Treatment Effect Risk: Bounds and Inference

NATHAN KALLUS, Netflix and Cornell University

Since the average treatment effect (ATE) measures the change in social welfare, even if positive, there is a risk of negative effect on, say, some 10% of the population. Assessing such risk is difficult, however, because any one individual treatment effect (ITE) is never observed so the 10% worst-affected cannot be identified, while distributional treatment effects only compare the first deciles within each treatment group, which does not correspond to any 10%-subpopulation. In this paper we consider how to nonetheless assess this important risk measure, formalized as the conditional value at risk (CVaR) of the ITE-distribution. We leverage the availability of pre-treatment covariates and characterize the tightest-possible upper and lower bounds on ITE-CVaR given by the covariate-conditional average treatment effect (CATE) function. We then proceed to study how to estimate these bounds efficiently from data and construct confidence intervals. This is challenging even in randomized experiments as it requires understanding the distribution of the unknown CATE function, which can be very complex if we use rich covariates so as to best control for heterogeneity. We develop a debiasing method that overcomes this and prove it enjoys favorable statistical properties even when CATE and other nuisances are estimated by black-box machine learning or even inconsistently. Studying a hypothetical change to French job-search counseling services, our bounds and inference demonstrate a small social benefit entails a negative impact on a substantial subpopulation.

Additional Key Words and Phrases: Program evaluation, Individual treatment effect, Conditional average treatment effect, Conditional value at risk, Partial identification, Debiased machine learning

ACM Reference Format:
Nathan Kallus. 2022. Treatment Effect Risk: Bounds and Inference. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3531146.3533087

1 INTRODUCTION

Policymakers and project managers regularly conduct randomized experiments ("A/B tests") to assess potential changes to policy or product. A key metric is the *average treatment effect (ATE)*, the difference in the population-average outcome when everyone or no one is treated. ATEs are easily estimated by differences in the sample-average outcome within treatment groups, barring interference. Estimation from observational data is also possible under appropriate assumptions, *e.g.*, unconfoundedness [27]. Identifying an individual's outcome with their utility – as we will throughout this paper – the ATE is the difference in social welfare in these two counterfactual scenarios. By linearity, this coincides with the population-average of each *individual*'s treatment effect, the difference in their own utility in the two counterfactual scenarios.

It is widely recognized, however, that treatment effects can vary widely between individuals [15, 25]. Thus, even if the ATE is positive, there is a *risk* that many individuals are harmed by the proposed change. Crucially, *distributional* treatment effects (DTEs), which compare the two counterfactual utility distributions beyond their means, *cannot* capture this risk. Indeed, Imbens and Wooldridge [28] note "quantile effects are defined as differences between quantiles of the two marginal potential outcome distributions, and not as quantiles of the unit level effect." They nonetheless advocate for the former because policy "choice should be governed by preferences of the policymaker over these distributions."

Manuscript submitted to ACM

However, such rational-decision-making framing presumes a policymaker facing a choice between lotteries drawing at random from individual outcomes. Instead, concerned with equity beyond social welfare, we should worry about the individuals, not the policymaker. Hypothetically, harm to some individuals is possible even when the "treat-all" utility distribution first-order-dominates "treat-none" so that *any* expected-increasing-utility-function-maximizer would choose "treat-all."

One way to gain further insight into heterogeneity and hence inequities is to consider conditional ATEs (CATEs) given pre-treatment covariates. For example, if we observe a discrete sensitive attribute (*e.g.*, race), we can simply compare the CATE in each attribute-value group.[1] But it may not always be clear what are relevant such attributes and whether we are omitting important ones. Given rich and continuous covariates, we can still reliably learn the CATE function by leveraging recent advances in causal machine learning [4, 26, 38, 42, 45, 56]. It may still not be clear, nonetheless, whether the covariates are relevant for fairness considerations, what groups are captured in this way, and/or how to summarize the many individual predictions of complex machine-learned CATEs.

It is therefore particularly appealing to focus directly on the distribution of *individual* treatment effects (ITEs), such as the average effects among the worst-affected 10%, 20%, *etc.*, corresponding to the conditional value at risk (CVaR) of this distribution. The challenge is that no ITE can ever be observed – the so-called Fundamental Problem of Causal Inference. Nonetheless, regardless of whether covariates are meaningful for fairness considerations, if they control for heterogeneity, CATE may predict ITE well. In this paper, we leverage this to proxy these important but unidentifiable treatment-effect risk measures. Specifically, we provide the tightest-possible upper and lower bounds given by CATE on the CVaR of ITE. By construction these are functions of distributions of observables. What remains is inference from data, whether experimental or observational. Since the CATE function can be high-dimensional, especially if we use a lot of covariates to control for heterogeneity, inference is difficult and naïve plug-in approaches fail. We design debiased estimators and confidence intervals for our bounds that overcome this challenge by being exceedingly robust: given rough, machine-learned estimates of CATE and other nuisances, they behave as though we used perfect estimates; they remain consistent even when some nuisances are mis-estimated; and surprisingly they remain valid as bounds even when CATE is mis-estimated. We conclude by using our tools to illustrate treatment-effect risk in a case study of job-search-assistance benefits.

2 PROBLEM SET UP AND DEFINITIONS

Each individual in the population is associated with two potential outcomes, $Y^*(0)$, $Y^*(1) \in \mathbb{R}$, corresponding to individual utility under "treat-all" and "treat-none," respectively, and baseline covariates (observable characteristics), $X \in \mathcal{X}$. The ITE, ATE, and CATE are, respectively,

$$\delta = Y^*(1) - Y^*(0), \qquad \bar{\tau} = \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] = \mathbb{E}\delta = \mathbb{E}\tau(X)$$
$$\tau(X) = \mathbb{E}[\delta \mid X] = \mu(X,1) - \mu(X,0), \quad \text{where } \mu(X,a) = \mathbb{E}[Y^*(a) \mid X].$$

We assume $\mathbb{E}\delta^2 < \infty$ throughout.

Of interest is the average effect among the $(100 \times \alpha)$%-worst affected, formalized by $\mathrm{CVaR}_\alpha(\delta)$, where for any Z [49][2]

$$\mathrm{CVaR}_\alpha(Z) = \sup_\beta \left(\beta + \frac{1}{\alpha}\mathbb{E}(Z-\beta)_- \right), \tag{1}$$

[1] We may still make some inferences on these even if we do not observe such attributes; see [11, 33].
[2] CVaR is sometimes defined for the right tail, corresponding to our $-\mathrm{CVaR}_\alpha(-Z)$.

where $(u)_- = u \wedge 0$. The sup is attained by β equal the α-quantile:

$$F_Z^{-1}(\alpha) = \inf\{\beta : F_Z(\beta) \geq \alpha\}, \quad \text{where } F_Z(z) = \mathbb{P}(Z \leq z). \tag{2}$$

Provided $F_Z(F_Z^{-1}(\alpha)) = \alpha$ (*e.g.*, Z continuous), then $\mathrm{CVaR}_\alpha(Z) = \mathbb{E}\left[Z \mid Z \leq F_Z^{-1}(\alpha)\right]$. Otherwise, $\mathrm{CVaR}_\alpha(Z) \in [\mathbb{E}\left[Z \mid Z < F_Z^{-1}(\alpha)\right], \mathbb{E}\left[Z \mid Z \leq F_Z^{-1}(\alpha)\right]]$, and, unlike these two endpoints, $\mathrm{CVaR}_\alpha(Z)$ is continuous in α and coherent [2]. It is therefore the *correct* generalization of "average of the $(100 \times \alpha)$%-lowest values" when ambiguous due to discontinuities.

