MALTS - JMLR Review Response

Harsh Parikh

October 2020

1 To EICs

Please route this paper to Russ Greiner, who handled a previous version of this work and recommended that we submit it as a new manuscript. Thank you.

2 Response to Editor

Dear Russ and Reviewers,
We appreciate the invitation to resubmit our paper on MALTS to JMLR. As you recommended, we have updated the manuscript and responded to all comments. Please see our responses below in blue.

Editor's comment: "Thank you for your re-submission. While it does address some of the shortcomings of the earlier manuscript, Reviewer 2's detailed evaluation points to many significant limitations of this revised version – enough that acceptance is not an option. Given JMLR's policy, we must therefore reject this manuscript. If you authors can address these important issues, JMLR can certainly consider a new manuscript."

Based on the reviewers' comments we have worked on the following:

- Updated the Definitions and Theorems (along with proofs) to ensure consistency of matching on smooth distance metrics.
- Added further experiments discussing MALTS performance in comparison with other methods as the number of units increases, as the number of covariate increases, as the number of irrelevant covariates increases and as the overlap decreases.
- We have also worked on notations in the framework and the theory sections to avoid confusion.

We provide detailed in line-response to Reviewer 2's comments. Thank you again.

Regards,
Harsh Parikh, Cynthia Rudin and Alexander Volfovsky

3 Reviewer 1

Thank you for your detailed changes in response to the previous reviews.

4 Reviewer 2

4.1 Definition 1, Lemma 2 and Theorem 1

My last assessment hinged on my complaint that

> Definition 1 is satisfied only in trivial, unrealistic cases: it requires that units possessing the same covariates and treatment assignments can only be observed to have precisely the same outcomes.

Version 2 of the manuscript does modify the problematic definition. With this version's relaxed Definition 1, units sharing values of covariate and treatment variables are permitted to have different outcomes, but the probability of any such difference is required to be bounded away from 1. For distance metric d, Definition 1's upper limit on the probability of a nonzero difference in outcomes for units sharing values of x and the treatment variable is termed β_d; we may take $\beta_d \in [0, 1]$ for any metric d, but for d to be a "Smooth Distance Metric" it is now required that $\beta_d < 1$. (Version 1 had in effect required $\beta_d = 0$.) So far, so good. For anyone who've taken advanced calculus, " ϵ and δ evoke a game where ϵ is set to an arbitrary small positive quantity and δ is adjusted to a correspondingly small positive position. In Lemma 2, ϵ is an arbitrary nonnegative quantity, but the $\delta(\epsilon, M, n)$ that is paired to it does not generally tend to zero in tandem with epsilon; rather, inspection of the proof 1 shows this δ to be no smaller than Definition 1's β_d. That is, Lemma 2 limits the probability in its conclusion statement (p.26, bottom) not to be less than an arbitrarily small positive quantity, depending on the value selected for ϵ, but only to be less than 1, whatever the value of ϵ. Theorem 1 is affected in the same way: its conclusion of form $P(|\hat{\tau} - \tau| \geq \epsilon) \leq \delta$ in fact only asserts $P(|\hat{\tau} - \tau| \geq \epsilon) \leq \beta_d$. So it's entirely misleading to say, as inline summary statement following the statment of Theorem 1 says:

> "Theorem 1 follows from Lemma 2 in the appendix which proves that we can estimate counterfactual outcomes y correctly with high probability using nearest neighbor matching under a smooth distance metric."

The only thing that's been shown is that the probability in question is larger than $1 - \beta_d$, i.e. larger than 0.

