Response to the Official Review of Paper9135

Anonymous Author(s)

July 2022

We greatly appreciate the reviewers' positive feedback and valuable comments. In the revision, we have carefully addressed all the points the reviewers have raised and improved the paper accordingly, as detailed in the point-by-point response below. We first reply to the reviewers' concerns of our weakness that are itemized with W#, and then respond to their questions one by one that are itemized with Q#.

Response to Reviewer hbGq

W1. *Using observational data and historical controls, we can directly and accurately estimate potential control outcomes $\widehat{Y}^{(0)}$ with a large amount of historical control data. The conditional average treatment effect on the treated group can be easily obtained:* $\mathrm{CATT} = E(Y \mid D = 1, X) - E(\widehat{Y}^{(0)} \mid D = 1, X)$.

Thanks for your insightful comment. In the presence of unmeasured confounding, we in fact *cannot* accurately estimate potential control outcomes $\widehat{Y}^{(0)}$ with observed covariates and historical outcomes. We provide a detailed explanation of why we cannot in the response to your first question, which we hope will dispel your doubts.

W2. *In Lines 98-102, proper stratification [26] can help mitigate confounding effects from observed confounders but can't reduce the bias from unmeasured confounders that are independent with observed confounders.*

Thank you for pointing out this issue that may cause confusion, and we add the sentence "by exploiting historical controls that contain information of unmeasured confounding" in the revised version to avoid misleading. We agree with what you said, but it does not conflict with our method. Different from classical matching methods that only adjust for X, our method finds a partition of the covariate space (stratification) in which unmeasured confounders U can also be partially adjusted for because (1) control outcomes for treated units can be "observed" historically and thus historical confounding effects can be identified, and (2) confounding effects are assumed to change mildly with respect to time (Assumption 2 in our revised article) such that confounding information can be transferred from historical data to current data. In other words, the stratification depends not only on observed covariates X but also unmeasured confounders U through historical controls.

W3. *In practical, time-invariant covariates and unmeasured confounders assumption is hardly satisfied in a causal scenario.*

Thanks for your valuable comment. We make some revision and now we allow unmeasured confounders to be time-varying. The time-invariant assumption of covariates is standard in the literature of causal inference, and serves to simplify our formal definition and theoretical analysis of partial identification of CATT. In fact, this assumption is not as restrictive as it seems to be. In spite of measured confounders being time-invariant, their effects on outcomes can be time-varying; see Assumption 2 in our revised article. Besides, measured confounders that are time-invariant, or at least varying mildly with respect to time, are very common, e.g., sex, age, BMI, etc. When samples are collected over a small time span, i.e., $t_m - t_1$ being small, the time-invariant assumption of confounders can be viewed to hold. Relaxation of this assumption is possible by considering functional regressions (e.g., Ju and Salibián-Barrera 2021, arXiv:2109.02989).

W4. *In Lines 104-107, without any prior knowledge, neither constant function nor polynomial function guarantees robust results.*

Thanks for your comment. Piece-wise constant functions, or step functions, enjoy point-wise universal approximation ability to the space of measurable functions. In fact, trees, as an example of step functions, have shown a great success in various practical applications. Indeed, our partial identification bound takes the issue you raised into consideration; see Theorem 1. The second term of the upper bound, $\max_{1\le j\le q} \sup_{x\in Q_j^{**}} \left|\eta^*(x) - \eta^*(Q_j^{**})\right|$, measures the distance between our target function $\eta^*(x)$ and its step function approximator $\eta^*(Q_j^{**})$, which would be large if step functions are not suitable.

Surely, to obtain reasonable theoretical guarantees, certain smoothness condition, e.g., bounded variation or high-order differentiability, should be imposed on the target function $\eta^*(x)$, but this is not the key issue we aim to solve in this paper.

W5. *Checklist in Lines 492-493: The Assumption 1 is not the full set of assumptions. SUTVA, positive/overlap assumption are missing in this paper. The authors should reframe the assumptions required.*

Thanks for your helpful comment. As you suggest, we impose the SUTVA and consistency assumption to formalize the Rubin causal model; see Lines 90–91. We also impose the condition that for $d \in \{0, 1\}$, $P(D = d \mid X) > 0$, and $Y_{t_m}^{(d)} \perp\!\!\!\perp D \mid (X, U)$.

Notice that our positivity condition is weaker than the conventionally assumed version $P(D = d \mid X, U) > 0$ for $d = 0, 1$. This is indeed not surprising because we also require the transportability condition that there exists a partition of observable covariate space, and in each local region, confounding effects change mildly across time; thus, $P(D = d \mid X) > 0$ is enough for partial identification.

W6. *Figure 1 and Figure 2 in Appendix, what data was used to draw this picture? the authors did not give a clear explanation.*

Thanks for your valuable comments. The data used to draw the pictures in Appendix are generated in the same way as the simulation. We have added more detailed descriptions of figures in the revised version.

Figure 1 shows the effectiveness of a single tree of GBCT on reducing confounding bias. Let $\{\widehat{Q}_j^{\mathrm{GBCT}}\}_{j=1}^{q_1}$ be the partition of a single tree of GBCT, and let $\{\widehat{Q}_j^{\mathrm{GRF}}\}_{j=1}^{q_2}$ be that of GRF. Given our generated data, the y-axis denotes $\sum_{j=1}^{q_1} \left|\widehat{\mathbf{b}}_{t_m}(\widehat{Q}_j^{\mathrm{GBCT}})\right|$ and $\sum_{j=1}^{q_2} \left|\widehat{\mathbf{b}}_{t_m}(\widehat{Q}_j^{\mathrm{GRF}})\right|$, respectively, where

$$\widehat{\mathbf{b}}_{t_m}(Q_j) = \sum_{i=1}^{n} \frac{\mathbf{I}\{X_i \in Q_j, D_i = 1\} Y_{i,t_m}^{(0)}}{|\{i : X_i \in Q_j, D_i = 1\}|} - \frac{\mathbf{I}\{X_i \in Q_j, D_i = 0\} Y_{i,t_m}^{(0)}}{|\{i : X_i \in Q_j, D_i = 0\}|},$$

and $Y_{i,t_m}^{(0)}$ is available for $D_i = 1$ since it is a simulation. Here, we use an ℓ_1-type loss because it is the largest norm for vectors such that small confounding bias can be highlighted.