We consider data from a randomized experiment or observational study. Each individual is associated with a treatment $A \in \{0, 1\}$, and we observe the *factual* outcome $Y = Y^*(A)$ (never $Y^*(1-A)$). The data is $(X_i, A_i, Y_i) \sim (X, A, Y)$, $1 \leq i \leq n$. We assume unconfoundedness throughout: $Y^*(a) \perp\!\!\!\perp A \mid X$.[3] Randomized experiments (our focus) ensure this by design (often with $X \perp\!\!\!\perp A$). Our results nonetheless extend to observational settings assuming unconfoundedness. Under unconfoundedness, ATE and CATE are identifiable, *i.e.*, are functions of the (X, A, Y)-distribution: $\mu(X, a) = \mathbb{E}[Y \mid X, A = a]$, $\tau(X) = \mu(X, 1) - \mu(X, 0)$, $\bar{\tau} = \mathbb{E}\tau(X)$ $(= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$ if $X \perp\!\!\!\perp A)$. Define also the propensity score $e(X) = \mathbb{P}(A = 1 \mid X)$ and marginal-outcome regression $\bar{\mu}(X) = \mathbb{E}[Y \mid X] = e(X)\mu(X, 1) + (1 - e(X))\mu(X, 0)$.

We now illustrate treatment-effect risk and its *un*identifiability, which motivates us to consider the tightest-possible *identifiable* bounds (Section 3) and inference thereon (Section 4).

Example 2.1 (Simple Example). Suppose

$$\begin{pmatrix} Y^*(0) \\ Y^*(1) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(0) \\ \mu(1) \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \ \mu(1) \geq \mu(0), \ \rho \in [-1, 1].$$

If $\bar{\tau} = \mu(1) - \mu(0) > 0$, the $Y^*(1)$-distribution *first-order-dominates* $Y^*(0)$. If $\mu(1) = \mu(0)$, the distributions are *indistinguishable*. However, the ITE-distribution depends on ρ: $\delta \sim \mathcal{N}(\mu(1) - \mu(0), \sqrt{2 - 2\rho})$, $\mathrm{CVaR}_{0.1}(\delta) = \bar{\tau} - 1.75\sqrt{2 - 2\rho}$. The unidentifiability of $\mathrm{CVaR}_{0.1}(\delta)$ follows because the (A, Y)-distribution is fixed given just $\mu(0), \mu(1), \mathbb{P}(A = 1)$ while $\mathrm{CVaR}_{0.1}(\delta)$ varies with ρ.

Remark 1 (Covariate-conditional policies)**.** Treat (*i.e.*, rollout to) all or none is often the choice faced by project managers, but given covariates we can learn covariate-conditional treatment policies [5, 31, 36, 40, 46, 57]. Learning aside, treating only when $\tau(X) > 0$ ensures all covariate-defined groups have nonnegative group-average effects.[4] Personalizing on all available covariates is however generally infeasible due to operational, non-stationarity, and/or ethical/reputational concerns. Nonetheless, given any policy $\pi : \mathcal{X} \to \{0, 1\}$, we may simply redefine ITE as $Y(\pi(X)) - Y(0)$ and our results still apply. This is especially relevant when π personalizes on some covariates and the rest explain heterogeneity conditionally thereon.

Remark 2 (Risk of observed vs unobserved variables)**.** CVaR is an example of coherent risk measures [2], which are used to assess distributions beyond expectations and are equivalent to distributionally-robust worst-case expectations [51]. For example, CVaR is the worst-case expectation among distributions with Radon-Nikodym derivative to the given distribution bounded by $1/\alpha$. Other distributional divergences can also define ambiguity sets [*e.g.*, 9, 10, 20]. Alternative approaches limit the *complexity* of subpopulations [37, 43]. In both finance [41], distributionally-robust supervised learning [6], demographics-free fair learning [43], and CVaR-DTEs [32], the variable whose risk is of interest is *always observed. E.g.*, model loss on each training example is observed. In contrast, we consider risk of an *unobserved variable*,

[3] And $Y = Y^*(A)$ assumes non-interference [50].
[4] However, even this ideal can induce disparate impacts [35].

hence we study bounds in Section 3. For inference, we are uniquely concerned with risk of an *unknown function*, hence we develop learning-robust methods in Section 4.

3 BOUNDS

3.1 Upper Bound: The CATE-CVaR

An upper bound on $\mathrm{CVaR}_\alpha(\delta)$ is crucial: if negative or substantially below ATE, the change poses certifiable risk or inequity to an $(100\times\alpha)\%$-subpopulation.

Theorem 3.1 (Upper Bound by CATE-CVaR).

$$\mathrm{CVaR}_\alpha(\delta) \le \mathrm{CVaR}_\alpha(\tau(X)). \tag{3}$$

Moreover, given any X-distribution and integrable $\tau : \mathcal{X} \to \mathbb{R}$, some (X,δ)-distribution has the given X-marginal, $\tau(X) = \mathbb{E}[\delta \mid X]$, and Eq. (3) holding with equality.

Since $\tau(X)$ represents our *best guess* for δ (in squared error), imputing the unknown δ with $\tau(X)$ seems reasonable. Theorem 3.1 shows this in fact provides an upper bound.[5] If $\tau(X)$ is continuous, $\mathrm{CVaR}_\alpha(\tau(X)) = \mathbb{E}[\delta \mid \tau(X) \le F^{-1}_{\tau(X)}(\alpha)]$, and Eq. (3) is intuitive: $\mathrm{CVaR}_\alpha(\delta)$ is worst average effect among *all* $(100\times\alpha)\%$-subpopulations, while $\mathrm{CVaR}_\alpha(\tau(X))$ only among X-defined subpopulations. This bound is also tight: given just $\tau(X)$, it cannot be improved.[6]

Theorem 3.1 implies an ordering:

$$\mathrm{CVaR}_{\alpha_1}(\delta) \le \mathrm{CVaR}_{\alpha_2}(\delta) \le \mathrm{CVaR}_{\alpha_2}(\tau(X)) \le \mathrm{CVaR}_{\alpha_3}(\tau(X)) \le \bar\tau \;\forall\, 0 < \alpha_1 \le \alpha_2 \le \alpha_3 \le 1. \tag{4}$$

Remark 3 (CVaR as summary of CATE)**.** Aside from being a bound, $\mathrm{CVaR}_\alpha(\tau(X))$ is of independent interest as a summary of effect heterogeneity along meaningful covariates X of explicit equity concern. When X is more than a few discrete groups, understanding the many facets of estimated heterogeneity is challenging, both interpretationally and statistically. We could test for X-heterogeneity [15, 16, 22, 52].[7] *E.g.*, omnibus test $H_0 : 0 \in \operatorname{argmin}_\gamma \mathbb{E}(\tau(X) - \bar\tau - \gamma^\top(X - \mathbb{E}X))^2$ [13]. This, however, may detect minor heterogeneity in small subpopulations, may not assess magnitude or direction, and may be inappropriate if we expect heterogeneity. In contrast, $\mathrm{CVaR}_\alpha(\tau(X))$ is a simple, meaningful summary of $\tau(X)$. Inference, however, is a challenge. We tackle this in Section 4.

Remark 4 (*Who* is negatively affected?)**.** Suppose we find $\mathrm{CVaR}_\alpha(\tau(X)) < 0$ while $\bar\tau > 0$, where α is "substantial" – the social-welfare benefit of the proposal is borne by some substantial negatively-impacted subpopulation. While that may already cool enthusiasm for the proposal, we may wonder *who* are the harmed individuals, *e.g.*, to help design a new, better treatment.