Thank you so much for your comments. We have updated Definition 1 (definition of smooth distance metric) to make it more realistic and in line with the literature. The new definition for smooth distance metric states that $\mathbf{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a smooth distance metric if there exists a monotonically increasing bounded function $\delta_{\mathbf{d}}(\cdot)$ with zero intercept, such that $\forall z_i, z_j \in \mathcal{Z}$ if $t_i = t_j$ and $\mathbf{d}(\mathbf{x}_i, \mathbf{x}_j) \leq a$ then

$$|E[Y_i|X_i = \mathbf{x}_i, T = t_i] - E[Y_j|X_j = \mathbf{x}_j, T = t_j]| \leq \delta_{\mathbf{d}}(a).$$

Further, as per your comment the Lemma 2 and Theorem 1 are also updated. The updated version of Lemma 2 proves the following statement:

> "Let $\{\mathcal{S}_n\}_{n=1}^{\infty}$ be a sequence of nested datasets, each of which includes n iid samples from $\mu(\mathcal{Z})$, $n = 1..\infty$. For a fixed covariate vector $\mathbf{x}$, and fixed treatment indicator t', for any $\alpha > 0$ want to be able to choose a small enough value of "a" and a large enough value of N such that when $\alpha > \delta_{\mathbf{d}_{\mathcal{M}}}(a)$ there are enough units in $\mathcal{K}_n^{(t')}(\mathbf{x}) = \{z_k : d(\mathbf{x}_k, \mathbf{x}) < a, t_k = t', z_k \in \mathcal{S}_n\}$ for all $n \geq N$ such that
>
> $$P(|E[Y^{(t')}|\mathbf{x}] - \widehat{Y}_{\mathbf{x}}^{(t')}| > \alpha) \leq \exp(-|\mathcal{K}_n^{(t')}(\mathbf{x})|(\alpha - \delta_{\mathbf{d}_{\mathcal{M}}}(a))^2/2\mathbf{C}_y)$$
>
> where $\delta_{\mathbf{d}_{\mathcal{M}}}(a)$ is the bound from Definition 1 (definition of smooth distance metric). As the above choice of a holds for all $n \geq N$, we have that the bound goes to zero as $n \to \infty$".

