Causal Effect Inference for Structured Treatments

Jean Kaddour*
Centre for Artificial Intelligence
University College London

Yuchen Zhu
Centre for Artificial Intelligence
University College London

Qi Liu
Department of Computer Science
University of Oxford

Matt J. Kusner
Centre for Artificial Intelligence
University College London

Ricardo Silva
Department of Statistical Science
University College London

Abstract

We address the estimation of conditional average treatment effects (CATEs) for structured treatments (e.g., graphs, images, texts). Given a weak condition on the effect, we propose the *generalized Robinson decomposition*, which (i) isolates the causal estimand (reducing regularization bias), (ii) allows one to plug in arbitrary models for learning, and (iii) possesses a quasi-oracle convergence guarantee under mild assumptions. In experiments with small-world and molecular graphs we demonstrate that our approach outperforms prior work in CATE estimation.

1 Introduction

Estimating feature-level causal effects, so-called *conditional average treatment effects* (CATEs), from observational data is a fundamental problem across many domains. Examples include understanding the effects of non-pharmaceutical interventions on the transmission of COVID-19 in a specific region [12], how school meal programs impact child health [13], and the effects of chemotherapy drugs on cancer patients [51]. Supervised learning methods face two challenges in such settings: (i) *missing interventions*, the fact that we only observe one treatment for each individual means models must extrapolate to new treatments without access to ground truth, and (ii) *confounding factors* that affect both treatment assignment and the outcome means that extrapolation from observation to intervention requires assumptions. Many approaches have been proposed to overcome these issues [1, 2, 3, 4, 5, 6, 7, 9, 10, 15, 18, 19, 21, 22, 23, 25, 27, 29, 33, 39, 41, 42, 45, 51, 55, 56, 59, 63, 66].

In many cases, treatments are naturally *structured*. For instance, a drug is commonly represented by its molecular structure (graph), the nutritional content of a meal as a food label (text), and geographic regions affected by a new policy as a map (image). Taking this structure into account can provide several advantages: (i) higher data-efficiency, (ii) capability to work with many treatments, and (iii) generalizing to unseen treatments during test time. However, the vast majority of prior work operates on either binary or continuous scalar treatments (structured treatments are rarely considered, a notable exception to this trend is Harada & Kashima [16] which we describe in Section 2).

To estimate CATEs with structured interventions, our contributions include:

- **Generalized Robinson decomposition (GRD):** A generalization of the Robinson decomposition [47] to treatments that can be vectorized as a continuous embedding. This GRD reveals a learnable

*Correspondence to jean.kaddour.20@ucl.ac.uk

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

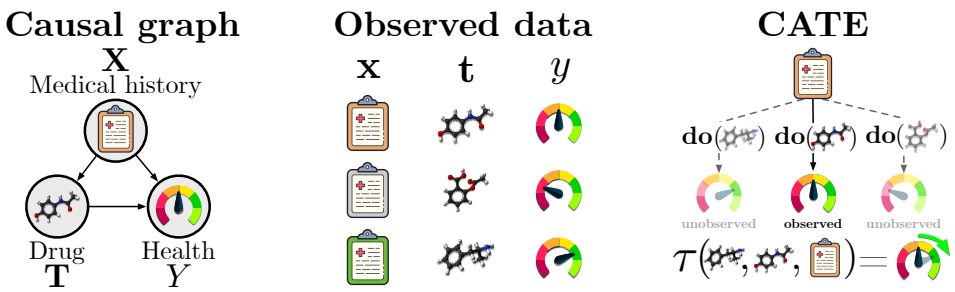

Figure 1: **Illustration of CATE estimation with structured treatments (e.g., molecular graphs).** *Left*: Problem setup with features $\mathbf{X}$, treatment $\mathbf{T}$, and outcome Y. *Center*: Observations the estimator has access to, typically containing only one outcome per individual. *Right*: The CATE is the difference between the expected outcomes given a fixed individual and a pair of treatments.

pseudo-outcome target that isolates the causal component of the observed signal by partialling out confounding associations. Further, it allows one to learn the nuisance and target functions using any supervised learning method, thus extending recent work on *plug-in estimators* [42, 29].

- **Quasi-oracle convergence guarantee:** A result that shows that given access to estimators of certain nuisance functions, as long as the estimates converge at an $O(n^{-1/4})$ rate, the target estimator for the CATE achieves the same error bounds as an oracle who has ground-truth knowledge of both nuisance components, the propensity features, and conditional mean outcome.
- **Structured Intervention Networks (SIN)**: A practical algorithm using GRD, representation learning, and alternating gradient descent. Our PyTorch [43] implementation is online.[2]
- **Evaluation metrics** designed for structured treatments. Since previous evaluation protocols of CATE estimators have mostly focused on binary or scalar-continuous treatment settings, we believe that our proposed evaluation metrics can be useful for comparing future work.
- **Experimental results** with graph treatments in which SIN outperforms previous approaches.

2 Related Work

Closest to our work is GraphITE [16], a method that learns representations of graph interventions for CATE estimation. They propose to minimize prediction loss plus a regularization term that aims to control for confounding based on the Hilbert-Schmidt Independence Criterion (HSIC) [14]. This technique suffers from two drawbacks: (i) the HSIC requires multiplication of kernel matrices and scales quadratically in the batch size; (ii) selecting the HSIC kernel hyper-parameter is not straightforward, as ground-truth CATEs are never observed, and empirical loss does not bound CATE estimation error [1]. We discuss other related work not on structured treatments in Appendix A.