Figure 2 shows the ablation experiments to demonstrate the gain of the confounding entropy regularization (defined in Equation 10) and the trick of subtracting the empirical pre-treatment confounding bias. "GBCT-ND" means without confounding entropy regularization $\widehat{\mathrm{H}}(\boldsymbol{Q})$, "GBCT-NR" means that the pre-treatment confounding bias is not subtracted when predicting the effect, and "GBCT-NR-ND" means neither confounding entropy loss nor subtraction of the pre-treatment confounding bias. The y-axis means $\mathrm{MAE}_{\mathrm{CATE}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{\eta}(X_i) - \eta_i|$ defined in Line 293. In addition, the performance of "GBCT-ND" has been updated in Table 1 in our revised version.

W7. *Not Reproducible. The data generation mechanism, and estimators $\alpha(\cdot), \beta(\cdot), f(\cdot), g(\cdot)$ are not detailed in this paper.*

Thanks for pointing our missing details. We have provided the code of generating simulated data in supplementary material to facilitate replication, and the complete code for implementing our proposed method is in the process of approval and will be released as soon as possible. Estimators $\alpha(\cdot), \beta(\cdot), f(\cdot), g(\cdot)$ are specified in responding to your fourth question.

Q1. *Using observational data and historical controls, can we directly and accurately estimate potential control outcomes with a large amount of historical control data? Is the estimation of conditional average treatment effect on the treated group still a challenge?*

Thanks for your question. We *cannot* estimate CATT or potential control outcomes $Y_{t_2}^{(0)}$ by simply adjusting for X and historical controls $Y_{t_1}^{(0)}$ $(t_1 < t_2)$ *when unmeasured confounders directly affect both current potential outcomes and treatment assignment.* As an illustrative example, consider the following directed acyclic graph (DAG) in Fig. R.1 that encodes causal relationships of $(Y_{t_2}, Y_{t_1}, D, X, U)$, where U denotes some unmeasured confounders, X denotes observed covariates, D denotes the treatment, and Y_{t_1}, Y_{t_2} denotes historical and current outcomes, respectively. For the ideal case where $Y_{t_2}^{(0)}$ can be accurately estimated, without doubt, CATT can be estimated directly by plugging the estimated $Y_{t_2}^{(0)}$ in the definition (Line 97 in our revised article). However, since (X, Y_{t_1}) cannot block the backdoor of (D, Y_{t_2}), directly using (X, Y_{t_1}) to estimate or predict $Y_{t_2}^{(0)}$ *cannot* fully adjusted for unmeasured confounders U, which would lead to a biased estimate of $Y_{t_2}^{(0)}$ and further an erroneous CATT estimator; e.g., Simpson paradox would occur even when the sample size is large.

In deep contrast, our method can give partial identifiability of CATT; specifically, we quantify the distance between CATT and an estimable function $\eta(x)$ (defined in Equation 3) in Theorem 1, and the distance can

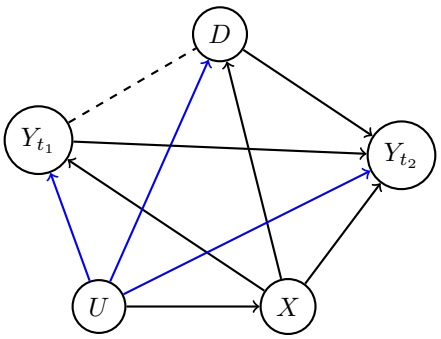

Figure R.1: A DAG for $(Y_{t_2}, Y_{t_1}, D, X, U)$. Dashed line --- denotes unspecified causal relationships.

be asymptotically negligible under some conditions. Different from classical matching methods that only adjust for X, our method finds a partition of the covariate space (stratification) in which unmeasured confounders U can also be partially adjusted for because (1) control outcomes for treated units can be 'observed' historically and thus historical confounding effects can be identified, and (2) confounding effects are assumed to change mildly with respect to time (Assumption 2 in our revised article) such that confounding information can be transferred from historical data to current data. These chronologically sampled data do facilitate our identification and estimation of CATT. Moreover, empirically, we compared several methods such as meta learners and causal forests which exploit historical controls as covariates to estimate CATT with our proposed method GBCT. GBCT yields smaller mean absolute errors than these methods, and the gap becomes more pronounced as impact of unmeasured confounding increases; see details at Lines 298–302 in our revised article.

Q2. *Lines 126-127, what is the relationship between Theorem 1 and the identification of instrumental variables?*

Thanks for pointing out the unclarity in our presentation. We meant to highlight that CATT is partially identifiable (identifiable up to a band) in our case (Theorem 1), and we mentioned instrumental variables as an example that ATE can be partially identified using IV-based methods.

Q3. *Why not use observed outcomes on the treated group when estimating CATT?*

We did use observed outcomes on the treated group to estimate CATT; see Lines 179–180 and Equation 12 in our revised article.

