Mindsets (NLSM) [Yeager et al., 2019], a large-scale randomized trial of a behavioral intervention, to emulate an observational study. For information on the dataset, please see Section 2 from Carvalho et al. [2019]. Due to privacy concerns, the workshop organizers only released limited information on the data generating process, as well as the simulated dataset, available at https://github.com/grf-labs/grf/tree/master/experiments/acic18. Although the focus of this workshop was on heterogeneous treatment effects, the organizers did not evaluate whether the submissions cover the ITE or CATE. In this subsection, we shall fill in this gap by comparing our method with Causal Forest, X-learner and BART, just as we did in Section 3.6.

Obviously, we must know the ground truth in order to evaluate coverage. Therefore, we generated synthetic datasets based on the available information from Carvalho et al. [2019]. First, we split data into two folds $\mathcal{Z}_1$ and $\mathcal{Z}_2$, with $|\mathcal{Z}_1| = 2079$ including 20% of the samples and $|\mathcal{Z}_2| = 8312$ including the remaining 80%. In our numerical experiments, we generate the covariate vector X by sampling from $\mathcal{Z}_2$ with replacement. To generate the potential outcomes, we apply random forest from `R` `randomForest` package on $\mathcal{Z}_1$ to fit $\mathbb{E}[Y(0)]$. Denote the output by $\hat{m}_0(x)$. Then we generate $\mathbb{E}[Y(1)]$ by adding the CATE function $\tau(x)$ (equation (1) of Carvalho et al. [2019]) to $\hat{m}_0(x)$. To account for heteroscedasticity, we apply quantile random forest from `R` `grf` package to fit the 25% and 75% conditional quantiles of $Y(0)$ and $Y(1)$ and compute the conditional interquartile ranges $\hat{r}_0(x)$ and $\hat{r}_1(x)$. Given a covariate vector X_i, we subsequently generate $Y_i(1)$ and $Y_i(0)$ as

$$Y_i(1) = \hat{m}_0(X_i) + \tau(X_i) + 0.5\hat{r}_1(X_i)\epsilon_{i1}, \quad Y_i(0) = \hat{m}_0(X_i) + 0.5\hat{r}_0(X_i)\epsilon_{i0}, \quad \epsilon_{i1}, \epsilon_{i0} \overset{i.i.d.}{\sim} N(0,1).$$

Finally, we generate propensity scores by applying random forest from `R` `randomForest` package on $\mathcal{Z}_1$

and truncating the estimated propensity score $\hat{e}(x)$ at 0.1 and 0.9 to guarantee overlap. For each X_i, we generate T_i as a Bernoulli random variable with parameter $\hat{e}(X_i)$.

For each run, we first generate 6000 quadruples $(X_i, T_i, Y_i(1), Y_i(0))$ sampled according to the above data generating process. We then randomly select $n = 1000$ samples as the training set to induce an observational study with observations $(X_i, T_i, Y_i^{\text{obs}})$. For the remaining 5000 testing samples, only the covariates X_i are accessible to the analyst; the triple $(T_i, Y_i(1), Y_i(0))$ is solely used to evaluate coverage. It goes without saying that each method will take the training set as the input and, for each testing sample, produce 95% intervals for ITE.

For weighted split-CQR, we apply the naive procedure as described in Section 4.1 and the nested procedure with both exact and inexact calibration as described in Algorithm 3. For all procedures, we apply BART to fit conditional quantiles, as in Section 3.6, and apply gradient boosting from `R` `gbm` package to fit propensity scores. For the exact nested method, we set $\alpha = \gamma = 0.025$. All methods are implemented in `R` `cfcausal` package, available at https://github.com/lihualei71/cfcausal.

Note that neither the naive nor the nested approaches are limited to CQR. Therefore, we also wrap both methods around BART as competitors, where only the inexact version is considered for the nested approach since BART cannot produce exact counterfactual intervals. Finally, since Causal Forest and X-learner cannot produce counterfactual intervals, except in the special case from Section 3.6 where $Y(0) \equiv 0$, we do not wrap the naive or the nested methods around them but directly report their confidence intervals as benchmarks instead. The R programs to replicate the results are available at https://github.com/lihualei71/cfcausalPaper.

Figure 5 presents the coverage and the average length of intervals estimated on the testing set by repeating the above procedure 100 times. As expected, the naive methods with both CQR and BART are conservative. The exact nested method with CQR is also conservative although the coverage is only guaranteed to be above 95% in the worst case. In contrast, the inexact nested methods with CQR or BART are less conservative. Notably, BART fails to achieve the desired coverage but CQR, with BART as the learner, calibrates it successfully. Also, we can see from the right panel that the average length of intervals of inexact-CQR is just slightly above that of inexact-BART while significantly lower than that of either the naive or the exact nested methods. Moreover, in accordance with the observations from Section 3.6, Causal Forest and X-learner have poor coverage and, thus, their (short) intervals are misleading. As in Section 3.6, we compute the average length of oracle intervals as $3.92\mathbb{E}[\sigma(X)]$ where $\sigma(x) = 0.5\sqrt{\hat{r}_1(x)^2 + \hat{r}_0(x)^2}$. Having said this, we are in a different situation here since the true conditional quantiles of ITE can never be identified without assumptions on the joint distribution of $Y(1)$ and $Y(0)$. Therefore, this oracle length cannot be achieved in general no matter how powerful the model fitting. Keeping this cautionary remark in mind, we nonetheless see that weighted split-CQR still produces reasonably short intervals.

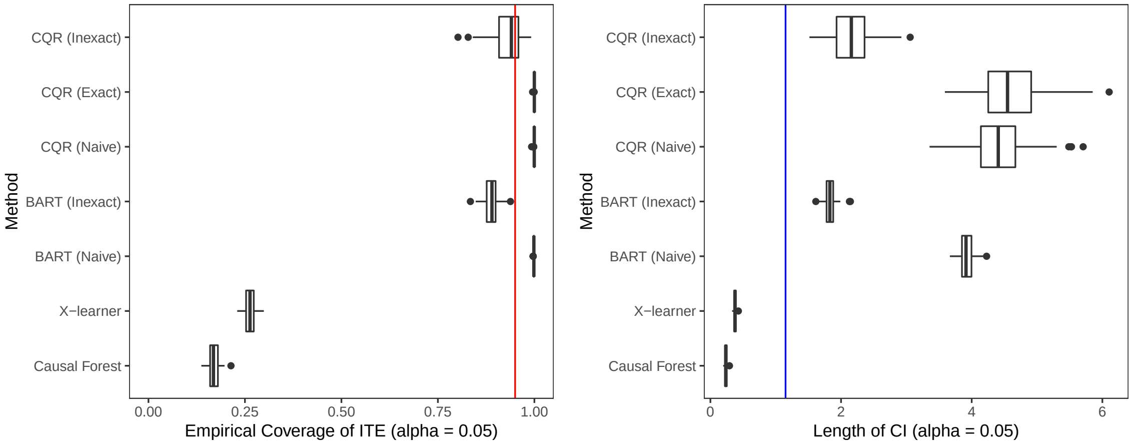

Figure 5: Coverage (left) and average length (right) of intervals for ITE on synthetic data generated from the NLSM data. The red vertical line corresponds to the target coverage 95% and the blue vertical line corresponds to the average length of oracle intervals.