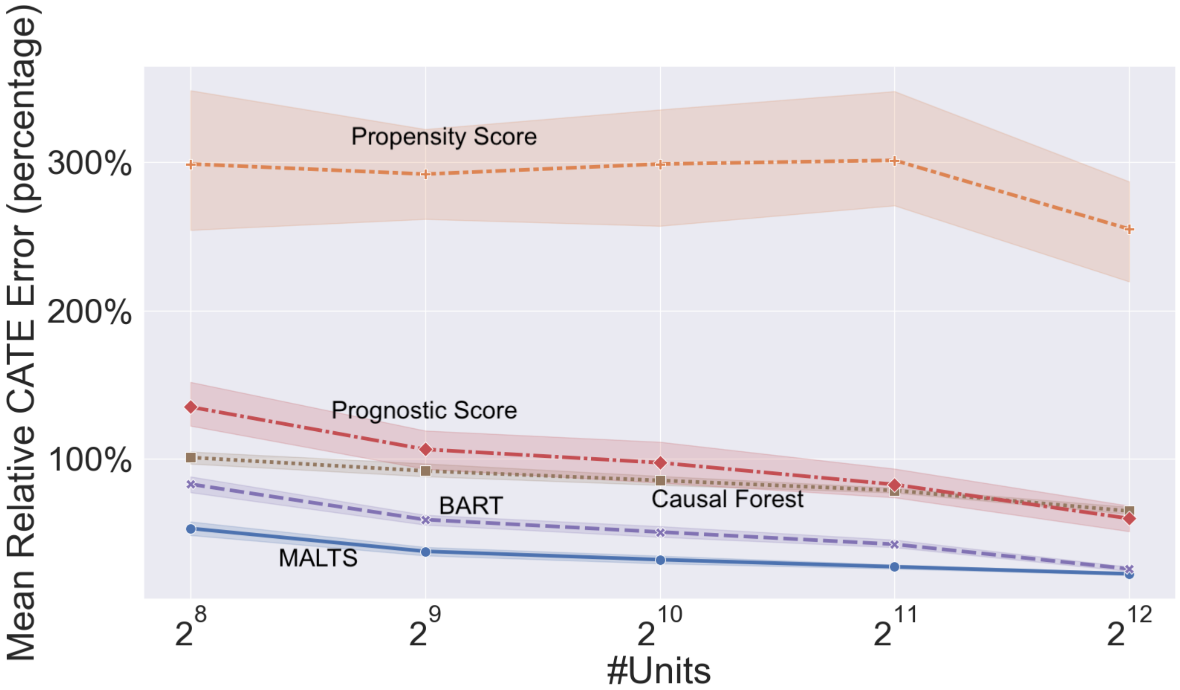


Figure 1: Comparative performance in estimating CATE using causal inference methods as the dataset size increases. The number of covariates is fixed throughout this simulation: $p = 20$. Each point in this plot is based on 50 simulations.

4.2 Simulations reported in "Experiments" section

The revision letter suggests that the paper's theoretical considerations are secondary to its main argument, "primarily motivational". Despite this being at odds with the large and prominent share of the paper devoted to these arguments, I did review the later sections of the paper this time. I agree that in the simulations and example presented in the paper, MALTS is competitive with a good selection of existing self-standing covariance adjustment and conditional treatment effect estimation routines. Additionally it offers advantages in terms of interpretability of pairings. Relative to other methodological papers built around simulation studies appearing in high-profile journals, however, this paper's simulation settings are limited in variety, considering only a few sample sizes n, a small range of ratios of p (the dimension of the covariate) to n, and essentially just one covariate overlap setting. This last limitation is of particular, somewhat subtle, significance, as I'll argue presently; for now I note that the paper's simulations would be rather thin for a JMLR paper to stand on.
Thank you. We have further added experiments studying the performance of different causal inference methods under regimes with different values of n, p and overlap. (1) Figure 1 compares MALTS performance with other methods as we increase number of units with $p = 20$, (2) Figure 2 compares MALTS performance with other methods while keeping $n = 2^{11}$ constant and changing p from 2 to 2^9 such that half of them are relevant, $k = p/2$, (3) Figure 3 compares MALTS performance with other methods while keeping $n = 2^{11}$, number of relevant covariates $k = 8$ and changing p from 8 to 256 and (4) finally, Figure 5 compares their performance under different overlap scenarios reported as (a) $\epsilon_{i,treat}$ (the level of noise in treatment assignment equation of the DGP) and (b) standard differences of means of covariates between the control and the treated groups. Figure 4 shows the relationship between $\epsilon_{i,treat}$ and standard differences of means of covariates between the control and the treated groups. We run each of the experiments 50 times to produce the above mentioned trend plots (Figure 1, Figure 2, Figure 5).