We are concerned that our presentation cannot adequately convey the rationale of our method which leads to your misunderstanding. Our regularized empirical risk minimization, i.e., Equation 12, consists of two parts. The first term measures the distance between our model and observed current outcomes, and the second term is estimated confounding entropy. The second term encourages to find a partition in which $\eta(x) \approx$ CATT as illustrated in replying your first question. For the first term, we use a tree $T(x; \boldsymbol{Q}, \widehat{\boldsymbol{\mu}}^{(1)}(\boldsymbol{Q}))$ to learn $\sum_{j=1}^{q} E(Y_{t_m}^{(1)} \mid D = 1, X \in Q_j^*)\mathbf{I}\{x \in Q_j^*\}$, where we use observed outcomes on the treated group, and another tree $T(x; \boldsymbol{Q}, \widehat{\boldsymbol{\mu}}^{(0)}(\boldsymbol{Q}))$ to learn $\sum_{j=1}^{q} E(Y_{t_m}^{(1)} \mid D = 1, X \in Q_j^*)\mathbf{I}\{x \in Q_j^*\}$. Thus, informally $\widehat{\eta}(x) \equiv T(x; \widehat{\boldsymbol{Q}}, \widehat{\boldsymbol{\mu}}^{(1)}(\widehat{\boldsymbol{Q}})) - T(x; \widehat{\boldsymbol{Q}}, \widehat{\boldsymbol{\mu}}^{(0)}(\widehat{\boldsymbol{Q}})) \approx \eta(x) \approx$ CATT, where $\widehat{\boldsymbol{Q}}$ denotes the solution of Equation 12.

Q4. *The data generation mechanism, and estimators $\alpha(\cdot), \beta(\cdot), f(\cdot), g(\cdot)$ are not detailed in this paper.*

Thanks for pointing our missing details due to limited length of paper. We provide codes of generating data in supplementary materials. Estimators $\alpha(\cdot), \beta(\cdot), f(\cdot), g(\cdot)$ are specified as follows.

The interception function $\alpha(\cdot)$ is

$$\alpha(X, W) = \frac{\alpha_1(X \oplus W) + \alpha_2(X \oplus W) - \mu_\alpha}{\sigma_\alpha}\tilde{\sigma}_\alpha + \tilde{\mu}_\alpha,$$

where

$$\alpha_1(Z) = \frac{1}{p_z - 4}\sum_{i=0}^{p_z-4} 10\sin(\pi Z_i Z_{i+1}) + 20(Z_{i+2} - 0.5)^2 + 10Z_{i+3} + 5Z_{i+4},$$

$$\alpha_2(Z) = \frac{1}{p_z - 4}\sum_{i=0}^{p_z-4} \sqrt{Z_i^2 + \left[Z_{i+1}Z_{i+2} - 1/(Z_{i+1}Z_{i+3})\right]^2}, \quad Z \in \mathbb{R}^{p_z},$$

$\mu_\alpha = E\big[\alpha_1(X \oplus W) + \alpha_2(X \oplus W)\big]$ and $\sigma_\alpha^2 = \mathrm{var}\big[\alpha_1(X \oplus W) + \alpha_2(X \oplus W)\big]$, $X \oplus W = (X^T, W^T)^T$ represents concatenating two vectors, and $\tilde{\mu}_\alpha = 2$ and $\tilde{\sigma}_\alpha = 2$ are the pre-defined mean and standard deviation.

The function $\beta(\cdot)$ is

$$\beta(X, W) = \theta_\beta^T(X \oplus W) - E\big[\theta_\beta^T(X \oplus W)\big] + \tilde{\mu}_\beta, \quad \theta_\beta \sim \mathcal{N}_{2p}(0, I_{2p}),$$

where $\tilde{\mu}_\beta = 2$.

The time-varying related function $f(\cdot)$ is

$$f(W, \lambda_t) = (\Theta_f^T W)^T \lambda_t,$$

where Θ_f is a $p \times 3$ matrix and $\lambda_t \in \mathbb{R}^3$ denotes a time-varying factor. In addition, entries of Θ_f are independently sampled from a normal distribution $\mathcal{N}(0, 1)$.

The effect function $g(\cdot)$ is

$$g(X) = \frac{\alpha_1(X)/10}{\log(|\alpha_2(X)| + 1)}.$$

Response to Reviewer J9bc

W1. *It would be great if the context can be connected with former works closely. Readers may be very familiar with Athey's works on regression trees and causal forests, but the content in Section 2 is too dense.*

Thanks for your valuable suggestions. We agree that the comparisons between our method and former works on causal forests are important, and we compare our work with causal forests in detail in responding your first question. We emphasize these points in Lines 189–190 and 194–201 of our revised version. However, due to limited paper length, we cannot add all the comparisons in the main text. We decide to include our response into supplementary materials for interesting readers.

W2. *I don't think the improvement is as fair and significant as the paper states. I think the improvement should lie in comparing with the tree-based or forest-based causal ML methods.*

Thanks for your valuable suggestions. As you suggest, we add more simulation experiments to evaluate gains of regularizing with confounding entropy; see revised Table 1 in our article and our response to your fourth and sixth question. Also, conceptual differences between our method and forest-based causal models are illustrated in responding to your first question.

Q1. *What are the differences between your work and "Recursive partitioning for heterogeneous causal effects" "Estimation and inference of heterogeneous treatment effects using random forests" "Generalized random forests"?*

Thanks for your question. In the following, we use AI16, WA18, and ATW19 to denote papers you mentioned in the same order (Authors' initials + year of publication), and firstly give a brief review. Notably, the tree methods are built on the condition that there is no unmeasured confounding. From a methodological perspective, AI16, WA18, and ATW19 are closely related. WA18 developed causal forests which generates an ensemble of causal trees proposed in AI16; generalized random forests proposed in ATW19 fit any quantity of interest identified as the solution to a set of local moment equations, which is almost equivalent to causal forests in the case of estimating CATE. In addition to unconfoundedness condition, all of the three methods hinge on the so-called "honest splitting technique" that when we grow a specific tree, one data point is exclusively used either for splitting or estimation. Thus, they can only eliminate spurious correlations caused by using the same set of samples to simultaneously select model structure (partition of the covariate space) and estimate parameters based on this selected model structure.