Finally, as in Section 3.6, we investigate the conditional coverage as a function of the conditional variance $\sigma^2(x)$ and the CATE $\tau(x)$, respectively. For better visualization, we exclude Causal Forest and X-learner since they have poor marginal coverage. We can see from Figure 6 that inexact-CQR has desirable and relatively even conditional coverage while inexact-BART performs worse.

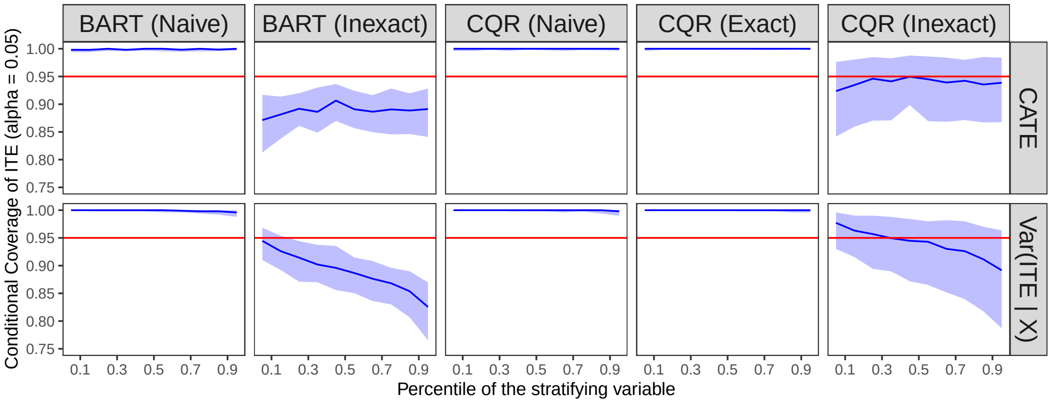

Figure 6: Estimated conditional coverage of ITE as a function of the conditional variance $\sigma^2(x)$ (upper row), and CATE $\tau(x)$ (lower row). The coverage is set to 95%. The blue curves correspond to the median and the blue confidence bands are the 5% and 95% quantiles of these estimates across 100 replicates.

4.5 Re-analysing NLSM data

In this subsection, we apply inexact-CQR with BART as the learner to re-analyze the NLSM data from Carvalho et al. [2019]. Since the ground truth is inaccessible, we only perform an exploratory analysis for the purpose of illustration.

To create informative testing points, we split the data into two folds $\mathcal{Z}_1$ and $\mathcal{Z}_2$. Then we apply

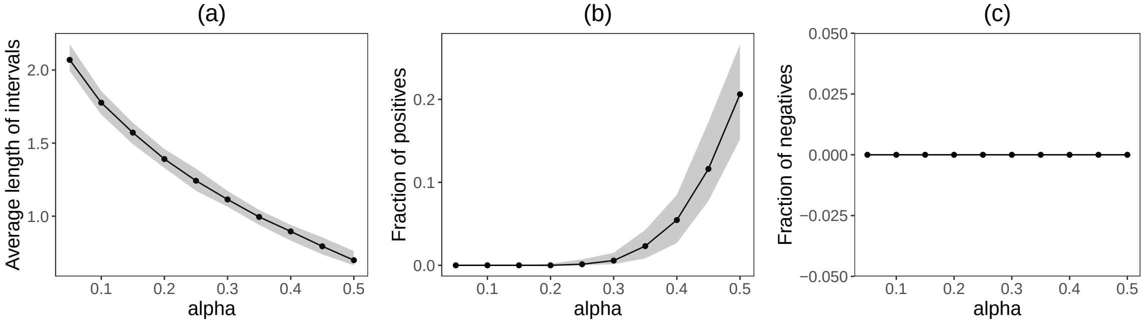

Figure 7: Re-analysis of NLSM data: (a) average length of intervals; (b) fraction of intervals with positive lower bounds; (c) fraction of intervals with negative upper bounds

weighted inexact-CQR on $\mathcal{Z}_1$ to produce intervals for ITE for each point in $\mathcal{Z}_2$. Similarly, we apply weighted inexact-CQR on $\mathcal{Z}_2$ to produce intervals of ITE for each point in $\mathcal{Z}_1$. To account for the variability from data splitting, we repeat the above procedure 100 times. Figure 7 (a) displays the average length of intervals as a function of level α with the upper and lower envelopes being respectively the 95% and 5% quantiles across 100 runs.

With these intervals, we can make inferential claims on ITE with confidence. For instance, we can decide to assign treatment to a patient if the lower confidence bound of her ITE interval is positive. Since these intervals are guaranteed to have desired coverage, it is likely that they produce fewer false positives on the average, although we leave a theoretical investigation of this intuition to future work. Figure 7 (b) and (c) show the fractions of intervals that only cover positive and negative values, respectively. We see some evidence of positive ITE when α is above 0.25 while no evidence of any negative ITE even when $\alpha = 0.5$.

5 From Potential Outcomes to Other Causal Frameworks

We proposed a method based on weighted conformal inference which produces interval estimates of counterfactuals and individual treatment effects under the potential outcome framework. For randomized experiments with perfect compliance, our method has guaranteed coverage in finite samples without any modeling assumptions on the data generating process. For randomized experiments with ignorable compliance or general observational studies under the strong ignorability assumption, our method is doubly robust in the sense that the coverage is asymptotically guaranteed if either the conditional quantiles of potential outcomes or the propensity scores are consistently estimated. In contrast, existing methods may suffer from a significant coverage deficit even in simple models. Furthermore, our framework naturally extends to other populations by modifying the weight function according to Table 1.

The key observation is the covariate shift together with the invariance of the conditional distribution: the observed distribution of $(X, Y(1))$ is $P_{X|T=1} \times P_{Y(1)|X}$ under ignorability while the target distribution is $P_X \times P_{Y(1)|X}$. Now the invariance of the conditional is also the enabling property in other frameworks of causal inference. As a consequence, our method can be naturally extended to those settings, and we close this paper by discussing two of them.

5.1 Causal diagram framework

Judea Pearl in his pioneering work [Pearl, 1995] introduced a general framework of causal inference based on graphical models. This framework does not rely on the notion of counterfactuals but instead defines causal effects through the `do` operator, which modifies the observed distribution by removing the causal paths that directly point to the intervention variable. We refer the readers to Pearl and Mackenzie [2018] for the philosophy and basics of causal diagrams and to Pearl et al. [2016] for the mathematical foundation of this framework.

We here only discuss the case where T is the intervention variable, Y is the outcome variable and X is a set of variables that satisfies the back-door criterion. Roughly speaking, this means that X includes all confounders and excludes all post-treatment variables. In this case, the foundational result in Pearl [1995] shows that

$$P_{(X,Y)|\mathrm{do}(T=t)} = P_X \times P_{Y|X,T=t}.$$

This is the target distribution to be inferred. In contrast, the observed distribution of (X, Y) given $T = t$ is

$$P_{(X,Y)|T=t} = P_{X|T=t} \times P_{Y|X,T=t}.$$

Clearly, this has exactly the same structure as in the potential outcome framework. As a consequence, weighted split-CQR can be applied without any modification to produce doubly robust interval estimates of Y under the `do` intervention.