Assuming continuity, $\mathrm{CVaR}_\alpha(\tau(X))$ is the ATE among individuals with $\tau(X) \le F^{-1}_{\tau(X)}(\alpha)$ – an *identifiable* group. A question is interpretation. This is easy if $\tau(X)$ is linear or tree (or estimated using such models, which still gives a bound per Theorem 4.5). We can also consider summaries of this group, *e.g.*, fraction belonging to sensitive groups, or learn simpler models to explain membership [44, 47]. Alternatively, given we detect substantial inequities, we can *separately* investigate which variables negatively modulate treatment effect by, *e.g.*, studying $\operatorname{argmin}_\gamma \mathbb{E}(\tau(X) - \bar\tau - \gamma^\top X)^2$ [13, 38].

[5] Equation (3) extends to any coherent risk by writing $\delta = \tau(X) + (\delta - \tau(X))$ and using sub-additivity.
[6] The bound need not be tight given the (X, A, Y)-distribution, which characterizes more than the mean of the $(\delta \mid X)$-distribution, as described by the Fréchet-Hoeffding bounds. We focus on best bounds given just by CATE, which is the common tool to understand effect heterogeneity in practice.
[7] There are also tests for heterogeneity *not* explained by X [17, 18]. These, like us, leverage bounds on unidentifiable quantities.

3.2 Lower Bounds under Limited Residual Heterogeneity Range

Much as we try to best control for heterogeneity, disparate effect-predictiveness of covariates may mean some negative ITEs are averaged out and hidden while others are singled out. A remedy when concerned about disproportionate predictiveness among sensitive groups (*e.g.*, race) would be to include these (or proxies) within X. But, we may always worry about missing something. A lower bound can provide assurances about what the upper bound may be missing.

This depends on how much residual heterogeneity remains. Our first set of lower bounds limit the range of residual heterogeneity, *i.e.*, almost-sure bounds on $\delta - \tau(X)$, while our second set of lower bounds limit its variance, *i.e.*, bounds on $\mathrm{Var}(\delta \mid X) = \mathbb{E}(\delta - \tau(X))^2$.

Theorem 3.2. *Suppose* $|\tau(X) - \delta| \le b$. *Then*

$$\mathrm{CVaR}_\alpha(\delta) \ge \sup_\beta \left(\beta + \frac{1}{2\alpha}\mathbb{E}[(\tau(X) - b - \beta)_-] + \frac{1}{2\alpha}\mathbb{E}[(\tau(X) + b - \beta)_-] \right). \tag{5}$$

Moreover, given any X-distribution and integrable $\tau : \mathcal{X} \to \mathbb{R}$, some (X, δ)-distribution has the given X-marginal, $\tau(X) = \mathbb{E}[\delta \mid X]$, $|\tau(X) - \delta| \le b$, and Eq. (5) holding with equality.

The right-hand side of Eq. (5) is the α-CVaR of the equal-mixture distribution of $\tau(X) - b$ and $\tau(X) + b$. It reduces to $\mathrm{CVaR}_\alpha(\tau(X))$ when $b = 0$ (equivalent to $\delta = \tau(X)$). When $\alpha = 1$, it becomes $\bar\tau$ for any $b \ge 0$ (as necessary for tightness). The lower bound is established via weak semi-infinite duality and its tightness by exhibiting the equal-mixture distribution.

Since $(\tau(X) \pm b - \beta)_- \ge (\tau(X) - \beta)_- - b$, Eq. (5) upper bounds $\mathrm{CVaR}_\alpha(\tau(X)) - b$. This simpler bound is tight if we only assume a one-sided-bounded range.

Theorem 3.3. *Suppose* $\tau(X) - \delta \le b$. *Then*

$$\mathrm{CVaR}_\alpha(\delta) \ge \mathrm{CVaR}_\alpha(\tau(X)) - b. \tag{6}$$

Moreover, for $\alpha < 1$, given any $\varepsilon > 0$, X-distribution, and integrable $\tau : \mathcal{X} \to \mathbb{R}$, some (X, δ)-distribution has the given X-marginal, $\tau(X) = \mathbb{E}[\delta \mid X]$, $\tau(X) - \delta \le b$, and Eq. (6) holding with equality up to ε-error.

The lower bound is immediate and its tightness given by exhibiting a skewed two-point-mass distribution. For $\alpha = 1$, Eq. (6) simply reads $\bar\tau \ge \bar\tau - b$, but for *any* $\alpha < 1$, Eq. (6) is actually *tight*.

3.3 Lower Bounds under Limited Residual Heterogeneity Variance

Limiting residual heterogeneity within a range may be implausible, or plausible only with large constants, yielding a weak bound. We next explore the implication of the residual ITE-variance after controlling for X, which we can bound given observables.

Theorem 3.4. *Suppose* $\mathrm{Var}(\delta \mid X) \le \bar\sigma^2(X)$ *for some integrable* $\bar\sigma^2 : \mathcal{X} \to \mathbb{R}_+$. *Then*

$$\mathrm{CVaR}_\alpha(\delta) \ge \sup_\beta \left(\beta + \frac{1}{2\alpha}\mathbb{E}\left[\tau(X) - \beta - \sqrt{(\tau(X) - \beta)^2 + \bar\sigma^2(X)} \right] \right). \tag{7}$$

Moreover, given any $\varepsilon > 0$, X-distribution, and integrable $\tau : \mathcal{X} \to \mathbb{R}$, some (X, δ)-distribution has the given X-marginal, $\tau(X) = \mathbb{E}[\delta \mid X]$, $\mathrm{Var}(\delta \mid X) \le \bar\sigma^2(X)$, and Eq. (7) holding with equality up to ε-error.

The proof of Theorem 3.4 leverages strong duality for convex semi-infinite optimization. Note Eq. (7) equals $\mathrm{CVaR}_\alpha(\tau(X))$ whenever $\bar\sigma^2(X) = 0$ and $\bar\tau$ whenever $\alpha = 1$. Since $|\delta - \tau(X)| \le b \implies \mathrm{Var}(\delta \mid X) \le b^2$, plugging

$\bar{\sigma}^2(X) = b^2$ into Eq. (7) must be looser than Eq. (5) by tightness. Triangle inequality verifies this directly: $\sum_{\pm}(\tau(X) \pm b - \beta)_- = \tau(X) - \beta - \frac{1}{2}\sum_{\pm}|\tau(X) \pm b - \beta| \geq \tau(X) - \beta - \sqrt{(\tau(X)-\beta)^2 + b^2}$.

A residual-variance bound is both more plausible and easier to calibrate than an absolute bound. Letting $\rho(X) = \mathrm{Corr}(Y(0), Y(1) \mid X) \in [-1, 1]$, we have

$$\mathrm{Var}(\delta \mid X) = \mathrm{Var}(Y \mid X, A=0) + \mathrm{Var}(Y \mid X, A=1) - 2\rho(X)\,\mathrm{Var}^{1/2}(Y \mid X, A=0)\,\mathrm{Var}^{1/2}(Y \mid X, A=1), \tag{8}$$

where all terms but $\rho(X)$ are identifiable. Thus, postulating different potential-outcome correlations, we obtain different bounds. Equation (8) is maximized for $\rho(X) = -1$, which is tight, as all correlations are realizable. Thus, plugging $\bar{\sigma}^2(X) = (\mathrm{Var}^{1/2}(Y \mid X, A=0) + \mathrm{Var}^{1/2}(Y \mid X, A=1))^2$ into Eq. (7) yields a tight lower bound on ITE-CVaR, given conditional expectations and variances. We may obtain better bounds if we postulate larger $\rho(X)$.