The differences between our work and the three you mentioned are three-fold. The most fundamental and significant difference is that our method can deal with unmeasured confounding but they cannot. In the presence of unmeasured confounding, their methods would lead to an erroneous CATT/CATE estimator even when the sample size is large, e.g., Simpson paradox. In contrast, our method gives partial identifiability of CATT; specifically, we quantify the distance between CATT and an identifiable function $\eta(x)$ (defined in Equation 3) in Theorem 1, and the distance can be asymptotically negligible under some conditions. Moreover, even in the absence of unmeasured confounding where assumptions of AI16, WA18 and ATW19 are satisfied, our method is more robust to imbalance of treatments; see Section 3 for details. This property enables our method to exploit historical controls under a weaker positivity condition on treatment assignment. As pointed by Reviewer hbGq, one can use historical controls and X

to estimate current potential controls $Y_{t_m}^{(0)}$, which, in the context of responding to your question, can be done using causal forests. Notably, however, for identifiability of $E\big(Y_{t_m}^{(0)} \mid D=1, X, Y_{t_k}, 1 \le k \le m-1\big)$, a typical condition is $P(D=d \mid X, Y_{t_k}, 1 \le k \le m-1) > 0$, which is stronger than our assumption $P(D=d \mid X) > 0$, especially when the number of historical timestamps, i.e., $m-1$, is large. Lastly, WA18 and ATW19 combine multiple causal trees to generate a random forest, aiming at variance reduction, but we choose the boosting framework to improve performance of a single tree. This different preference for ensemble techniques results from a deep methodological distinction. Methods proposed in WA18 and ATW19 need to reduce variances of trees because trees they construct are very deep, i.e., each leaf contains a very small amount of samples. However, our proposed debiased causal trees cannot be very deep to assure confounding entropy to be accurately estimated, and extra bias would be introduced owing to regularization; therefore, we exploit gradient boosting to further reduce bias. See detailed discussion at Lines 195–203 in our revised paper.

In fact, we compared our method GBCT and generalized random forests (GRF) empirically; see Table 1 in our paper. In both cases (absence/presence of unmeasured confounding), GBCT outperforms GRF with 95% confidence level, which supports our analysis.

Q2. *Why is confounding entropy valid? Is it like the individual version of the integral probability metric? Is there any connection?*

Thanks for your insightful question. The rationale behind our method is that (1) control outcomes for treated units are observable historically and thus historical confounding effects can be identified, and (2) confounding effects are assumed to change mildly with respect to time (Assumption 2 in our revised article) such that confounding information can be transferred from historical data to current data.

Conventionally, effects introduced by unmeasured confounding at time t are characterized by $\mathbf{b}_t = E(Y_t^{(0)} \mid D=1, X) - E(Y_t^{(0)} \mid D=0, X)$; see Equation 1 for details. Noting that $\mathbf{b}_t$ only accounts for differences at the first moment, in our work, we exploit symmetric cross entropy H_t to model differences between $P(Y_t^{(0)} \mid D=1, X)$ and $P(Y_t^{(0)} \mid D=0, X)$, and we expect that H_t can adjust for high-order variations introduced by unmeasured confounders that $\mathbf{b}_t$ cannot. To adapt this idea to tree-based models, we obtain averaged $\widehat{\mathrm{H}}_t(\boldsymbol{Q})$ across $t \in \{t_1, \ldots, t_{m-1}\}$ as a function of partition $\boldsymbol{Q} = \{Q_j\}_{j=1}^q$ in Equation 10, which is used as a regularization term in Equation 12. It adjusts for unmeasured confounding by encouraging to find a partition $\widehat{\boldsymbol{Q}} = \{\widehat{Q}_j\}_{j=1}^q$ such that $P(Y_t^{(0)} \mid D=1, X \in \widehat{Q}_j) \approx P(Y_t^{(0)} \mid D=0, X \in \widehat{Q}_j)$.

We remark that symmetric cross-entropy is *not statistically necessary* and in principle we can use other alternatives such as integral probability metric (IPM) and Csiszár's f-divergence to measure the discrepancy between distributions. IPM with respect to a function class $\mathcal{F}$ is defined as $d_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \big|E_P[f(X)] - E_Q[f(X)]\big|$, and Csiszár's f-divergence is defined as $\mathrm{div}_f(P,Q) = E_Q[f((\mathrm{d}P/\mathrm{d}Q)(X))]$. By definition, symmetric cross entropy is known to be closely related to Kullback–Leibler (KL) divergence that is an f-divergence with $f(t) = t\log t$; more specifically, the symmetric cross entropy of P, Q with density function being p, q is $\mathrm{H}(P,Q) = E_P[\log p(X)] + E_Q[\log q(X)] + E_P[(q(X)/p(X))\log(q(X)/p(X))] + E_Q[(p(X)/q(X))\log(p(X)/q(X))] = \mathrm{Entropy}(P) + \mathrm{Entropy}(Q) + \mathrm{KL}(Q,P) + \mathrm{KL}(P,Q)$. Therefore, our proposed confounding entropy can be viewed as a symmetric version of KL-divergence between localized distributions, i.e., $Y_t^{(0)} \mid D=d, X \in Q_j, d=0,1$. Moreover, this reformulation of symmetric cross entropy raises an intriguing interpretation of our proposed confounding entropy: regularizing with confounding entropy can not only find a partition that reduce differences between the conditional outcome distribution of the treated and control groups, but also control their model complexity and thus avoid overfitting.