5.2 Invariant prediction framework

Invariant prediction is another framework proposed by Peters et al. [2016]. It is particularly powerful when there are multiple data sources under different interventions, such as in gene knockout experiments. Consider an outcome variable Y, a set of interventions or covariates X and an environment variable E that indicates the source of data. Then one basic setting under this framework assumes that $Y \perp\!\!\!\perp E \mid X$ while X may depend on E with $X \mid E \sim P_X^E$. The goal is to predict the outcome under a new environment. For simplicity, we assume that a dataset $(X_{ij}, Y_{ij})_{i=1}^{n_j}$ is available for each environment $e_1, \ldots, e_J$ and a testing dataset $(X_{i0})_{i=1}^{n_0}$ for the target environment e_0. Due to the invariance assumption, on the j-th dataset, the observed distribution of (X, Y) is $P_X^{e_j} \times P_{Y|X}$ while the target distribution to be inferred on is $P_X^{e_0} \times P_{Y|X}$.

Again, this is reduced to a problem of constructing valid predictive intervals under covariate shifts. When $J = 1$, this has exactly the same structure as in the potential outcome framework and thus weighted split-CQR with weight function $dP_X^{e_0}(x)/dP_X^{e_1}(x)$ produces doubly robust intervals of Y under environment e_0. When $J > 1$, we can in principle apply the general weighted conformal inference techniques from Barber et al. [2019a]. However the weight function becomes much more complicated than that in Algorithm 1. An alternative approach is to create a weighted population from environments $e_1, \ldots, e_J$ with observed distribution

$$\left(\sum_{j=1}^{J} q_j P_X^{e_j}\right) \times P_{Y|X},$$

and apply Algorithm 1 on this pseudo dataset with weight $w(x) = 1/\sum_{j=1}^{J} q_j (dP_X^{e_j}/dP_X^{e_0})(x)$. The weights q_j can be chosen through certain balancing procedures that forces the covariate distribution to approximate $P_X^{e_0}$. We leave the formal development of this idea to future work.

Acknowledgment

E. C. was supported by Office of Naval Research grant N00014-20-12157. L. L. was supported by NSF OAC grant 1934578. These results were presented at Stanford on May 7, 2020. The authors are grateful to Stephen Bates, Jelena Bradic, Michael Celentano, Nikolaos Ignatiadis, Zhimei Ren and Chiara Sabatti for their constructive feedback.

References

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. *arXiv preprint arXiv:1812.09970*, 2018.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *arXiv preprint arXiv:1710.10251*, 2017.

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

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications.* Newnes, 2014.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019a.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*, 2019b.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, 2019c.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*, 2018.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019.

Niall Bolger, Katherine S Zee, Maya Rossignac-Milon, and Ran R Hassin. Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, 148(4):601, 2019.

Jennie E Brand and Yu Xie. Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, 75(2):273–302, 2010.

Richard Breen, Seongsoo Choi, and Anders Holm. Heterogeneous causal effects and sample selection bias. *Sociological Science*, 2:351–369, 2015.

Carlos Carvalho, Avi Feller, Jared Murray, Spencer Woody, and David Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge. *arXiv preprint arXiv:1907.07592*, 2019.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017.

Habiba Djebbari and Jeffrey Smith. Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145(1-2):64–80, 2008.

Vincent Dorie. aciccomp2016: Atlantic causal inference conference competition 2016 simulation. *R package version 0.1-0*, 2017.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

Jean-Pierre Florens, James J Heckman, Costas Meghir, and Edward Vytlacil. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206, 2008.

Margherita Fort. Unconditional and conditional quantile treatment effect: Identification strategies and interpretations. In *Topics in Theoretical and Applied Statistics*, pages 15–24. Springer, 2016.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. *gbm: Generalized Boosted Regression Models*, 2019. URL https://CRAN.R-project.org/package=gbm. R package version 2.1.5.

Justin Grimmer, Solomon Messing, and Sean J Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

Richard P Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

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.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

Han Hong, Michael P Leung, and Jessie Li. Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1):32–47, 2020.

Kosuke Imai and Aaron Strauss. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):1–19, 2011.

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

Joseph Kang and Joseph Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Adam Kapelner and Justin Bleich. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016.

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

Danijel Kivaranovic, Robin Ristl, Martin Posch, and Hannes Leeb. Conformal prediction intervals for the individual treatment effect. *arXiv preprint arXiv:2006.01474*, 2020.

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

Sren Knzel, Allen Tang, Ling Xie, Theo Saarinen, Peter Bickel, Bin Yu, and Jasjeet Sekhon. *causalToolbox: Toolbox for Causal Inference with emphasize on Heterogeneous Treatment Effect Estimator*, 2020. R package version 0.0.2.000.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Max GSell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Lihua Lei and Emmanuel Candès. Theory of weighted conformal inference. *Unpublished manuscript*, 2020.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, pages 1–10, 2019.

Stephen L Morgan. Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. *Sociology of education*, 74(4):341–374, 2001.

Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5:465–472, 1923/1990. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczyc, Tom X (1923): 1–51 (Annals of Agricultural Sciences).

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3538–3548, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. *arXiv preprint arXiv:2006.02544*, 2020.

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, 1983.

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

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Donald B Rubin. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292, 1990.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

Julie Tibshirani, Susan Athey, and Stefan Wager. *grf: Generalized Random Forests*, 2019. URL https://CRAN.R-project.org/package=grf. R package version 1.0.1.

Elizabeth Tipton. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266, 2013.

Elizabeth Tipton. How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, 39(6):478–501, 2014.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36:299–303, 1965.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490, 2012.

Vladimir Vovk. Transductive conformal predictors. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 348–360. Springer, 2013.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2): 9–28, 2015.

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

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590, 2009.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

Stephanie Winkelbeiner, Stefan Leucht, John M Kane, and Philipp Homan. Evaluation of differences in individual treatment response in schizophrenia spectrum disorders: a meta-analysis. *JAMA psychiatry*, 76(10):1063–1073, 2019.

Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

David S Yeager, Paul Hanselman, Gregory M Walton, Jared S Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, Barbara Schneider, Chris S Hulleman, and Cintia P Hinojosa. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369, 2019.

A Double Robustness of Weighted Split-CQR

This section establishes the double robustness of general weighted split-CQR.

Theorem 3. *Let $(X_i, Y_i) \overset{i.i.d.}{\sim} (X, Y) \sim P_X \times P_{Y|X}$ and Q_X be another distribution on the domain of X. Set $N = |\mathcal{Z}_{\mathrm{tr}}|$ and $n = |\mathcal{Z}_{\mathrm{ca}}|$. Further, let $\hat{q}_{\beta,N}(x) = \hat{q}_{\beta,N}(x; \mathcal{Z}_{\mathrm{tr}})$ be an estimate of the β-th conditional quantile $q_\beta(x)$ of Y given $X = x$, $\hat{w}_N(x) = \hat{w}_N(x; \mathcal{Z}_{\mathrm{tr}})$ be an estimate of $w(x) = (dQ_X/dP_X)(x)$, and $\hat{C}_{N,n}(x)$ be the resulting conformal interval from Algorithm 1. Assume that $\mathbb{E}[\hat{w}_N(X) \mid \mathcal{Z}_{\mathrm{tr}}] = 1$ and $\mathbb{E}[w(X)] = 1$, where $\mathbb{E}$ denotes expectation over $X \sim P_X$ and $\mathcal{Z}_{\mathrm{tr}}$. Assume that either B1 or B2 (or both) is satisfied:*