Theorem 3.4 also implies a simpler but looser bound.

Corollary 3.5.

$$0 \leq \mathrm{CVaR}_\alpha(\tau(X)) - \mathrm{CVaR}_\alpha(\delta) \leq \frac{1}{2\alpha}\mathbb{E}\left[\mathrm{Var}^{1/2}(\delta \mid X)\right] \tag{9}$$

$$\leq \frac{1}{2\alpha}\mathbb{E}\left[\mathrm{Var}^{1/2}(Y \mid X, A=0) + \mathrm{Var}^{1/2}(Y \mid X, A=1)\right] \tag{10}$$

$$\leq \frac{1}{2\alpha}\sqrt{\mathbb{E}\left[(Y-\mu(X,A))^2 \mid A=0\right]} + \frac{1}{2\alpha}\sqrt{\mathbb{E}\left[(Y-\mu(X,A))^2 \mid A=1\right]}. \tag{11}$$

Equation (9) more transparently bounds the slack in Eq. (3) in terms of residual effect variance. However, it is not tight, as can be seen for $\alpha = 1$. Equation (11) is even looser but appealing as it avoids $\mathrm{Var}(Y \mid X, A)$, depending only on the root-mean-squared error of regressing Y on X for each $A \in \{0, 1\}$ (*i.e.*, the numerator of nonparametric R^2).

4 INFERENCE

We next turn to estimating the bounds developed in Section 3 and constructing confidence intervals. Recall our data $(X_i, A_i, Y_i) \sim (X, A, Y)$, $1 \leq i \leq n$, may be experimental or observational. The only relevant technical difference between these two cases is whether propensity, $e(X) = \mathbb{P}(A = 1 \mid X)$, is known or not. While it matters not here, note that $e(X)$ is usually constant in experiments ($A \perp\!\!\!\perp X$). In observational settings $e(X)$ may be estimated.

We focus here on inference on CATE-CVaR. We provide analogous procedures for the lower bounds of Theorems 3.2 to 3.4 and Corollary 3.5 in Appendix A. Fix α. Our inferential target is

$$\Psi = \mathrm{CVaR}_\alpha(\tau(X)) = \beta^* + \frac{1}{\alpha}\mathbb{E}(\tau(X) - \beta^*)_-, \quad \text{where } \beta^* = F^{-1}_{\tau(X)}(\alpha) = \inf\{\beta : \mathbb{P}(\tau(X) \leq \beta) \geq \alpha\}.$$

Since $\tau(X)$ is not directly observed, the first step is fitting it. Fortunately, recent advances in causal machine learning provide excellent tools for this [4, 26, 38, 42, 45, 56]. Given an estimate $\hat{\tau}$, we might consider a plug-in approach: $\hat{\Psi}^{\text{plug-in}} = \sup_\beta(\beta + \frac{1}{n\alpha}\sum_{i=1}^n(\hat{\tau}(X_i) - \beta)_-)$. Unfortunately, the statistical behavior of $\hat{\Psi}^{\text{plug-in}}$ depends heavily on that of $\hat{\tau}$: if $\hat{\tau}$ converges slowly and/or has non-negligible bias, as occurs when fit by flexible machine-learning methods, both estimation rates and valid inference may be imperiled for $\hat{\Psi}^{\text{plug-in}}$.

Instead, we develop a debiasing approach that is *insensitive* to CATE-estimation, accommodating both misspecified parametric models and flexible-but-imprecise machine-learning CATE-estimators. The main challenge is estimating β^*, which cannot be expressed by an estimating equation in $X, Y(0), Y(1)$, so its efficient/orthogonal estimation is unclear, unlike the case of quantile/CVaR treatment effects [8, 21, 32]. Fortunately, we care only about Ψ, not β^*, and special optimization structure in Ψ gives robustness to perturbations. so even rough estimates suffice. Our approach is therefore

Algorithm 1 Point estimate and confidence interval for $\mathrm{CVaR}_\alpha(\tau(X))$

Input: Level $\alpha \in (0,1)$, data $\{(X_i, A_i, Y_i) : i = 1, \dots, n\}$, number of folds K, e, μ, τ-estimators
1: **for** $k = 1, \dots, K$ **do**
2: Estimate $\hat{e}^{(k)}, \hat{\mu}^{(k)}, \hat{\tau}^{(k)}$ using data $\{(X_i, A_i, Y_i) : i \not\equiv k-1 \pmod K\}$
3: Set $\hat{\beta}^{(k)} = \inf\{\beta : \sum_{i \not\equiv k-1 \pmod K} (\mathbb{I}[\hat{\tau}^{(k)}(X_i) \le \beta] - \alpha) \ge 0\}$
4: **for** $i \equiv k-1 \pmod K$ **do** set $\phi_i = \phi(X_i, A_i, Y_i; \hat{e}^{(k)}, \hat{\mu}^{(k)}, \hat{\tau}^{(k)}, \hat{\beta}^{(k)})$
5: **end for**
6: Set $\hat{\Psi} = \frac{1}{n}\sum_{i=1}^n \phi_i$, $\hat{\mathrm{se}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^n (\phi_i - \hat{\Psi})^2}$
7: Return $\hat{\Psi}$ as point estimate and $[\hat{\Psi} \pm \Phi^{-1}((1+\gamma)/2)\hat{\mathrm{se}}]$ as γ-confidence intervals

unique: we treat both τ and β^* as nuisance parameters, together with e, μ, and ensure simultaneous orthogonality to all four nuisances.

Algorithm 1 summarizes our procedure. It proceeds by approximating the sample average of $\Psi = \mathbb{E}\phi(X, A, Y, e, \mu, \tau, \beta^*)$, where, we define

$$\phi(X, A, Y; \check{e}, \check{\mu}, \check{\tau}, \check{\beta}) = \check{\beta} + \frac{1}{\alpha}\mathbb{I}\left[\check{\tau}(X) \le \check{\beta}\right]\left(\check{\mu}(X,1) - \check{\mu}(X,0) + \frac{A - \check{e}(X)}{\check{e}(X)(1-\check{e}(X))}\left(Y - \check{\mu}(X,A)\right) - \check{\beta}\right). \tag{12}$$

We first estimate the unknown (e, μ, τ, β^*). We do so using "cross-fitting" over K even folds so that nuisance estimates are independent of samples where applied [12, 53, 58].[8] As we discuss in detail in Section 4.3, we treat τ as a separate nuisance even though $\tau(x) = \mu(x,1) - \mu(x,0)$. For one, this enables the use of specialized CATE-learners. We also treat β^* as a separate nuisance (not as a parameter as in [32]) and fit it as the quantile of $\hat{\tau}(X)$ in the out-of-fold data. As simple regressions, e and μ can be fit by parametric regression or standard machine-learning methods such as random forests, gradient boosting, neural networks, *etc.*.

Remark 5 (Comparing different levels)**.** To assess disparities, we may want to compare to ATE (equivalently, $\mathrm{CVaR}_1(\tau(X))$). To get good confidence intervals on $\mathrm{CVaR}_\alpha(\tau(X)) - \mathrm{CVaR}_{\alpha'}(\tau(X))$, we can replace ϕ_i in Line 4 of Algorithm 1 with the difference of ϕ_i's for α and α' (using the same nuisances except $\hat{\beta}^{(k)}$). This will correctly yield smaller confidence intervals on $\bar{\tau} - \mathrm{CVaR}_\alpha(\tau(X))$ for α closer and closer to 1. We may also consider covariances of ϕ_i's corresponding to many α-levels for constructing simultaneous intervals.