Confounding entropy does not belong to a family of IPMs. In fact, IPMs and f-divergences are intrinsically different; the family of f-divergences and the family of IPMs intersect only at the total variation distance (see, Sriperumbudur et al., Electron. J. Statist. 6: 1550-1599 (2012)). Surely, extension of confounding entropy to a more general f-divergences and IPMs is possible; however, several issues exist in choosing function class when using IPMs, and in choosing $f(\cdot)$ when using Csiszár's f-divergence, e.g., overfitting due to too complex function class which may take noises as signals, misspecification raised by an unsuitable $f(\cdot)$, and potentially expensive time cost for computing these alternatives. These issues can be left for future work.

Q3. *Why do you need to compare GBCT with meta learners? They are just ways to estimate treatment effects, but the base models can be chosen as linear/tree/neural nets. Similarly, why DML-RF? DML is a doubly robust estimator, and the base models for nuisance parameters can be chosen as linear/tree/neural nets, and you choose RF. In my view, you only need to compare all the regression trees and random forests models that are related to causal inference and analyze the source of gain produced by confounding entropy.*

Thank you for this insightful comment and suggestion. We agree that it is essential to assess the gain of our proposed confounding entropy and we have added a detailed discussion on this issue, see the answer

for question 6 below.

When implementing the conventional methods, we include historical outcomes as covariates, just as suggested by the first reviewer for comparison. Therefore, our comparison with the conventional methods is intended to evaluate the performances of two different strategies for utilizing historical controls, one through our proposed confounding entropy and the other by directly including historical outcomes as covariates. The results confirm that our proposed strategy for integrating the information of historical controls can give a much better performance. We have made this point more clear in our new version, see Lines 258–260, and thank you for pointing out this issue that may cause confusion.

Q4. *In Section 4.2, I don't find any source of data. Where is the real-world dataset and code to reproduce your experiments? I hope code/datasets can be released in the rebuttal stage.*

The real-world dataset involves users' personal and transactions information, and is not publicly available. We have provided the code of generating simulated data in supplementary material to facilitate replication, and the complete code for implementing our proposed method is in the process of approval and will be released as soon as possible.

Q5. *I am confused about the Biased dataset in Table 3. Is the control group increasing the credit line by 0? Why increasing the credit line for the low-risk group can bring bias into both low-risk and medium-risk groups? How does the RCT experiments conduct? Is it randomly assigning 0/2k/3k/6k to some people, where 0 is control and 2k/3k/6k is treat?*

Thank you for pointing out these issues that may cause confusion. The reall-world dataset contains 3 million observations from an RCT, which randomly assigns four treatment arms (increasing credit line by 0, 2000, 3000 or 6000) to users within the low-risk and medium-risk strata, respectively. We take the units with a credit line increase of 0 as the control group, and a credit line increase of 2000, 3000 or 6000 as three different treated groups.

In real-world financial scenarios, however, the financial companies always favor low-risk users, and are unwilling to increase credit line for those with relatively high risks. Therefore, we artificially constructed a biased observational dataset by retaining only the control group of the medium-risk stratum and the treated groups of the low-risk stratum. In this case, the risk status affects both the treatment assignment and the outcome, and thus serves as a substantial confounding. In data analysis, the risk status of each individual is blinded for investigators, and we implement our proposed method by including the observed 87-dimensional covariates as well as outcome values over the last eight months.

Q6. *In my view, the most important part is to analyze the gain of confounding entropy, which is missing in other tree/forest-based causal ML models. You are not developing a new triply robust estimator or a novel meta learner, so I think there is no need to compare with doubly robust estimators or other meta-learners. So, I think 4-5 well-known tree models for causalML are enough as baselines.*

We very much appreciate your helpful comments and suggestions. We agree that it is essential to assess the impact of our proposed confounding entropy more in-depth. In the new version, we made a concerted effort to demonstrate the the gain of confounding entropy by adding discussions and simulations, see Lines 298–302, Lines 316–317 and Table 1. For convenience, we also present the numerical results in Table R.1. The results confirm that our proposed confounding entropy leads to more substantially accurate and stable estimates across different scenarios.

Response to Reviewer ZPaK

W1. *I am somewhat concerned about the transition from Equation 8 to Equation 9. As written, the assertion in line 159 about the log density of $P_{t,1,j}$ being $-(y_t-\theta_j^{*(d)})^2$ is true only when y_t is normally distributed with unit variance. You seem to implicitly assume that no matter the partition Q_j or value of D, the outcomes are normally distributed with known and fixed variance. I find the fixed & known variance assumption somewhat unrealistic.*

Thanks for pointing out issues in our paper that may cause misunderstanding. We do *not* require that log-density of $Y_t^{(0)} \mid D=1, X\in Q_j$ and $Y_t^{(0)} \mid D=0, X\in Q_j$ are exactly normally distributed, but use Gaussian distribution to approximate them. Specifying $P_{t,d,j}$ as a normal distribution $\mathcal{N}(\theta_{t,j}^{*(d)}, \sigma_{t,j}^{*(d)})$ is to only consider the first and second moment of $Y_t^{(0)} \mid D=d, X\in Q_j$. In this case, the symmetric cross

entropy $\widehat{\mathrm{H}}_{j,t}$ between $\mathcal{N}(\theta^{*(0)}_{t,j}, \sigma^{*(0)}_{t,j})$ and $\mathcal{N}(\theta^{*(1)}_{t,j}, \sigma^{*(1)}_{t,j})$ can be estimated by