B1 $\lim_{N\to\infty} \mathbb{E}|\hat{w}_N(X) - w(X)| = 0$*;*

B2 *(1)* $\alpha_{\mathrm{hi}} - \alpha_{\mathrm{lo}} = 1 - \alpha$*,*

(2) there exists $r, b_1, b_2 > 0$ such that $\mathbb{P}(Y = y \mid X = x) \in [b_1, b_2]$ uniformly over all (x, y) with $y \in [q_{\alpha_{\mathrm{lo}}}(x) - r, q_{\alpha_{\mathrm{lo}}}(x) + r] \cup [q_{\alpha_{\mathrm{hi}}}(x) - r, q_{\alpha_{\mathrm{hi}}}(x) + r]$,

(3) there exists $\delta > 0$ such that $\limsup_{N\to\infty} \mathbb{E}\left[\hat{w}_N(X)^{1+\delta}\right] < \infty$,

(4) $\lim_{N\to\infty} \mathbb{E}[\hat{w}_N(X) H_N(X)] = \lim_{N\to\infty} \mathbb{E}[w(X) H_N(X)] = 0$, where

$$H_N(x) = \max\{|\hat{q}_{\alpha_{\mathrm{lo}},N}(x) - q_{\alpha_{\mathrm{lo}}}(x)|, |\hat{q}_{\alpha_{\mathrm{hi}},N}(x) - q_{\alpha_{\mathrm{hi}}}(x)|\}.$$

Then

$$\lim_{N,n\to\infty} \mathbb{P}_{(X,Y)\sim Q_X\times P_{Y|X}}(Y \in \hat{C}_{N,n}(X)) \geq 1-\alpha. \tag{A.1}$$

*Furthermore, under **B**2, for any $\epsilon > 0$,*

$$\lim_{N,n\to\infty} \mathbb{P}_{X\sim Q_X}\left(\mathbb{P}(Y \in \hat{C}_{N,n}(X) \mid X) \leq 1-\alpha-\epsilon\right) = 0. \tag{A.2}$$

A.1 Proof of Theorem 3 under assumption B1

Let Quantile$(\beta; F)$ denote the β-th quantile of a distribution function F, i.e.

$$\text{Quantile}(\beta; F) = \inf\{z : F(z) \geq \beta\} = \sup\{z : F(z) < \beta\}.$$

We start with two lemmas.

Lemma 1 (Equation (2) in Lemma 1 from Barber et al. [2019a])**.** *Let $v_1, \ldots, v_{n+1} \in \mathbb{R}$ and $(p_1, \ldots, p_{n+1}) \in \mathbb{R}$ be non-negative reals summing to 1. Then for any $\beta \in [0,1]$ and*

$$v_{n+1} \leq \text{Quantile}\left(\beta; \sum_{i=1}^{n+1} p_i\delta_{v_i}\right) \quad\Longleftrightarrow\quad v_{n+1} \leq \text{Quantile}\left(\beta; \sum_{i=1}^{n} p_i\delta_{v_i} + p_{n+1}\delta_\infty\right).$$

Lemma 2 (Equation (10) from Berrett et al. [2019])**.** *Let $d_{\text{TV}}(Q_{1X}, Q_{2X})$ denote the total-variation distance between Q_{1X} and Q_{2X}. Then*

$$d_{\text{TV}}(Q_{1X} \times P_{Y|X}, Q_{2X} \times P_{Y|X}) = d_{\text{TV}}(Q_{1X}, Q_{2X}).$$

Returning to the proof of Theorem 3, index the calibration fold by $\{1, \ldots, n\}$ and let $(X_{n+1}, Y_{n+1}) \sim Q_X \times P_{Y|X}$. Write Z_i for (X_i, Y_i) and V for $(V_1, \ldots, V_{n+1})$. For notational convenience, we suppress the subscripts N and n in $\hat{q}, \hat{w}, \hat{C}$ as well as in $\hat{p}_i$ and η. Next, for any permutation π on $\{1, \ldots, n+1\}$ and $v^* \in \mathbb{R}^{n+1}$, let $v^*_\pi = (v^*_{\pi(1)}, \ldots, v^*_{\pi(n+1)})$. Further, let $\ell(z)$ be the joint density of $\mathcal{Z} = (Z_1, \ldots, Z_{n+1})$ and $p(z)$ be the density of Z_1 (with respect to a dominating measure). Letting $\mathcal{E}(v)$ denote the unordered set of v, it is easy to see that

$$(V \mid \mathcal{E}(V) = \mathcal{E}(v^*), \mathcal{Z}_{\text{tr}}) \overset{d}{=} v^*_\Pi, \tag{A.3}$$

where Π is a random permutation with

$$\mathbb{P}\left(\Pi = \pi \mid \mathcal{Z}_{\text{tr}}\right) = \frac{p(z^*_\pi)}{\sum_\pi p(z^*_\pi)} = \frac{w(X_{\pi(n+1)})}{\sum_\pi w(X_{\pi(n+1)})} = \frac{w(X_{\pi(n+1)})}{n!\sum_{i=1}^{n+1} w(X_i)}.$$

As a result, for any $j \in \{1, 2, \ldots, n+1\}$,

$$\mathbb{P}\left(\Pi(n+1) = j \mid \mathcal{Z}_{\text{tr}}\right) = \frac{w(X_j)}{\sum_{i=1}^{n+1} w(X_i)} = p_j,$$

where p_{n+1} denotes p_∞ for notational convenience. This gives

$$(V_{n+1} \mid \mathcal{E}(V) = \mathcal{E}(v^*), \mathcal{Z}_{\mathrm{tr}}) \overset{d}{=} v^*_{\Pi(n+1)} \sim \sum_{i=1}^{n+1} p_i \delta_{v_i^*}. \tag{A.4}$$

Note that p_i involves the true likelihood ratio function $w(x)$ and thus is different from $\hat{p}_i$.

Let $\tilde{Q}_X$ be a measure with

$$d\tilde{Q}_X(x) = \hat{w}(x) dP_X(x).$$

Since $\mathbb{E}_{X \sim P_X}[\hat{w}(X)] = 1$, $\tilde{Q}_X$ is a probability measure. Consider now a new sample $(\tilde{X}_{n+1}, \tilde{Y}_{n+1}) \sim \tilde{Q}_X \times P_{Y|X}$. Let $\tilde{V}_{n+1}$ denote the non-conformity score of $(\tilde{X}_{n+1}, \tilde{Y}_{n+1})$ and set $\tilde{V} = (V_1, \ldots, V_n, \tilde{V}_{n+1})$. Using the same argument as for (A.4), we have

$$\left(\tilde{V}_{n+1} \mid \mathcal{E}(\tilde{V}) = \mathcal{E}(v^*), \mathcal{Z}_{\mathrm{tr}}\right) \sim \sum_{i=1}^{n+1} \hat{p}_i \delta_{v_i^*}. \tag{A.5}$$