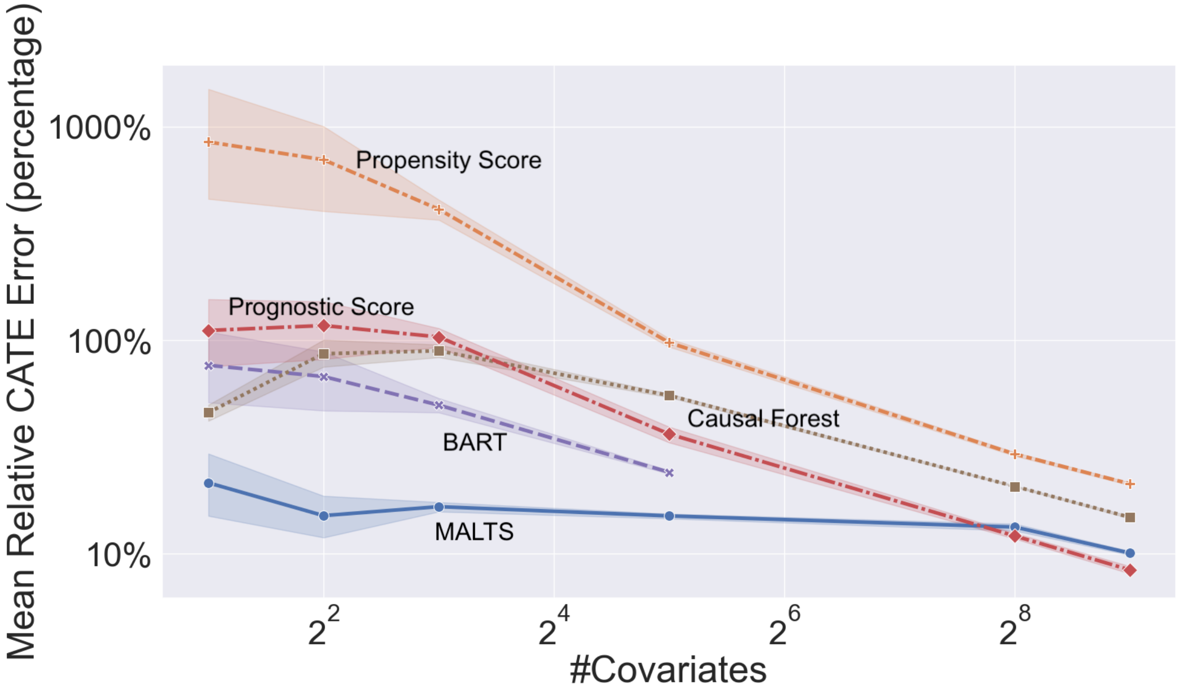


Figure 2: Comparative performance in estimating CATE using causal inference methods as the number of covariates to dataset size increases. The number of units is fixed: $n = 2^{11}$. (For the given n, BART doesn't return CATE estimate for all units when $p > 2^7$. Prognostic scores use BART for $p \leq 2^7$ and random forests for $p > 2^7$.)