$$\begin{aligned}
\widehat{\mathrm{H}}_{j,t} &= -\sum_{i:X_i\in Q_j}\left\{\frac{D_i\log(p(Y_{i,t};\widehat{\theta}^{(0)}_{t,j},\widehat{\sigma}_0))}{|\{i:X_i\in Q_j, D_i=1\}|}+\frac{(1-D_i)\log(p(Y_{i,t};\widehat{\theta}^{(1)}_{t,j},\widehat{\sigma}_1))}{|\{i:X_i\in Q_j, D_i=0\}|}\right\}\\
&= -\log(2\pi)+\log(\widehat{\sigma}_1)+\log(\widehat{\sigma}_0)+\sum_{i:X_i\in Q_j}\left\{\frac{D_i(Y_{i,t}-\widehat{\theta}^{(0)}_{t,j})^2\widehat{\sigma}_0^{-2}}{|\{i:X_i\in Q_j, D_i=1\}|}+\frac{(1-D_i)(Y_{i,t}-\widehat{\theta}^{(1)}_{t,j})^2\widehat{\sigma}_1^{-2}}{|\{i:X_i\in Q_j, D_i=0\}|}\right\}\\
&= \underbrace{-\log(2\pi)+\log(\widehat{\sigma}_1)+\log(\widehat{\sigma}_0)}_{\text{entropy}}+\underbrace{\frac{\widehat{\sigma}_1^2+(\widehat{\theta}^{(0)}_{t,j}-\widehat{\theta}^{(1)}_{t,j})^2}{\widehat{\sigma}_0^2}+\frac{\widehat{\sigma}_0^2+(\widehat{\theta}^{(0)}_{t,j}-\widehat{\theta}^{(1)}_{t,j})^2}{\widehat{\sigma}_1^2}}_{\text{KL-divergence}},
\end{aligned}$$

where $\widehat{\sigma}_1, \widehat{\sigma}_0$ denote sample variances and $\widehat{\theta}^{(1)}_{t,j}, \widehat{\theta}^{(0)}_{t,j}$ denote sample mean of $Y^{(0)}_t \mid D=1, X\in Q_j$ and $Y^{(0)}_t \mid D=0, X\in Q_j$, respectively. Noting that variances $\widehat{\sigma}_1, \widehat{\sigma}_0$ are denominators, $\widehat{\mathrm{H}}_{j,t}$ is very sensitive to magnitude of $\widehat{\sigma}_1$ and $\widehat{\sigma}_0$. More specifically, $\widehat{\theta}^{(0)}_{t,j}-\widehat{\theta}^{(1)}_{t,j}$ is unlikely to be penalized towards zero when $\widehat{\sigma}_1$ and $\widehat{\sigma}_0$ are large, which may give rise to a Q_j where $Y^{(0)}_t \mid D=d, X\in Q_j$ $(d=0,1)$ are very noisy (likely to happen when sample size in Q_j is small) and $\widehat{\theta}^{(0)}_{t,j}-\widehat{\theta}^{(1)}_{t,j}$ is relatively large. In order to emphasize the role played by $\widehat{\theta}^{(0)}_{t,j}-\widehat{\theta}^{(1)}_{t,j}$ in finding the partition, we simply set the denominators as a constant, which coincidentally equals to the empirical symmetric cross entropy between $\mathcal{N}(\theta^{*(0)}_{t,j}, 1)$ and $\mathcal{N}(\theta^{*(1)}_{t,j}, 1)$.

Therefore, a fixed and known variance assumption is not imposed but serves as an approximation of true distributions of $Y^{(0)}_t \mid D=d, X\in Q_j$ $d=0,1$. Empirically, this approximation works well; see Table 1 in our revised article for detailed simulation results.

W2. *Additionally, since $\theta^{*(d)}_{t,j}$ is typically unknown, you plug in the sample mean in Equation 8. This may be reasonable when Q_j contains a large number of treated and control observations. However, it seems much less reasonable when there are not many of either. As such the expression in Equation 9 may not be a particularly accurate estimate of the true symmetric cross-entropy between $P_{t,1,j}$ and $P_{t,0,j}$. I'd go further and say that the expression in Equation 9 represents the symmetric cross-entropy between $P_{t,1,j}$ and $P_{t,0,j}$ under extremely restrictive conditions, namely that the $y_{t,j}$'s are centered around the corresponding sample mean with variance 1 for all choices of Q_j.*

Thanks for your comments. Our explanation for the issue that "$y_{t,j}$'s are centered around the corresponding sample mean with variance 1 for all choices of Q_j" can be found in the response to your first concern. We agree with you that when sample size in Q_j is small, plug-in estimation of the symmetric cross entropy between $P_{t,1,j}$ and $P_{t,0,j}$ can be problematic. However, applying some techniques such as pruning, our single debiased causal tree is not very deep such that sample size in each Q_j is not very small, and thus accuracy of estimating confounding entropy is acceptable. The price to pay of maintaining a large sample size in Q_j is the risk of underfitting which can introduce a non-vanishing bias. To eliminate the underfitting bias, we further combine multiple debiased causal trees through gradient boosting.

W3. *I found the motivation for regularizing based on historical cross-entropy loss somewhat lacking. The identification result suggests that one should strive to make $\sum_{k=1}^{m-1}\sum_{j=1}^{q}\mathbf{b}^2_{t_k}(Q^{**}_j)$ as small as possible. You write in lines 150-151 that minimizing this expression can "give rise to partitions that yield estimators with large variance." It is not at all clear that your suggested regularizer avoids this pathology.*

Thanks for your comments. This comment is closely related to your first question, and responses can be found therein.