As a consequence,

$$\begin{aligned}
\mathbb{P}\left(\tilde{Y}_{n+1} \in \hat{C}(\tilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}\right) &= \mathbb{P}\left(\tilde{V}_{n+1} \le \eta \mid \mathcal{Z}_{\mathrm{tr}}\right) \\
&= \mathbb{P}\left(\tilde{V}_{n+1} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \hat{p}_i \delta_{V_i^*} + \hat{p}_\infty \delta_\infty\right) \mid \mathcal{Z}_{\mathrm{tr}}\right) \\
&\overset{(i)}{=} \mathbb{P}\left(\tilde{V}_{n+1} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \hat{p}_i \delta_{V_i^*} + \hat{p}_\infty \delta_{\tilde{V}_{n+1}}\right) \mid \mathcal{Z}_{\mathrm{tr}}\right) \\
&= \mathbb{E}\mathbb{P}\left(\tilde{V}_{n+1} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \hat{p}_i \delta_{V_i^*} + \hat{p}_\infty \delta_{\tilde{V}_{n+1}}\right) \mid \mathcal{E}(\tilde{V}), \mathcal{Z}_{\mathrm{tr}}\right) \\
&\overset{(ii)}{\ge} 1-\alpha,
\end{aligned}$$

where (i) uses Lemma 1 and (ii) uses the definition of Quantile($\beta; F$). By Lemma 2,

$$d_{\mathrm{TV}}\left(Q_X \times P_{Y|X}, \tilde{Q}_X \times P_{Y|X}\right) = d_{\mathrm{TV}}\left(Q_X, \tilde{Q}_X\right),$$

and as a consequence,

$$\left|\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}, \mathcal{Z}_{\mathrm{ca}}\right) - \mathbb{P}\left(\tilde{Y}_{n+1} \in \hat{C}(\tilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}, \mathcal{Z}_{\mathrm{ca}}\right)\right| \le d_{\mathrm{TV}}(Q_X, \tilde{Q}_X),$$

which implies that

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}, \mathcal{Z}_{\mathrm{ca}}\right) \ge \mathbb{P}\left(\tilde{Y}_{n+1} \in \hat{C}(\tilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}, \mathcal{Z}_{\mathrm{ca}}\right) - d_{\mathrm{TV}}(Q_X, \tilde{Q}_X).$$

Taking expectation over $\mathcal{Z}_{\mathrm{ca}}$, we have

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}\right) \ge \mathbb{P}\left(\tilde{Y}_{n+1} \in \hat{C}(\tilde{X}_{n+1}) \mid \mathcal{Z}_{\mathrm{tr}}\right) - d_{\mathrm{TV}}(Q_X, \tilde{Q}_X) \ge 1 - \alpha - d_{\mathrm{TV}}(Q_X, \tilde{Q}_X).$$

Using the integral definition of total-variation distance,

$$d_{\mathrm{TV}}(Q_X, \tilde{Q}_X) = \frac{1}{2}\int |\hat{w}(x)dP_X(x) - dQ_X(x)| = \frac{1}{2}\int |\hat{w}(x)dP_X(x) - w(x)dP_X(x)| = \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)|.$$

Taking expectation over $\mathcal{Z}_{\mathrm{tr}}$, we have

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1})\right) \geq 1 - \alpha - \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)|.$$

The proof follows from assumption **B**1.

A.2 Proof of Theorem 3 under assumption B2

We start with the following Rosenthal-type inequality for sums of independent random variables with finite $(1+\delta)$-th moments.

Proposition 2 (Theorem 2 of von Bahr and Esseen [1965])**.** *Let Z_i be independent mean-zero random variables. Then for any $\delta \in [0, 1)$,*

$$\mathbb{E}\left|\sum_{i=1}^{n} Z_i\right|^{1+\delta} \leq 2\sum_{i=1}^{n}\mathbb{E}|Z_i|^{1+\delta}.$$

For notational convenience, we suppress the subcripts N and n in $\hat{q}, \hat{w}, \hat{C}$ as well as in $\hat{p}_i$ and η. We will prove the following result on η:

$$\lim_{N,n\to\infty}\mathbb{P}(\eta \geq -\epsilon) = 1, \quad \text{for any } \epsilon \in (0, r/2). \tag{A.6}$$

Let $(\tilde{X}, \tilde{Y})$ denote a generic random vector drawn from $Q_X \times P_{Y|X}$, which is independent of the data. If (A.6) holds, then for any fixed $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}(\tilde{Y} \in \hat{C}(\tilde{X}) \mid \tilde{X}) \\
&= \mathbb{P}\left(\min\{\hat{q}_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - \hat{q}_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq \eta \mid \tilde{X}\right) \\
&\geq \mathbb{P}\left(\min\{q_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - q_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq \eta - H(\tilde{X}) \mid \tilde{X}\right) \\
&\geq \mathbb{P}\left(\min\{q_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - q_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq -\epsilon - H(\tilde{X}) \mid \tilde{X}\right) - \mathbb{P}(\eta < -\epsilon \mid \tilde{X}) \\
&= \mathbb{P}\left(\min\{q_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - q_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq -\epsilon - H(\tilde{X}) \mid \tilde{X}\right) - \mathbb{P}(\eta < -\epsilon) \\
&\geq \mathbb{P}\left(\min\{q_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - q_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq -\epsilon - H(\tilde{X})I(H(\tilde{X}) \leq \epsilon) \mid \tilde{X}\right) - I(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta < -\epsilon) \\
&\overset{(i)}{\geq} \mathbb{P}\left(\min\{q_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - q_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq 0\right) - b_2(\epsilon + H(\tilde{X})I(H(\tilde{X}) \leq \epsilon)) - I(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta < -\epsilon) \\
&\geq \mathbb{P}\left(\min\{q_{\alpha_{\mathrm{lo}}}(\tilde{X}) - \tilde{Y}, \tilde{Y} - q_{\alpha_{\mathrm{hi}}}(\tilde{X})\} \leq 0\right) - b_2(\epsilon + H(\tilde{X})) - I(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta < -\epsilon) \\
&\overset{(ii)}{=} 1 - \alpha - b_2(\epsilon + H(\tilde{X})) - I(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta < -\epsilon);
\end{aligned}$$

above, (i) uses the condition that $2\epsilon < r$ and assumption **B**2 (2), and (ii) uses assumption **B**2 (2) that $\tilde{Y}$ has positive conditional densities at $q_{\alpha_{\mathrm{lo}}}(x)$ and $q_{\alpha_{\mathrm{hi}}}(x)$. Since $dQ_X(x) = w(x)dP_X(x)$ and $\mathbb{E}_{X\sim P_X}[w(X)] = 1$,

$$\mathbb{E}[H(\tilde{X})] = \mathbb{E}_{X\sim P_X}[w(X)H(X)].$$

By Markov's inequality, $\mathbb{P}(H(\tilde{X}) > \epsilon) \le \mathbb{E}_{X\sim P_X}[w(X)H(X)]/\epsilon$. By (A.6) and assumption **B**2 (4),

$$\lim_{N,n\to\infty} \mathbb{P}_{(X,Y)\sim Q_X\times P_{Y|X}}(Y \in \hat{C}(X)) = \lim_{N,n\to\infty} \mathbb{E}[\mathbb{P}(\tilde{Y} \in \hat{C}(X) \mid \tilde{X})] \ge 1 - \alpha - b_2\epsilon.$$

Since this holds for arbitrary $\epsilon > 0$, this implies (3.5). Similarly, if (A.6) holds, for sufficiently large N and n, $\mathbb{P}(\eta < -\epsilon) \le b_2\epsilon$ and thus

$$\mathbb{P}\left(\mathbb{P}(\tilde{Y} \in \hat{C}(\tilde{X}) \mid \tilde{X}) \le 1 - \alpha - 3b_2\epsilon\right) \le 2\mathbb{P}(H(\tilde{X}) \ge \epsilon).$$

By Markov's inequality,

$$\mathbb{P}(H(\tilde{X}) \ge \epsilon) \le \frac{\mathbb{E}[H(\tilde{X})]}{\epsilon} = \frac{\mathbb{E}_{X\sim P_X}[w(X)H(X)]}{\epsilon}.$$

By assumption **B**2 (4),

$$\lim_{N,n\to\infty} \mathbb{P}\left(\mathbb{P}(\tilde{Y} \in \hat{C}(\tilde{X}) \mid \tilde{X}) \le 1 - \alpha - 3b_2\epsilon\right) = 0.$$

Since $(\tilde{X}, \tilde{Y}) \sim Q_X \times P_{Y|X}$, (3.6) is proved by replacing $3b_2\epsilon$ with ϵ.