Remark 6 (Partial-identification intervals)**.** While Algorithm 1 focuses on CATE-CVaR, which upper bounds ITE-CVaR, in Appendix A we provide inference procedures for lower bounds on ITE-CVaR. These can be combined to construct intervals containing ITE-CVaR with probability γ. By union bound, we can simply combine the one-sided $(1+\gamma)/2$-confidence intervals for the lower and upper bounds. But coverage may be conservative ($> \gamma$) for the partial-identification interval given by the bounds. For calibrated γ-coverage (asymptotically), we must account for correlation between lower- and upper-bound estimates, given by the correlation between ϕ_i's for each procedure. Then, we can construct calibrated intervals following Appendix A.4 of Kallus et al. [33].

Remark 7 (Monotonicity)**.** While $\mathrm{CVaR}_\alpha(\tau(X))$ is monotone in α, Algorithm 1's output for different α may not be due to estimation errors. We can post-process to ensure monotonicity using rearrangement [24], which only improves estimation and does not affect inference [14]. We use this in Section 5.

[8] We may avoid cross-fitting and fit nuisances once on the whole sample if we assume estimates belong to a Donsker class with probability tending to 1; we omit this option for brevity.

4.1 Local Robustness and Confidence Intervals

We now establish favorable guarantees for Algorithm 1. First, we show it is insensitive to slow but consistent estimation of nuisances, having first-order behavior as if we used true values.

We will need some minimal regularity.

Assumption 1 (Regularity). $\bar{e} \le e \le 1-\bar{e}$ and $|Y| \le B$ for positive constants $\bar{e}, B > 0$. $F_{\tau(X)}$ is continuously differentiable at $F^{-1}_{\tau(X)}(\alpha)$.

The first condition ensures that the X-distributions of experimental groups *overlap*. It is usually guaranteed in randomized experiments by setting $e(X)$ constant ($A \perp\!\!\!\perp X$). In unconfounded observational studies, it is a standard assumption. The second condition requires bounded outcomes and is largely technical to make analysis tractable. The third condition prohibits degeneracy of the quantile. The same is needed for asymptotic normality of sample quantiles of *observed* variables. If $\tau(X)$ is discrete, the condition may be replaced by $\exists \varepsilon > 0 : F^{-1}_{\tau(X)}(\alpha-\varepsilon) = F^{-1}_{\tau(X)}(\alpha+\varepsilon)$, yielding superefficient quantile estimation. The only problematic case is multiplicity of $\{\beta : F_{\tau(X)}(\beta) = \alpha\}$, but only finitely-many such "bad" α's exist. Since the focus is on X being rich, we focus on the continuous case and the condition in Assumption 1.

We first show how, under Assumption 1, estimation rates for $\hat{\tau}^{(k)}$ translate to rates for $\hat{\beta}^{(k)}$.

Lemma 4.1. *Suppose Assumption 1 holds. Then, for each $k = 1, \dots, K$, $\hat{\beta}^{(k)}$ in Line 3 of Algorithm 1 satisfies*

$$|\hat{\beta}^{(k)} - \beta^*| = O_p(n^{-1/2} \vee \|\hat{\tau}^{(k)} - \tau\|_r^{\frac{r}{r+1}}) \quad \forall r \in [1, \infty].$$

We now show robust oracle-like behavior for $\hat{\Psi}$.

Theorem 4.2. *Suppose Assumption 1 holds and that for $k = 1, \dots, K$, $\|\hat{e}^{(k)} - e\|_2 = o_p(1)$, $\|\hat{\mu}^{(k)} - \mu\|_2 = o_p(1)$, $\|\hat{e}^{(k)} - e\|_2 \|\hat{\mu}^{(k)} - \mu\|_2 = o_p(n^{-\frac{1}{2}})$, $\|\hat{\tau}^{(k)} - \tau\|_\infty = o_p(n^{-\frac{1}{4}})$, $\mathbb{P}(\|\hat{\mu}^{(k)}\|_\infty \le B) \to 1$, and $\mathbb{P}(\bar{e} \le \hat{e}^{(k)} \le 1-\bar{e}) \to 1$. Then $\hat{\Psi}$, $\hat{\text{se}}$ in Line 6 of Algorithm 1 satisfy*

$$\hat{\Psi} = \frac{1}{n}\sum_{i=1}^n \phi(X, A, Y; e, \mu, \tau, \beta^*) + o_p(n^{-1/2}) = \Psi + O_p(n^{-1/2}),$$
$$\mathbb{P}(\Psi \in [\hat{\Psi} \pm \Phi^{-1}((1+\gamma)/2)\hat{\text{se}}]) \to \gamma \ \forall \gamma.$$

The rate assumptions on e and μ are lax: it suffices to have $o_p(n^{-1/4})$-rates on both or no rate on μ at all if e is known. This parallels standard conditions in double-machine-learning ATE-estimation, achievable by a variety of machine-learning methods [12]. We explore the condition on τ in Section 4.3.

4.2 Double Robustness and Double Validity

Theorem 4.2 guarantees good performance if all nuisances are estimated slowly, but still consistently. But even if nuisances are inconsistent, we perform well.

First, we establish a property mirroring doubly-robust ATE-estimation [48]: even if e or μ are inconsistent, we remain consistent, provided τ is consistently estimated, albeit slowly.

Theorem 4.3 (Double robustness). *Fix any $\tilde{e}, \tilde{\mu}$ with $\bar{e} \le \tilde{e} \le 1-\bar{e}$, $\|\tilde{\mu}\|_\infty \le B$. Let $r_n \to 0$ be a deterministic sequence. Suppose Assumption 1 holds and that for $k = 1, \dots, K$, $\|\hat{e}^{(k)} - \tilde{e}\|_2 = o_p(1)$, $\|\hat{\mu}^{(k)} - \tilde{\mu}\|_2 = o_p(1)$, $\|\hat{\tau}^{(k)} - \tau\|_\infty = O_p(r_n^{1/2})$, $\mathbb{P}(\|\hat{\mu}^{(k)}\|_\infty \le B) \to 1$, $\mathbb{P}(\bar{e} \le \hat{e}^{(k)} \le 1-\bar{e}) \to 1$, and*

$$\text{either} \quad \|\hat{e}^{(k)} - e\|_2 = O_p(r_n) \quad \text{or} \quad \|\hat{\mu}^{(k)} - \mu\|_2 = O_p(r_n).$$

Then $\hat{\Psi}$ in Line 6 of Algorithm 1 satisfies:

$$\hat{\Psi} = \Psi + O_p(r_n \vee n^{-1/2}).$$

Theorem 4.3 is particularly strong in experiments (e known): we can get away with $\hat{\mu}^{(k)} = 0$. We need only estimate CATE at $o_p(n^{-1/4})$-rates to ensure $O_p(n^{-1/2})$-consistency.

It would appear we must consistently estimate CATE to have hope of estimating its CVaR. While true, we next show that *even* if we mis-estimate CATE *and also* one of e, μ, we *still* get an upper bound on CATE-CVaR (hence on ITE-CVaR). This appears to be the second finding of a double-validity property since being first documented in sensitivity analysis [19].