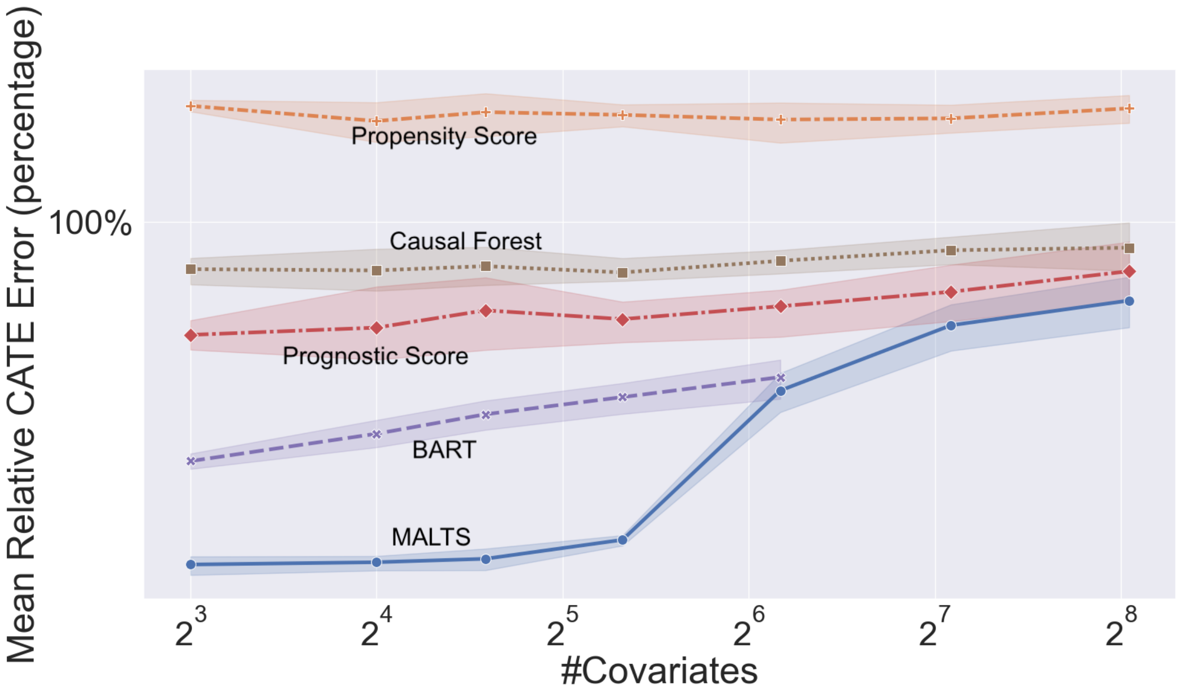


Figure 3: Comparative performance in estimating CATE using causal inference methods as the number of covariates increases keeping the number of relevant covariates constant and equal to 8. The number of units is fixed: $n = 2^{11}$. (For the given n, BART doesn't return CATE estimate for all units when $p \geq 2^7$. Prognostic scores use BART for $p \leq 2^7$ and random forests for $p \geq 2^7$.)