Now we prove (A.6). Let G denote the cdf of the random distribution $\sum_{i=1}^n \hat{p}_i\delta_{V_i} + \hat{p}_\infty\delta_\infty$. Again, G implicitly depends on N and n. Then (A.6) is equivalent to

$$\lim_{N,n\to\infty} \mathbb{P}(G(-\epsilon) < 1 - \alpha) = 1, \quad \text{for any } \epsilon > 0. \tag{A.7}$$

Note that $G(-\epsilon)$ is a weighted sum of indicator functions:

$$G(-\epsilon) = \sum_{i=1}^n \hat{p}_i I(V_i \le -\epsilon).$$

Let $G^*(-\epsilon)$ denote the expectation of $G(-\epsilon)$ conditional on $\mathcal{D} = \{\mathcal{Z}_{\mathrm{tr}}, (X_i)_{i=1}^n\}$, namely,

$$G^*(-\epsilon) = \mathbb{E}[G(-\epsilon) \mid \mathcal{D}] = \sum_{i=1}^n \hat{p}_i \mathbb{P}(V_i \le -\epsilon \mid \mathcal{D}).$$

Then

$$G(-\epsilon) - G^*(-\epsilon) = \sum_{i=1}^n \hat{p}_i \left(I(V_i \le -\epsilon) - \mathbb{P}(V_i \le -\epsilon \mid \mathcal{D})\right).$$

Conditional on $\mathcal{D}$, $G(-\epsilon) - G^*(-\epsilon)$ is σ^2 sub-Gaussian with

$$\sigma^2 = \sum_{i=1}^n \hat{p}_i^2 \le \max_i \hat{p}_i \left(\sum_{i=1}^n \hat{p}_i\right) \le \max_i \hat{p}_i.$$

For any $t > 0$,

$$\mathbb{P}\left(G(-\epsilon) - G^*(-\epsilon) \geq t \mid \mathcal{D}\right) \leq 2\exp\left(-\frac{t^2}{2\max_i \hat{p}_i}\right)$$
$$\Longrightarrow \quad \mathbb{P}\left(G(-\epsilon) - G^*(-\epsilon) \geq t\right) \leq 2\mathbb{E}\exp\left(-\frac{t^2}{2\max_i \hat{p}_i}\right). \quad \text{(A.8)}$$

By definition of $H(x)$,

$$V_i \geq \max\{q_{\alpha_{\mathrm{lo}}}(X_i) - Y_i, Y_i - q_{\alpha_{\mathrm{hi}}}(X_i)\} - H(X_i) \triangleq V_i^* - H(X_i).$$

By assumption **B**2 (1) and (2), $\mathbb{P}(V_i^* \leq 0 \mid \mathcal{D}) = \alpha_{\mathrm{hi}} - \alpha_{\mathrm{lo}} = 1 - \alpha$. Conditioning on $\mathcal{D}$, $H(X_i)$ is deterministic. As a result, when $\epsilon < 2r$,

$$\begin{aligned}
G^*(-\epsilon) &\leq \sum_{i=1}^n \hat{p}_i \left(I\left(H(X_i) \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(V_i^* \leq -\frac{\epsilon}{2} \mid \mathcal{D}\right)\right) \\
&\leq \sum_{i=1}^n \hat{p}_i \left(I\left(H(X_i) \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(V_i^* \leq -\frac{\epsilon}{2} \mid \mathcal{D}\right)\right) \\
&\leq \sum_{i=1}^n \hat{p}_i \left(I\left(H(X_i) \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(V_i^* \leq 0 \mid \mathcal{D}\right) - \frac{\epsilon b_1}{2}\right) \\
&\leq \left(1 - \alpha - \frac{\epsilon b_1}{2}\right) + \sum_{i=1}^n \hat{p}_i I\left(H(X_i) \geq \frac{\epsilon}{2}\right) \\
&\leq \left(1 - \alpha - \frac{\epsilon b_1}{2}\right) + \frac{2}{\epsilon}\sum_{i=1}^n \hat{p}_i H(X_i). \qquad \text{(A.9)}
\end{aligned}$$

Combining (A.8) and (A.9) with $t = \epsilon b_1/4$ gives

$$\begin{aligned}
\mathbb{P}(G(-\epsilon) \geq &1 - \alpha) \\
&\leq \mathbb{P}\left(G(-\epsilon) - G^*(-\epsilon) \geq \frac{\epsilon b_1}{4}\right) + \mathbb{P}\left(\sum_{i=1}^n \hat{p}_i H(X_i) \geq \frac{\epsilon^2 b_1}{8}\right) \\
&\leq 2\mathbb{E}\exp\left(-\frac{\epsilon^2 b_1^2}{32\max_i \hat{p}_i}\right) + \mathbb{P}\left(\sum_{i=1}^n \hat{p}_i H(X_i) \geq \frac{\epsilon^2 b_1}{8}\right) \\
&\leq 2\exp\left(-\frac{\epsilon^2 b_1^2}{32}\log n\right) + 2\mathbb{P}\left(\max_i \hat{p}_i \geq \frac{1}{\log n}\right) + \mathbb{P}\left(\sum_{i=1}^n \hat{p}_i H(X_i) \geq \frac{\epsilon^2 b_1}{8}\right). \qquad \text{(A.10)}
\end{aligned}$$

Recall that

$$\hat{p}_i \leq \frac{\hat{w}(X_i)}{\sum_{i=1}^n \hat{w}(X_i)}$$

and, therefore,

$$\mathbb{P}\left(\max_i \hat{p}_i \geq \frac{1}{\log n}\right) \leq \mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i) \leq \frac{n}{2}\right) + \mathbb{P}\left(\max_i \hat{w}(X_i) \geq \frac{n}{2\log n}\right). \qquad \text{(A.11)}$$

Assume without loss of generality that $0 < \delta < 1$. Since $\mathbb{E}[\hat{w}(X_i) \mid \mathcal{Z}_{\mathrm{tr}}] = 1$, Markov's inequality and Proposition 2 give

$$\mathbb{P}\left(\sum_{i=1}^{n} \hat{w}(X_i) \le \frac{n}{2}\right) \le \mathbb{P}\left(\left|\sum_{i=1}^{n}(\hat{w}(X_i) - 1)\right| \ge \frac{n}{2}\right) \le \frac{2^{1+\delta}}{n^{1+\delta}}\mathbb{E}\left|\sum_{i=1}^{n}(\hat{w}(X_i) - 1)\right|^{1+\delta}$$
$$\le \frac{2^{2+\delta}}{n^{1+\delta}}\sum_{i=1}^{n}\mathbb{E}|\hat{w}(X_i) - 1|^{1+\delta} = \frac{2^{2+\delta}\mathbb{E}|\hat{w}(X) - 1|^{1+\delta}}{n^{\delta}} \le \frac{2^{2+2\delta}\left(\mathbb{E}[\hat{w}(X)^{1+\delta}] + 1\right)}{n^{\delta}}. \quad \text{(A.12)}$$