We first establish the population-level bound behavior and then state the implication for estimation.

Lemma 4.4. *Fix any $\tilde{\tau} : \mathcal{X} \to \mathbb{R}$. Let $\tilde{\beta} = F^{-1}_{\tilde{\tau}(X)}(\alpha)$. Suppose Assumption 1 holds with τ replaced with $\tilde{\tau}$. Then:*

$$\mathrm{CVaR}_\alpha(\tau(X)) \leq \tilde{\beta} + \frac{1}{\alpha}\mathbb{E}[\mathbb{I}[\tilde{\tau}(X) \leq \tilde{\beta}](\tau(X) - \tilde{\beta})]. \tag{13}$$

Theorem 4.5 (Double validity). *Fix any $\tilde{e}, \tilde{\mu}, \tilde{\tau}$ with $\bar{e} \leq \tilde{e} \leq 1 - \bar{e}$, $\|\tilde{\mu}\|_\infty \leq B$, $\|\tilde{\tau}\|_\infty \leq 2B$. Let $r_n \to 0$ be a deterministic sequence. Suppose Assumption 1 holds with τ replaced with $\tilde{\tau}$ and that for $k = 1, \ldots, K$, $\|\hat{e}^{(k)} - \tilde{e}\|_2 = o_p(1)$, $\|\hat{\mu}^{(k)} - \tilde{\mu}\|_2 = o_p(1)$, $\|\hat{\tau}^{(k)} - \tilde{\tau}\|_\infty = O_p(r_n)$, $\mathbb{P}(\|\hat{\mu}^{(k)}\|_\infty \leq B) \to 1$, $\mathbb{P}(\bar{e} \leq \hat{e}^{(k)} \leq 1 - \bar{e}) \to 1$, and*

$$\textit{either} \quad \|\hat{e}^{(k)} - e\|_2 = O_p(r_n) \quad \textit{or} \quad \|\hat{\mu}^{(k)} - \mu\|_2 = O_p(r_n).$$

Then $\hat{\Psi}$ in Line 6 of Algorithm 1 satisfies:

$$\hat{\Psi} \geq \Psi - O_p(r_n \vee n^{-1/2}).$$

Theorem 4.5 guarantees extensive robustness and suggests a practical, blackbox-free approach in experimental settings: set $\hat{\mu}^{(k)} = 0$ and use simple *misspecified* parametric models (*e.g.*, linear) for CATE-estimation, and we still estimate a valid ITE-CVaR bound at fast $O_p(n^{-1/2})$-rates.

4.3 CATE-Estimation and Rates

Algorithm 1 accepts separate learners for *both* μ *and* τ. So, while $\tau(x) = \mu(x, 1) - \mu(x, 0)$, we need *not* have $\hat{\tau}^{(k)}(X) = \hat{\mu}^{(k)}(x, 1) - \hat{\mu}^{(k)}(x, 0)$, and in fact we should not. Recent work advocates and provides specialized methods for *directly* estimating CATE [4, 26, 38, 42, 45, 56].

This is important because Algorithm 1 uses the μ- and τ-estimates differently and, correspondingly, our theoretical results impose different assumptions on each. The τ-estimate is used for approximating the event $\mathbb{I}[\tau(X) \leq \beta^*]$, which is crucial for targeting CVaR correctly. In contrast, the μ-estimate is just used in order to estimate a weighted-average treatment effect, given the weights $\mathbb{I}[\tau(X) \leq \beta^*]$, and is therefore interchangeable with propensity.

We next review different options for CATE-estimation and how these ensure the conditions of Theorems 4.2, 4.3 and 4.5. We emphasize that these need not be understood as exhaustive list of which learners to use: practically, the nuisance-estimation rates are high-level assumptions that generally say one may safely plug-in black-box machine-learning estimators to Algorithm 1: no restrictions are made but rates (no metric-entropy conditions), estimators can be flexible/nonparametric in that rates can be much slower than "parametric" $O_p(n^{-1/2})$-rates, and results are exceedingly robust to inconsistent estimation.

4.3.1 Experimental settings. A major issue with CATE-estimation by differencing outcome regressions is that effect signals are easily lost. CATE is generally simpler and less variable than baseline mean outcomes, $\mu(X, 0), \mu(X, 1)$. For

example, many variables often help predict outcomes, but few modulate the treatment effect. It is therefore imperative to learn CATE directly.

In experimental settings (e known) we can construct a pseudo-outcome $\Delta = \frac{A-e(X)}{e(X)(1-e(X))}Y$ and, since $\tau(X) = \mathbb{E}[\Delta \mid X]$, learn CATE by regressing Δ on X, using any supervised-learning method. One case that theoretically ensures $\|\hat{\tau}^{(k)} - \tau\|_\infty = o_p(n^{-1/4})$ is when $\tau(x)$ is more-than-$d/2$-smooth in $x \in \mathbb{R}^d$ [54, Theorem 1]. Another option is $\tau(x)$ linear with $o(\sqrt{n}/\log d)$-nonzero coefficients [8]. Note this works *regardless* of μ being nice.

Or, we may avoid black-box models (and cross-fitting) altogether by using simple linear regression of Δ on X to obtain a valid bound per Theorem 4.5.

To satisfy the other conditions, for Theorems 4.3 and 4.5 we can set $\mu = 0$, and for Theorem 4.2 we need only estimate μ consistently without rate. We can either estimate μ directly or only estimate $\bar{\mu}(X) = \mathbb{E}[Y \mid X]$ and set $\hat{\mu}^{(k)}(X, A) = \hat{\bar{\mu}}^{(k)}(X) + (A - e(X))\hat{\tau}^{(k)}(X)$. Consistency for either is immediate from $\mathbb{E}Y^2 < \infty$ [23].

4.3.2 Observational settings. When e is unknown, the pseudo-outcome-construction needs refinement. One option is DR-leaner [38]: regress $\Delta = \hat{\mu}(X,1) - \hat{\mu}(X,0) + \frac{A-\hat{e}(X)}{\hat{e}(X)(1-\hat{e}(X))}(Y - \hat{\mu}(X,A))$ on X, where $\hat{e}, \hat{\mu}$ are appropriately cross-fitted. Another is R-learner [45]: let $\hat{\tau}$ minimize the average of $\left(Y - \hat{\bar{\mu}}(X) - (A - \hat{e}(X))\hat{\tau}(X)\right)^2$, where $\hat{e}, \hat{\bar{\mu}}$ are appropriately cross-fitted. Kennedy [38, Corollary 3] provides rates for local-polynomial R-learners: if $e(x)$ is s_e-smooth in $x \in \mathbb{R}^d$, $\bar{\mu}(x)$ s_μ-smooth, and $\tau(x)$ more-than-$d/2$-smooth, then we obtain $o_p(n^{-1/4})$-rate pointwise error, provided $s_e \geq s_\mu$, $\frac{s_e+s_\mu}{2} > \frac{d}{8}$. To convert pointwise-error bounds to sup-norm-error bounds, we may follow the discretization approach of Stone [54], incurring only logarithms. Or, we can simply use linear R- or DR-learners and get a valid bound per Theorem 4.5.

5 CASE STUDY

We now demonstrate our bounds and inference.[9] While we consider a program-evaluation example, we believe our results are also particularly relevant to A/B testing on online platforms, where, after testing, product innovations are usually either scrapped/reworked or broadly rolled out, and where ATEs are often small, creating an opportunity for many users to be negatively impacted despite positive average effects. Little such data is public, however.