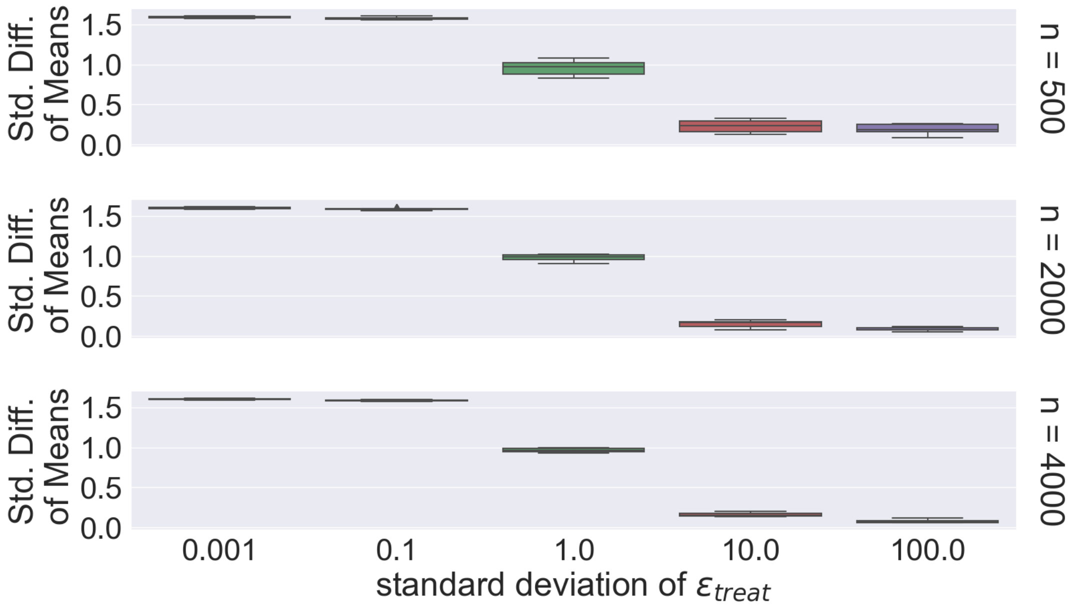
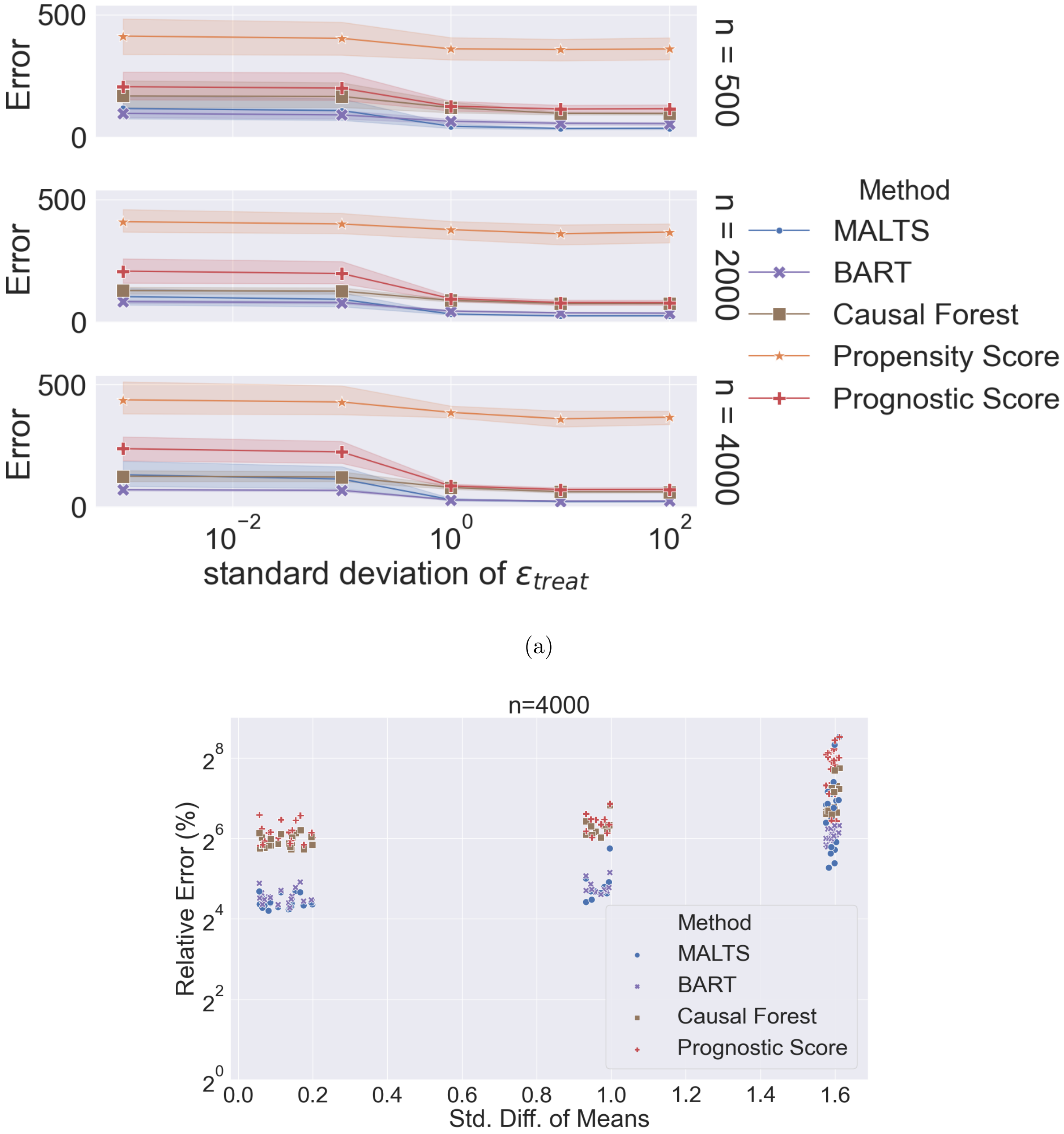


Figure 4: *Standardized difference of means between covariates of treated and control units decreases as* $\epsilon_{i,treat}$ *increases.* We increase the value of $\epsilon_{i,treat}$ in the DGP of treatment allocation which increases the overlap in the treated and control group. We generate data with p equal to 20 covariates and for values of $n \in \{500, 2000, 4000\}$.



(a)

(b)

Figure 5: (a) Trend plot comparison of MALTS performance measured as mean relative error for CATE estimation under different level of overlap measured as a function of $\epsilon_{i,treat}$ (the scale of noise in treatment allocation process). Higher values of the standard deviation of $\epsilon_{i,treat}$ correspond to more overlap between the control and the treated groups. (b) Scatter plot comparing MALTS' performance measured as mean relative error for CATE estimation under different level of overlap measured as standardized difference of means for different dataset sizes. Large value of standardized difference of means corresponds to less overlap between the control and treated groups.