Using Markov's inequality once more gives

$$\mathbb{P}\left(\max_i \hat{w}(X_i) \ge \frac{n}{2\log n}\right) \le \frac{2^{1+\delta}(\log n)^{1+\delta}}{n^{1+\delta}}\mathbb{E}\left[\left(\max_i \hat{w}(X_i)\right)^{1+\delta}\right]$$
$$\le \frac{2^{1+\delta}(\log n)^{1+\delta}}{n^{1+\delta}}\left(\sum_{i=1}^{n}\mathbb{E}\left[\hat{w}(X_i)^{1+\delta}\right]\right) = \frac{2^{1+\delta}(\log n)^{1+\delta}}{n^{\delta}}\mathbb{E}\left[\hat{w}(X)^{1+\delta}\right]. \quad \text{(A.13)}$$

Together, (A.11), (A.13) and assumption **B**2 (3) give

$$\lim_{N,n\to\infty}\mathbb{P}\left(\max_i \hat{p}_i \ge \frac{1}{\log n}\right) = 0. \quad \text{(A.14)}$$

Similar to (A.11),

$$\mathbb{P}\left(\sum_{i=1}^{n}\hat{p}_i H(X_i) \ge \frac{\epsilon^2 b_1}{8}\right) \le \mathbb{P}\left(\sum_{i=1}^{n}\hat{w}(X_i) \le \frac{n}{2}\right) + \mathbb{P}\left(\sum_{i=1}^{n}\hat{w}(X_i)H(X_i) \ge \frac{\epsilon^2 b_1 n}{16}\right),$$

and by Markov's inequality,

$$\mathbb{P}\left(\sum_{i=1}^{n}\hat{w}(X_i)H(X_i) \ge \frac{\epsilon^2 b_1 n}{16}\right) \le \frac{16}{n\epsilon^2 b_1}\sum_{i=1}^{n}\mathbb{E}[\hat{w}(X_i)H(X_i)] = \frac{16}{\epsilon^2 b_1}\mathbb{E}[\hat{w}(X)H(X)].$$

By (A.12) and assumption **B**2 (4),

$$\lim_{N,n\to\infty}\mathbb{P}\left(\sum_{i=1}^{n}\hat{p}_i H(X_i) \ge \frac{\epsilon^2 b_1}{8}\right) = 0. \quad \text{(A.15)}$$

(A.10), (A.14) and (A.15) complete the proof of (A.7). In turn, this establishes (A.6).

B Proofs of Other Results

B.1 Proof of Theorem 1

Since $\mathbb{E}[1/\hat{e}_N(X) \mid \mathcal{Z}_{\mathrm{tr}}] < \infty$ and $\mathbb{E}[1/e(X)] < \infty$, we can here set

$$\hat{w}_N(x) = \frac{1/\hat{e}_N(x)}{\mathbb{E}[1/\hat{e}_N(X) \mid \mathcal{Z}_{\mathrm{tr}}]},\ w(x) = \frac{1/e(x)}{\mathbb{E}[1/e(X)]} \quad \Longrightarrow \quad \mathbb{E}[w_N(X) \mid \mathcal{Z}_{\mathrm{tr}}] = 1 = \mathbb{E}[w(X)].$$

Thus, assumption **B**1 reduces to

$$\lim_{N\to\infty}\mathbb{E}\left|\frac{1/\hat{e}_N(X)}{\mathbb{E}[1/\hat{e}_N(X)\mid\mathcal{Z}_{\mathrm{tr}}]}-\frac{1/e(X)}{\mathbb{E}[1/e(X)]}\right|=0,$$

and assumptions **B**2 (3)-(4) reduce to

$$\limsup_{N\to\infty}\frac{\mathbb{E}[1/\hat{e}(X)^{1+\delta}]}{\mathbb{E}[1/\hat{e}(X)]^{1+\delta}}<\infty,\quad \lim_{N\to\infty}\frac{\mathbb{E}[H_N(X)/\hat{e}_N(X)]}{\mathbb{E}[1/\hat{e}_N(X)]}=\lim_{N\to\infty}\frac{\mathbb{E}[H_N(X)/e(X)]}{\mathbb{E}[1/e(X)]}=0.$$

Clearly, **A**2 implies **B**2 since $e(x),\hat{e}_N(x)\in[0,1]$. Now we prove that **A**1 implies **B**1. In fact,

$$\begin{aligned}
&\lim_{N\to\infty}\mathbb{E}\left|\frac{1/\hat{e}_N(X)}{\mathbb{E}[1/\hat{e}_N(X)\mid\mathcal{Z}_{\mathrm{tr}}]}-\frac{1/e(X)}{\mathbb{E}[1/e(X)]}\right|\\
\le&\limsup_{N\to\infty}\frac{1}{\mathbb{E}[1/\hat{e}_N(X)\mid\mathcal{Z}_{\mathrm{tr}}]}\mathbb{E}\left|\frac{1}{\hat{e}_N(X)}-\frac{1}{e(X)}\right|+\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{e(X)}\right]\mathbb{E}\left|\frac{1}{\mathbb{E}[1/\hat{e}_N(X)\mid\mathcal{Z}_{\mathrm{tr}}]}-\frac{1}{\mathbb{E}[1/e(X)]}\right|\\
\overset{(i)}{\le}&\limsup_{N\to\infty}\mathbb{E}\left|\frac{1}{\hat{e}_N(X)}-\frac{1}{e(X)}\right|+\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{e(X)}\right]\mathbb{E}\left|\mathbb{E}\left[\frac{1}{\hat{e}_N(X)}\mid\mathcal{Z}_{\mathrm{tr}}\right]-\mathbb{E}\left[\frac{1}{e(X)}\right]\right|\\
\le&\limsup_{N\to\infty}\mathbb{E}\left|\frac{1}{\hat{e}_N(X)}-\frac{1}{e(X)}\right|+\limsup_{N\to\infty}\mathbb{E}\left[\frac{1}{e(X)}\right]\mathbb{E}\left(\mathbb{E}\left[\left|\frac{1}{\hat{e}_N(X)}-\frac{1}{e(X)}\right|\mid\mathcal{Z}_{\mathrm{tr}}\right]\right)\\
\overset{(ii)}{=}&\left(1+\mathbb{E}\left[\frac{1}{e(X)}\right]\right)\limsup_{N\to\infty}\mathbb{E}\left|\frac{1}{\hat{e}_N(X)}-\frac{1}{e(X)}\right|;
\end{aligned}$$

(i) above uses the fact that $e(x),\hat{e}_N(x)\in[0,1]$ and (ii) uses assumption **A**1 and the condition that $\mathbb{E}[1/e(X)]<\infty$.