5.1 Background and Setup

Behaghel et al. [7] analyze a large-scale randomized experiment comparing assistance programs offered to French unemployed individuals. They compare three arms: individuals in the "control" arm receive the standard services of the Public Employment Services, in "public" receive an intensive counseling program run by a public agency, and in "private" a similar program run by private agencies.

We consider a hypothetical scenario where the private-run counseling program ($A = 0$) is currently being offered to the unemployed and we consider the change to a public-run program ($A = 1$).[10] We take reemployment within six months as our (binary) outcome.

The ATE is 1.22 percentage points (90%-CI $[-0.35, 2.8]$), a 4.9% increase in reemployment. This suggests a positive/neutral effect, so a policymaker might hypothetically consider this an acceptable policy change, *e.g.*, if the public-run program provided cost savings.[11]

[9] Replication code is available at [ANONYMIZED].
[10] Some individuals assigned to the additional counseling refused it. We nonetheless restrict our attention to intent-to-treat interventions, considering hypothetically making available either the public-run or private-run counseling to unemployed individuals, who may decline it.
[11] Behaghel et al. [7, section IV] discuss why public-run programs fare better.

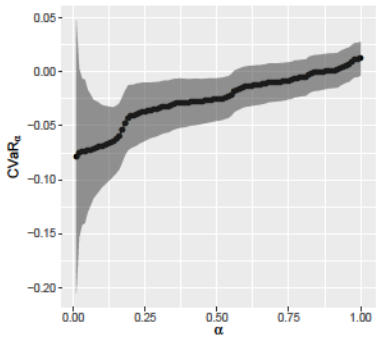

Fig. 1. $\mathrm{CVaR}_\alpha(\tau(X))$

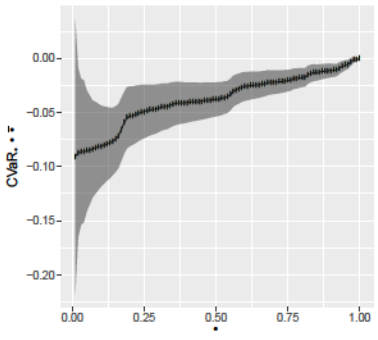

Fig. 2. $\mathrm{CVaR}_\alpha(\tau(X)) - \bar{\tau}$

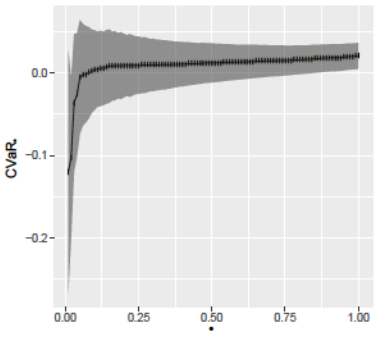

Fig. 3. $\mathrm{CVaR}_\alpha(\tau_1(X_1))$

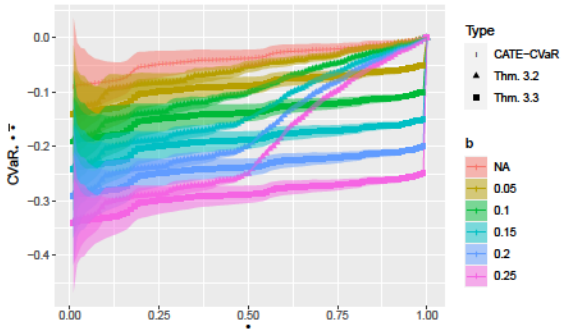

Fig. 4. Bounds based on residual-heterogeneity range

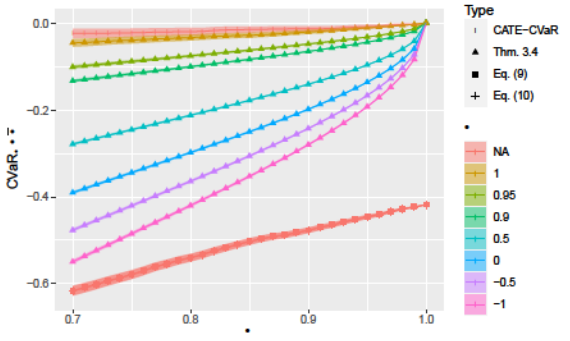

Fig. 5. Bounds based on residual-heterogeneity variance

To apply our methodology, we consider all pre-treatment covariates in table 2 of Behaghel et al. [7], except we treat as numeric (rather than dichotomize) age, number children, years experience, salary target, assignment timing, and number unemployment spells. Other variables quantify education, employment level and type, gender, martial status, national origin, region, unemployment reason, and long-term-unemployment risk. The propensity is constant. As recommended in Section 4.3.1, we fit CATE using a pseudo-outcome linear regression. We estimate μ using cross-fitted gradient-boosting machines.

5.2 Upper bounds

Figure 1 presents inference on CATE-CVaR using Algorithm 1 for $\alpha \in \{0.01, 0.02, \ldots, 1\}$. The line represents our point estimate, after rearrangement as recommended in Remark 7,[12] and the shaded region represents point-wise 90%-confidence intervals. Note uncertainty grows for smaller α.

We see that the ATE-estimate (right-most point) is positive with an interval containing zero. We find, however, that some 56%-sized X-defined-subpopulation has a negative effect at 90%-confidence.[13] This strongly suggests that the change, if enacted could materially negatively impact a large portion of the population, despite the positive/neutral

[12] We present the figure without rearrangement in Appendix B.

[13] Since outcome is binary, the *largest* fraction that can have a negative effect is $(50 \times (1 - \bar{\tau}))\%$, so either $\bar{\tau} < 0$ or at most half may be negatively affected. The ATE interval indeed contains zero with confidence only 90%.

ATE. Thus, considering treatment effect *risk* provides a crucial metric not reflected in the ATE. This risk is also *not* reflected in DTEs: the binary potential-outcome distributions are *fully* specified by just $\mathbb{E}[Y(0)]$, $\mathbb{E}[Y(1)]$.[14]

In Fig. 2 we focus on comparing CATE-CVaR to ATE following Remark 5. The only difference to Fig. 1 is a slight vertical shift and that confidence intervals (correctly) shrink to a point as $\alpha \to 1$, enabling more confident conclusions comparing subpopulations to the population.

In Fig. 3 we consider CATE-CVaR when we capture less heterogeneity, using only age, high-school dropout, African national origin, and Paris-region resident as covariates (X_1). This detects no significant risk.

5.3 Lower bounds

While the upper bounds show a significant subpopulation can be negatively harmed, being only bounds, it may be the subpopulation can be harmed even more or an even larger subpopulation can be harmed. Lower bounds help us understand how much greater the risk might be.

In Fig. 4 we consider our lower bounds (vs ATE) when limiting the residual-heterogeneity range given by Theorems 3.2 (two-sided range) and 3.3 (one-sided range).

Since it may be hard to justify and calibrate a limited range, in Fig. 5 we consider lower bounds given by Theorem 3.4 and Corollary 3.5 by limiting residual-heterogeneity variance. For the former, we fit $\mathrm{Var}(Y \mid A, X)$ using gradient-boosting machines and construct $\bar{\sigma}^2(X)$ per Eq. (8) by varying constant values of $\rho(X) = \rho \in [-1, 1]$. Recall $\rho = -1$ always yields an assumption-free bound. We use the same model to estimate the right-hand side of Eq. (10). We compute the cross-validated root-mean-squared prediction error to estimate the right-hand side of Eq. (11).

We observe that assuming perfectly-conditionally-correlated potential outcomes yields a lower bound very close to the upper bound. The bounds of Corollary 3.5 appear loose; indeed they are not tight.