4.3 Overlap/common support

This brings me to the matter of common support. The simulations discussed in the body of either version of the paper feature near-perfect overlap of treatment and control groups. (Treatment-control labels are generated according to the sign of the sum of an independent pair of covariates, each either $\mathcal{N}(0, 1.5)$ or a Rademacher variable scaled by 1/2, added together with an unobserved noise variable randomly drawn from $\mathcal{N}(0, 20)$. So the noise variation exceeds covariate variation by a factor of roughly 10.) At the behest of other reviewers, appendicial simulations with more limited overlap in the x-distributions of the treatment and control groups have been added. But the description of these simulations is so incomplete as to make it impossible to compare them to simulation settings in the main paper, much less any that might be reported elsewhere in the literature. The ps and ns of these supplemental simulations seem to have been 2 and 500, as opposed to the 25-40 and 2500 of the main simulations; we're not told how other settings may have differed. In particular, overlap is only described only in terms of a descriptive measure used to compare specific treatment and control samples; the method of stochastic allocation to treatment and control conditions is not given. This appendix compares MALTS to itself, under different variations of the (unexplained) selection mechanism, but to none of the competing methods considered in the main body of the paper. Plots of specific draws from that distribution are given, and based on these the selection mechanism of these simulations looks dramatically more severe that in the paper's main simulations – albeit modest in relation to selection mechanisms considered elsewhere in the matching literature.

Overlap may be a side-issue from this paper's perspective, given its goal of interpretable causal learning, but for extant matching literature it is anything but. For the originators of propensity score ideas in statistics and economics, it was the central concern, one calling for separate and distinct phase of analysis from any learning of $\mathbf{E}(Y^{(T)}|X)$ or $\mathbf{E}(Y^{(C)}|X)$; the purpose of propensity score matching is to play this preparatory, bias-reducing role (Rubin 1991; Rosenbaum 2002; Imbens and Rubin 2015). Propensity score matching is not be expected to compete well in the paper's main simulations, combining next to no selection with strongly prognostic covariates; that MALTS squarely beats it is no surprise.

We have now added an experiment in the appendix of the paper with 20 covariates with different levels of overlap reported both as difference of standardized means and the $\epsilon_{i,treat}$ values. Here, we compare the performance under limited overlap for different numbers of units, $n \in \{500, 2000, 4000\}$. The updated setup is comparable to experiments in the main text in the number of units and covariates. MALTS performs on par with BART and causal forest even as the common support decreases (see Figure 5). The primary reason for MALTS deterioration of performance under almost no-overlap compared to BART is because matching methods like MALTS can be conceptualized of as interpolation exercise unlike regression approaches, explicitly modeling the potential outcomes' surfaces, which are extrapolation exercises.

4.4 Additional comment for authors

The revision stops short of explicitly defining $y_i^{(T)}$, as I had suggested. I understand that you mean for it to be gleaned from the inline equation of $E(Y^{(T)} - Y^{(C)}|X = x_i)$ to $y^{(T)}(x_i) - y^{(C)}(x_i)$, with subsequent abbreviation of $y^{(T)}(x_i)$ and $y^{(C)}(x_i)$ as $y_i^{(T)}$ and y(C)i, on p.4; but then it's quite confusing that at the top of p.5 you say

> For units in the treatment set we know $y^{(T)}(X_i)$

which by your definition we of course do not. Were $y^{(T)}(x_i)$ to be read in the more ordinary sense of "subject i's potential outcome," as opposed to "population conditional mean potential outcome among subjects with covariates like i's covariates," then the statement would be quite correct; but this is precisely what your definition denies. The old Definition 1 would have (when assumed) entailed that a given subject's potential outcome and that conditional mean potential outcome to be the same, but ordinarily they're quite distinct, as the new Definition 1 rightly permits them to be. As a result, the inference from the displayed expression at bottom of p.4 to the one at the top of p.5 is not warranted; sample average loss is not representable in this manner.