B.2 Proof of Theorem 2

Let $n=|\mathcal{Z}_{\mathrm{ca}}|$ and (X_{n+1},C_{n+1}) be an independent copy of (X,C). Further let

$$V_{n+1}=\max\{\hat{m}^L(X_{n+1};\mathcal{Z}_{\mathrm{tr}})-C_{n+1}^L,C_{n+1}^R-\hat{m}^R(X_{n+1};\mathcal{Z}_{\mathrm{tr}})\}.$$

Conditional on $\mathcal{Z}_{\mathrm{tr}}$, $V_1,\ldots,V_n,V_{n+1}$ are exchangeable. Then

$$\mathbb{P}\left(V_{n+1}\le\mathrm{Quantile}\left(1-\alpha;\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{V_i}\right)\right)\ge1-\alpha.$$

By definition,

$$\mathrm{Quantile}\left(1-\alpha;\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{V_i}\right)=\mathrm{Quantile}\left((1-\alpha)\frac{n+1}{n};\frac{1}{n}\sum_{i=1}^{n}\delta_{V_i}\right)=\eta.$$

As a consequence,

$$\mathbb{P}\left(C_{n+1}\in\hat{\mathcal{C}}(X_{n+1})\right)=\mathbb{P}\left(V_{n+1}\le\eta\right)\ge1-\alpha.$$

C Additional Experimental Results

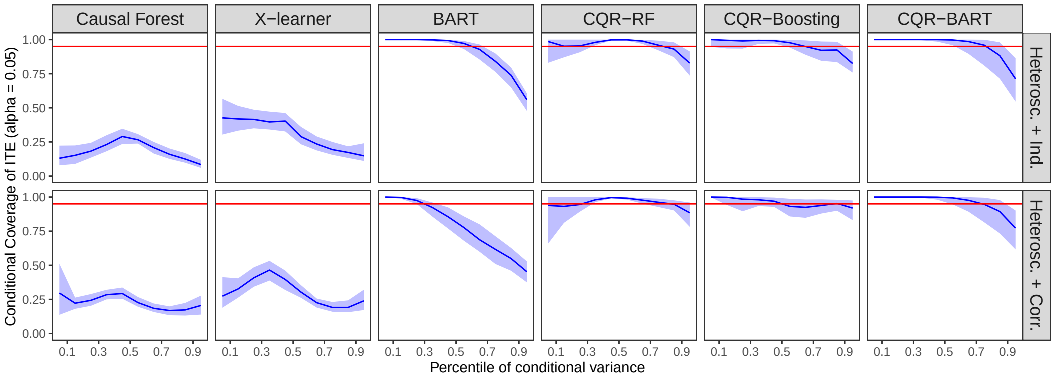

Figure 8: Estimated conditional coverage of ITE as a function of the conditional variance $\sigma^2(x)$ for the heteroscedastic cases from Section 3.6. Here, $d = 100$ and $\alpha = 0.05$. The blue curves correspond to the median and the boundaries of the blue confidence bands correspond to the 95% and 5% quantiles of these estimates across 100 replicates.

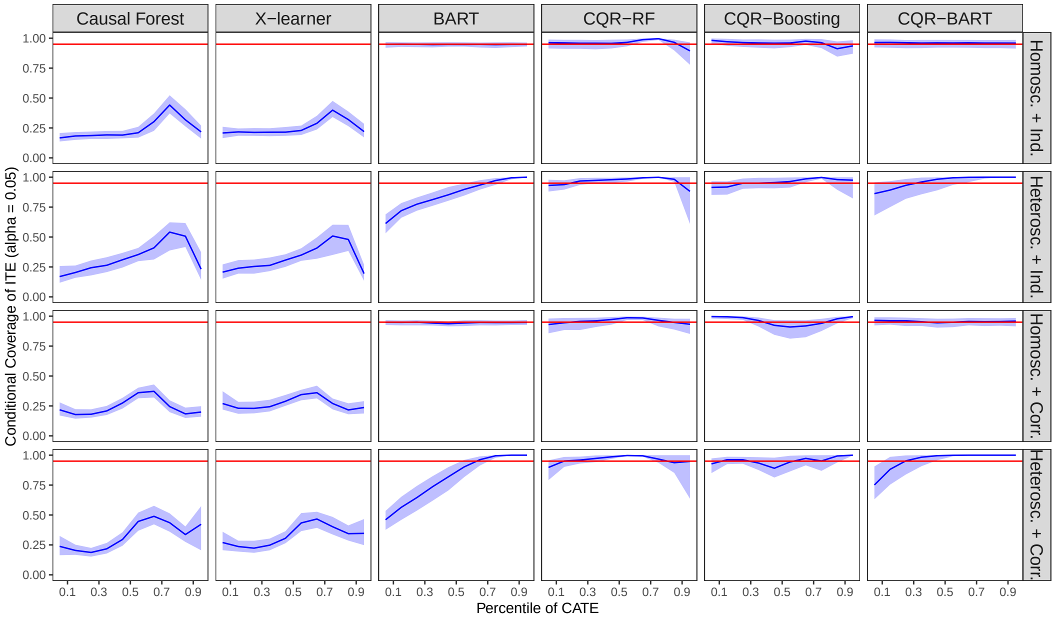

Figure 9: Estimated conditional coverage of ITE as a function of the CATE $\tau(x)$ for all scenarios with $d = 10$ from Section 3.6. Everything else is as in Figure 8.

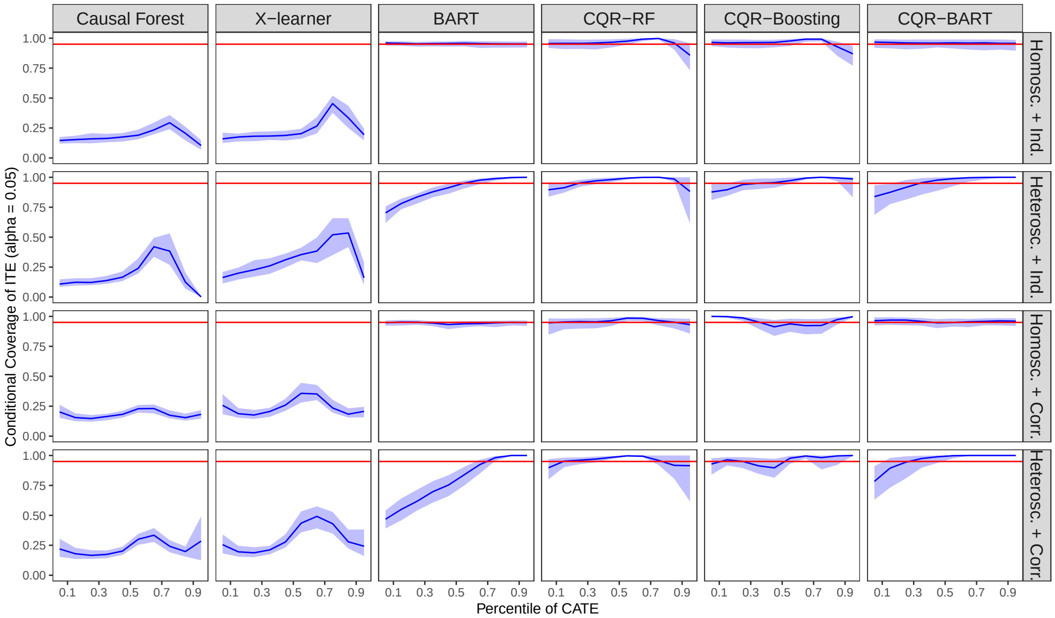

Figure 10: Estimated conditional coverage of ITE as a function of the CATE $\tau(x)$ for all scenarios with $d = 100$ from Section 3.6. Everything else is as in Figure 8. The difference with Figure 9 is the value of the dimension d. Yet, we can observe a very similar behavior.