6 CONCLUDING REMARKS

We study the average effect on those worst-affected by a proposed change as a measure of its *risk*, how to tightly bound it given covariates that explain some heterogeneity, and how to make robust inferences on these bounds even when this heterogeneity is roughly estimated. This provides very practical tools for assessing policy and product changes beyond their ATE and DTEs. We can safely use flexible yet biased/slow-to-converge machine learning, or we can avoid black-box models and easily get good bounds by considering only linear projections of heterogeneity . In the hypothetical case study this detected that, what appeared to be a positive/neutral change could actually very negatively impact a substantial subpopulation.

We focused on experimental (or, unconfounded observational) settings without interference, where risk is already unidentifiable *despite* randomization. A future direction is to consider the impact of interference [3, 29] or confounding Tan [55], where even ATEs are unidentifiable and fairness is harder to assess [30, 34, 39]. Interestingly, for partial identification under Tan [55]' model, X-conditional outcome-CVaR plays a crucial role [19]. Another direction may be to consider other risk measures, such as given by Kullback-Leibler ambiguity sets [1]. Per Footnote 5, the tight upper bound is still the risk measure applied to CATE, but it remains to compute lower bounds and design robust inference methods.

[14] In particular, the α-quantile DTE is uselessly *zero* for *all* $\alpha \in [0,1] \setminus \{1 - \mathbb{E}[Y(0)], 1 - \mathbb{E}[Y(1)]\}$ and the α-CVaR DTE is $\frac{1}{\alpha}(\mathbb{E}[Y(1)] - 1 + \alpha)_+ - \frac{1}{\alpha}(\mathbb{E}[Y(0)] - 1 + \alpha)_+$, which is not even monotonic. For illustration we plot it in Appendix B.

ACKNOWLEDGMENTS

I thank Netflix's Darío García García, Molly Jackman, Danielle Rosenberg, William Nelson, and Martin Tingley for very helpful conversations.

REFERENCES

[1] Amir Ahmadi-Javid. 2012. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications* 155, 3 (2012), 1105–1123.
[2] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1999. Coherent measures of risk. *Mathematical finance* 9, 3 (1999), 203–228.
[3] Susan Athey, Dean Eckles, and Guido W Imbens. 2018. Exact p-values for network interference. *J. Amer. Statist. Assoc.* 113, 521 (2018), 230–240.
[4] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.
[5] Susan Athey and Stefan Wager. 2017. Efficient policy learning. (2017).
[6] J Andrew Bagnell. 2005. Robust supervised learning. In *AAAI*.
[7] Luc Behaghel, Bruno Crépon, and Marc Gurgand. 2014. Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American economic journal: applied economics* 6, 4 (2014), 142–74.
[8] A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. 2017. Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica* 85, 1 (2017), 233–298.
[9] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59, 2 (2013), 341–357.
[10] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. 2018. Robust sample average approximation. *Mathematical Programming* 171, 1 (2018), 217–282.
[11] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAccT*.
[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21, 1 (2018), C1–C68.
[13] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. *Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.* Technical Report. National Bureau of Economic Research.
[14] Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. 2010. Quantile and probability curves without crossing. *Econometrica* 78, 3 (2010), 1093–1125.
[15] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90, 3 (2008), 389–405.
[16] AC Davison. 1992. Treatment effect heterogeneity in paired data. *Biometrika* 79, 3 (1992), 463–474.
[17] Peng Ding, Avi Feller, and Luke Miratrix. 2016. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2016), 655–671.
[18] Peng Ding, Avi Feller, and Luke Miratrix. 2019. Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* 114, 525 (2019), 304–317.
[19] Jacob Dorn, Kevin Guo, and Nathan Kallus. 2021. Doubly-Valid/Doubly-Sharp Sensitivity Analysis for Causal Inference with Unmeasured Confounding. (2021).
[20] Peyman Mohajerin Esfahani and Daniel Kuhn. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171, 1 (2018), 115–166.
[21] Sergio Firpo. 2007. Efficient semiparametric estimation of quantile treatment effects. *Econometrica* 75, 1 (2007), 259–276.
[22] M Gail and Richard Simon. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* (1985), 361–372.
[23] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. 2002. *A distribution-free theory of nonparametric regression.* Springer.
[24] Godfrey Harold Hardy, John Edensor Littlewood, George Pólya, and György Pólya. 1952. *Inequalities.* Cambridge university press.
[25] James J Heckman, Jeffrey Smith, and Nancy Clements. 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies* 64, 4 (1997), 487–535.
[26] Kosuke Imai and Marc Ratkovic. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7, 1 (2013), 443–470.
[27] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.
[28] Guido W Imbens and Jeffrey M Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of economic literature* 47, 1 (2009), 5–86.
[29] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. 2022. Experimental design in two-sided platforms: An analysis of bias. *Management Science* (2022).
[30] Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. 2020. Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 64–70.

[31] Nathan Kallus. 2018. Balanced policy evaluation and learning. In *NeurIPS*.
[32] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. 2019. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. (2019).
[33] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2021. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* (2021).
[34] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *ICML*.
[35] Nathan Kallus and Angela Zhou. 2019. Assessing disparate impacts of personalized interventions: Identifiability and bounds. *NeurIPS* (2019).
[36] Nathan Kallus and Angela Zhou. 2021. Minimax-optimal policy learning under unobserved confounding. *Management Science* 67, 5 (2021), 2870–2890.
[37] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*.
[38] Edward H Kennedy. 2020. Optimal doubly robust estimation of heterogeneous causal effects. (2020).
[39] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. 2020. The sensitivity of counterfactual fairness to unmeasured confounding. In *UAI*.
[40] Toru Kitagawa and Aleksey Tetenov. 2018. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica* 86, 2 (2018), 591–616.
[41] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. 2002. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk* 4 (2002), 43–68.
[42] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.
[43] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *NeurIPS*.
[44] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *AIES*.
[45] Xinkun Nie and Stefan Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 2 (2021), 299–319.
[46] Min Qian and Susan A Murphy. 2011. Performance guarantees for individualized treatment rules. *Annals of statistics* 39, 2 (2011), 1180.
[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *KDD*.
[48] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression-coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* 89, 427 (1994), 846–866.
[49] R Tyrrell Rockafellar and Stanislav Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of risk* 2 (2000), 21–42.
[50] Donald B Rubin. 1986. Comment: Which ifs have causal answers. *Journal of the American statistical association* 81, 396 (1986), 961–962.
[51] Andrzej Ruszczyński and Alexander Shapiro. 2006. Optimization of convex risk functions. *Mathematics of operations research* 31, 3 (2006), 433–452.
[52] Shlomo S Sawilowsky. 1990. Nonparametric tests of interaction in experimental design. *Review of Educational Research* 60, 1 (1990), 91–126.
[53] Anton Schick. 1986. On Asymptotically Efficient Estimation in Semiparametric Models. *Annals of Statistics* 14, 3 (09 1986), 1139–1151.
[54] Charles J Stone. 1982. Optimal global rates of convergence for nonparametric regression. *The annals of statistics* (1982), 1040–1053.
[55] Zhiqiang Tan. 2006. A Distributional Approach for Causal Inference Using Propensity Scores. *J. Amer. Statist. Assoc.* 101, 476 (2006), 1619–1637.
[56] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[57] Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. 2012. Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* 107, 499 (2012), 1106–1118.
[58] Wenjing Zheng and Mark J van der Laan. 2011. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*. Springer, 459–474.