Q1. *Can you prove hat regularizing with $\widehat{\mathrm{H}}$ instead of the sum of squared biases actually results in estimators with smaller variance? If not, you should provide, at a minimum, more justification for not using the sum of squared biases (i.e. the right-hand side of Equation 4) as the regularizer in Equation 12. Better would be an empirical demonstration that $\widehat{\mathrm{H}}$ yielded better point estimates or better uncertainty quantification than using the sum of squared biases.*

Thank you for your insightful comment and suggestion. Equation 9 of our article says that $\widehat{H}_t(Q_j)$ can be decomposed into a sum of squared biases and sample variances of local estimators, which suggests that, compared to only squared biases, regularizing with $\widehat{H}_t(Q_j)$ leads to estimators with smaller variances. In general, as we reply to your second question, symmetric cross entropy of P, Q with density function being p, q can be reformulated as $\mathrm{H}(P,Q)=\mathrm{Entropy}(P)+\mathrm{Entropy}(Q)+\mathrm{KL}(Q,P)+\mathrm{KL}(P,Q)$; therefore, regularizing with confounding entropy does have the ability to control model complexity.

As you suggest, we demonstrate empirically the gain of regularizing with $\widehat{\mathrm{H}}$ against the sum of squared biases in Table 1 of the new vision; we also present the results here, see Table R.1. Our experimental results confirm the superiority of regularizing with confounding entropy over squared biases in terms of estimation error, and the superiority is more significant as the level of unmeasured confounding and imbalance of treatments (in the absence of unmeasured confounding) increases, which supports our analysis.

Table R.1: MAE (mean±s.d.) of the ablation experiments. The variant `GBCT-ND` does not use the confounding entropy $\widehat{\mathrm{H}}(\boldsymbol{Q})$, and the other variant `GBCT-B` employs solely the first term $\widehat{\mathbf{b}}_t^2(Q_j)$ as $\widehat{\mathrm{H}}_t(Q_j)$ to construct the de-bias loss. Scenario I represents the absence of unmeasured confounding, whereas II represents the presence of unmeasured confounding. In addition, $\bullet/\circ$ indicates whether `GBCT` is statistically superior/inferior to its variants (pairwise t-test at 0.05 significance level).

	ϕ	GBCT	GBCT-ND	GBCT-B
I	0.2	1.05 ± 0.01	$1.10 \pm 0.01\bullet$	1.06 ± 0.01
	0.5	1.25 ± 0.03	$1.30 \pm 0.02\bullet$	$1.34 \pm 0.04\bullet$
	0.8	1.31 ± 0.02	$1.41 \pm 0.04\bullet$	$1.45 \pm 0.04\bullet$
	0.9	1.28 ± 0.03	$1.44 \pm 0.06\bullet$	$1.43 \pm 0.10\bullet$
	1	1.71 ± 0.06	$3.89 \pm 0.57\bullet$	$2.85 \pm 0.34\bullet$
II	0.2	1.07 ± 0.01	$1.11 \pm 0.01\bullet$	1.08 ± 0.02
	0.5	1.25 ± 0.02	$1.29 \pm 0.02\bullet$	$1.33 \pm 0.04\bullet$
	0.8	1.29 ± 0.04	$1.34 \pm 0.04\bullet$	$1.52 \pm 0.06\bullet$
	0.9	1.38 ± 0.09	$1.92 \pm 0.04\bullet$	$1.45 \pm 0.08\bullet$
	1	1.81 ± 0.10	$3.19 \pm 0.75\bullet$	$2.65 \pm 0.76\bullet$

Q2. *Relatedly, why did you try to minimize the symmetric cross-entropy between $P_{t,1,j}$ and $P_{t,0,j}$? There are several other choices of distributional differences/distances/divergences. Does closeness in this distance imply closeness in moments?*

Thanks for your question. Conventionally, effects introduced by unmeasured confounding at time t are characterized by $\mathbf{b}_t = E(Y_t^{(0)} \mid D = 1, X) - E(Y_t^{(0)} \mid D = 0, X)$, but it only accounts for differences at the first moment. We propose a general framework to model differences between the conditional outcome distribution of the treated and control groups, aiming to adjust for high-order variations introduced by unmeasured confounders that $\mathbf{b}_t$ cannot. Based on this motivation, symmetric cross entropy serves as a well-tested choice, but is *not statistically necessary*.

As pointed by you and reviewer J9bc, in principle we can use other alternatives such as integral probability metric (IPM) and Csiszár's f-divergence to measure the discrepancy between distributions. We highlight that symmetric cross entropy enjoys several advantages. In fact, symmetric cross entropy is closely related to Kullback–Leibler (KL) divergence that is an f-divergence with $f(t) = t \log t$; more specifically, the symmetric cross entropy of P, Q with density function being p, q is

$$\begin{aligned}
\mathrm{H}(P, Q) &= -E_P[\log q(X)] - E_Q[\log p(X)] \\
&= -E_P[\log p(X)] - E_Q[\log q(X)] + E_P[\{q(X)/p(X)\} \log\{q(X)/p(X)\}] \\
&\qquad + E_Q[\{p(X)/q(X)\} \log\{p(X)/q(X)\}] \\
&= \mathrm{Entropy}(P) + \mathrm{Entropy}(Q) + \mathrm{KL}(Q, P) + \mathrm{KL}(P, Q).
\end{aligned}$$

Noting that Shannon's entropy measures variability of random variables, this reformulation of symmetric cross entropy raises an intriguing interpretation of our proposed confounding entropy: regularizing with confounding entropy can *not only* reduce differences between the conditional outcome distribution of the treated and control groups, *but also* control their model complexity and thus avoid overfitting. Also, this observation suggests that closeness in symmetric cross entropy does not imply closeness in moments. Indeed, the maximum mean discrepancy (MMD) $d_K(\cdot, \cdot)$ with kernel function being $K(y_1, y_2) = \phi(y_1)^T \phi(y_2)$ $\phi(y) = (y, y^2)^T, y \in \mathbb{R}$ measures the closeness in moments since $d_K^2(P, Q) = (E_P[X] - E_Q[X])^2 + (E_P[X^2] - E_Q[X^2])^2$ for any P and Q.