In the updated version of the main text of the paper, we have corrected the mistakes in the notations. We use $E[Y_i^{(t')}|X_i = \mathbf{x}_i]$ to refer to conditional average potential outcome under treatment t' for covariate $X_i = \mathbf{x}_i$. Further, we denote the estimate of conditional average potential outcome for treatment t' and $X_i = \mathbf{x}_i$ as $\widehat{Y}_{\mathbf{x}_i}^{(t')}$. We further corrected the notation in the main text of the paper to avoid the confusion between the observed outcome y_i for unit $s_i = (\mathbf{x}_i, y_i, t_i)$ with treatment choice $t_i = t'$ and the estimate of conditional average potential outcome for treatment t' and $X_i = \mathbf{x}_i$ denoted as $\widehat{Y}_{\mathbf{x}_i}^{(t')}$.

- On p.7, you define $d_{\mathcal{M}_c} = |\mathcal{M}_c(a_c - b_c)|_2 = [(a_c - b_c)\mathcal{M}_c'\mathcal{M}_c(a_c - b_c)]^{1/2}$. On p.8 $d_{\mathcal{M}_c}$ seems to be the square of this quantity (or this quantity plus a squared Hamming distance).
  On p.8 we have corrected the definition of $d_{\mathcal{M}_c}$ in main text. Please see the update manuscript.

- In section 6, noise and covariates contribute to determination of treatment status with variance 2 and 20, resp. i.e. treatment assignment is almost totally random – an extremely friendly arrangement overlap-wise. The appendix discussion of overlap on pages 29 and 31 suggest that this 20 was reduced, but don't say how. Simulations discussed in this section differ in an incompletely specified manner from simulations discussed elsewhere, seeming to involve smaller samples and numbers of variables. The key σ parameter that determines the degree of overlap is not reported here, in contrast to other reported simulations. This evidence can't be counted as supporting the section's claim that "MALTS performs reasonably well under limited overlap."
  Appreciate the concern. We have added an experiment in the appendix of the paper with 20 covariates with different levels of overlap reported both as difference of standardized means and the $\epsilon_{i,treat}$ values. In this setup compares the performance for multiple values of number of units, $n \in \{500, 2000, 4000\}$. The update setup is comparable to experiments in the main text in the number of units and covariates. MALTS performs on par with BART and causal forest even as the common support decreases (see Figure 5).

- p.10 Definition 6 ends with a stipulation about δ_ϵ, but there's no δ_ϵ in the defining equation.
  In the updated version of the paper, we have cleaned Definition 6.

- p.10 Theorem 2 is a key claim, with Theorem 1 depending indirectly on it; but its statement is a mess. It makes reference to smooth distance metrics with "bounding function $\delta(\cdot)$," but there's no "$\delta(\cdot)$" in the current definition of "bounding function". It's unclear where to find explanations of key symbols: B appears not to be defined until the statement of subsequent Lemma 2; $\rho_\gamma^{(t')}$ is defined only within the proof, despite the symbol ρ having been used with a different meaning previously in Definition 2. The theorem statement suggests β is an arbitrary constant, but Definition 1 of smooth distance metrics also makes reference to β_d, and β's occurrence in the theorem statement adjacent to the assumption of a smooth distance metric suggests that β_d is what's intended.
  We will like to clarify that Theorem 2 doesn't inform anything for Theorem 1. I believe the reviewer intended to mean Lemma 2 instead. In the updated version of Definition 1, Definition 2, Theorem 1, Lemma 1 and Lemma 2, we have cleaned the notations to avoid confusion.

- p.11 In theorem 3, what do β and ϵ refer to?
  The statement of Theorem 3 has been updated in the main text to define and explain β and ϵ.