Another reason why we choose symmetric cross-entropy, is that we want to accommodate more types of data in a unified notion. For example, when observed outcomes are continuous random variables, we can approximate $P_{t,d,j}$ with Gaussian distributions, and for binary outcomes, we can approximate $P_{t,d,j}$ with Bernoulli distributions. Surely, extension of confounding entropy to a more general class of metrics is possible; however, several issues exist in choosing function class when using IPM and $f(\cdot)$ when

using Csiszár's f-divergence, e.g., overfitting due to too complex function class which may take noises as signals, misspecification raised by an unsuitable $f(\cdot)$, and potentially expensive time cost for computing these alternatives. These issues can be left for future work.

Q3. *To what extent does your framework allow treatment assignment to depend on the potential outcomes in the pre-treatment period? One can readily imagine scenarios where the decision to treat is based on past history. It would be useful to clarify explicitly whether you need to assume treatment is assigned independently of past outcomes.*

Thanks for you insightful question. In principle, our method allows dependency between historical outcomes and treatment assignment, but to highlight our main idea, we implicitly exclude this situation and assume that treatment assignment is independent of historical outcomes given covariates and unmeasured confounders (allowed to be time-variant); see Assumption 1 in our revised paper.

Consider the directed acyclic graph in Figure R.2 where past outcomes at $t = t_1 < t_2$ can directly affect treatment assignment D. Imagine the situation where we do not know the fact that $Y_{t_1} \to D$, and

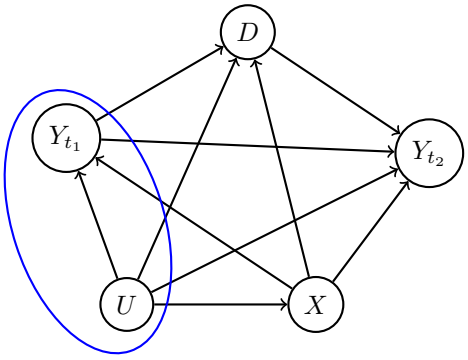

Figure R.2: A DAG for $(Y_{t_1}, Y_{t_2}, D, X, U)$. Here, U denotes some unmeasured confounders, X denotes observed covariates, D denotes the treatment, and Y_{t_1}, Y_{t_2} denotes historical and current outcomes, respectively.

naturally we do not include historical controls as our observed covariates. Accordingly, both historical outcomes Y_{t_1} and U serve as unmeasured confounders *in effect*. Our method is still applicable but unable to use information of Y_{t_1}, leading to efficiency and power loss. To highlight our main contribution, we decide to rule this situation out by imposing the condition that $Y_{t_m}^{(d)} \perp\!\!\!\perp D \mid (X, U)$ for $d \in \{0, 1\}$.

Q4. *Relatedly, what specific causal assumptions do you need to make in order to establish Theorem 1? Presumably you are assuming some sort of positivity assumption and SUTVA. You likely do not assume strong ignorability (i.e. $(Y(1), Y(0))$and D are conditionally independent given X) and it may be helpful to point this out.*

We very much appreciate your helpful comments and suggestions. As you suggest, we impose the SUTVA and consistency assumption to formalize the Rubin causal model; see Lines 88–90. In addition to Assumption 1 of our paper at the first version, we impose the condition that for $d \in \{0, 1\}$, $P(D = d \mid X) > 0$, and $Y_{t_m}^{(d)} \perp\!\!\!\perp D \mid (X, U)$.

Notice that our positivity condition is weaker than the conventionally assumed version $P(D = d \mid X, U) > 0$ for $d = 0, 1$. This is indeed not surprising because we also require the transportability condition that there exists a partition of observable covariate space, and in each local region, confounding effects change mildly across time; thus, $P(D = d \mid X) > 0$ is enough for partial identification.

Q5. *How does one set λ in practice (generally) and how did you set it in your experiments? Is it just through cross-validation?*

Thanks for your question. In our experiments, we directly set $\lambda = 1$, based on the consideration that the two terms in Equation 12 are comparable as both of them can be viewed as an empirical risk. We perform sensitivity analysis of tuning λ in simulation studies, and the results demonstrate that $\lambda = 1$ is an acceptable choice in practice, see the following Table R.2.

How to tune hyperparameters in the causal inference literature remains an important open problem. Cross-validation is commonly used to select hyperparameters in prediction tasks, but often faces difficulties in causal tasks; the reason is that the ground truth of treatment effect for each unit (no matter belongs to the training data or testing data) can never be known, since for each unit only one potential outcome can be observed.

Table R.2: MAE (mean±s.d.) of `GBCT` on the simulated data ($\phi = 1$). Scenario I represents the absence of unmeasured confounding ($\mathrm{MAE}_{\mathrm{CATT}}$) , whereas II represents the presence of unmeasured confounding ($\mathrm{MAE}_{\mathrm{CATE}}$).

λ	0	0.5	1	2	3
I	2.06 ± 0.34	1.80 ± 0.17	1.81 ± 0.10	1.76 ± 0.08	1.88 ± 0.18
II	3.89 ± 0.57	2.32 ± 0.39	1.71 ± 0.06	1.78 ± 0.14	1.82 ± 0.12

Q6. *Does your method provide any uncertainty quantification? As presented, I don't think it can easily. It may be useful to compare the proposed method to Bayesian Causal Forests, which does provide a type of uncertainty quantification while still flexibly targeting the CATE.*

Thanks for your question. As you point out, our method cannot provide uncertainty quantification directly, whereas the confidence interval can be constructed by the Bootstrap procedure. We add this point as one of the limitations of our method in the discussion, and leave it to the future work.