3 Preliminaries

3.1 Conditional Average Treatment Effects (CATEs)

Imagine a dataset where each example $(\mathbf{x}_i, \mathbf{t}_i, y_i) \in \mathcal{D}$ represents a hospital patient's medical history record $\mathbf{x}_i$, prescribed drug treatment $\mathbf{t}_i$, and health outcome y_i, as illustrated in Figure 1 (*Center*). Further, we wish to understand how changing the treatment changes a patient's health outcome. The CATE, $\tau\left(\mathbf{t}', \mathbf{t}_i, \mathbf{x}_i\right)$, describes the expected change in outcome for individuals with history $\mathbf{x}_i$, when treatment $\mathbf{t}_i$ is replaced by $\mathbf{t}'$, depicted in Figure 1 (*Right*). In real-world scenarios, we only observe one outcome for each patient at one treatment level. Further, the patient's pre-treatment health conditions $\mathbf{x}_i$ influence both the doctor's treatment prescription and outcome, thereby *confounding* the effect of the treatment on the outcome.

Formally, we have the dataset $\mathcal{D} = \left\{(\mathbf{x}_i, \mathbf{t}_i, y_i)\right\}_{i=1}^{n}$ sampled from a joint distribution $p\left(\mathbf{X}, \mathbf{T}, Y\right)$, where $Y = f\left(\mathbf{X}, \mathbf{T}\right) + \varepsilon$, as depicted in Figure 1 (*Left*). We define the causal effect of fixing

[2] https://github.com/JeanKaddour/SIN

treatment variable $\mathbf{T} \in \mathcal{T}$ to a value $\mathbf{t}$ on *outcome* variable $Y \in \mathbb{R}$ using the do-operator [44] as $\mathbb{E}\left[Y \mid \text{do}(\mathbf{T}=\mathbf{t})\right]$. Crucially, this estimate differs from the conditional expectation $\mathbb{E}\left[Y \mid \mathbf{T}=\mathbf{t}\right]$ in that it describes the effect of an external entity *intervening* on $\mathbf{T}$ by fixing it to a value $\mathbf{t}$ (removing the edge $\mathbf{X} \to \mathbf{T}$). We further condition on pre-treatment *covariates* $\mathbf{X}$ to define the conditional causal estimand $\mathbb{E}\left[Y \mid \mathbf{X}=\mathbf{x}, \text{do}(\mathbf{T}=\mathbf{t})\right]$. The *conditional average treatment effect* (CATE) is the difference between expected outcomes at different treatment values $\mathbf{t}, \mathbf{t}'$ for given covariates $\mathbf{x}$,

$$\tau(\mathbf{t}', \mathbf{t}, \mathbf{x}) \triangleq \underbrace{\mathbb{E}\left[Y \mid \mathbf{X}=\mathbf{x}, \text{do}(\mathbf{T}=\mathbf{t}')\right]}_{=:\mu_{\mathbf{t}'}(\mathbf{x})} - \underbrace{\mathbb{E}\left[Y \mid \mathbf{X}=\mathbf{x}, \text{do}(\mathbf{T}=\mathbf{t})\right]}_{=:\mu_{\mathbf{t}}(\mathbf{x})}, \tag{1}$$

where $\mu_{\mathbf{t}}(\mathbf{x})$ is defined as the *expected outcome* for a covariate vector $\mathbf{x}$ under treatment $\mathbf{t}$.

Because we do not observe both treatments $\mathbf{t}, \mathbf{t}'$ for a single covariate $\mathbf{x}$, we need to make assumptions that allow us to identify the CATE from observational data.

Assumption 1. *(Unconfoundedness) There are no confounders of the effect between* $\mathbf{T}$ *and* Y *beyond* $\mathbf{X}$*. Therefore,* $Pr\left(Y \leq y \mid \mathbf{x}, do(\mathbf{t})\right) = Pr\left(Y \leq y \mid \mathbf{x}, \mathbf{t}\right)$*, for all* $(\mathbf{x}, \mathbf{t}, y)$*.*

Assumption 2. *(Overlap) It holds that* $0 < p\left(\mathbf{t} \mid \mathbf{x}\right) < 1$*, for all* $(\mathbf{x}, \mathbf{t})$*.*

Assumption 2 means that all sub-populations have some probability of receiving any value of treatment (otherwise, some $\tau(\mathbf{t}', \mathbf{t}, \mathbf{x})$ may be undefined or impossible to estimate.) These assumptions allow us to estimate the causal quantity $\tau(\mathbf{t}', \mathbf{t}, \mathbf{x})$ through statistical estimands:

$$\tau\left(\mathbf{t}', \mathbf{t}, \mathbf{x}\right) = \mu_{\mathbf{t}'}(\mathbf{x}) - \mu_{\mathbf{t}}(\mathbf{x}) = \mathbb{E}\left[Y \mid \mathbf{X}=\mathbf{x}, \mathbf{T}=\mathbf{t}'\right] - \mathbb{E}\left[Y \mid \mathbf{X}=\mathbf{x}, \mathbf{T}=\mathbf{t}\right]. \tag{2}$$

While one can model $\mu_t(\mathbf{x})$ with regression models, such approaches suffer from bias [9, 26, 29] due to two factors: (i) associations between $\mathbf{X}$ and $\mathbf{T}$, due to confounding, makes it hard to identify the distinct contributions of $\mathbf{X}$ and $\mathbf{T}$ on Y, and (ii) regularization for predictive performance can harm effect estimation. Mitigating these biases relies on exposing and removing *nuisance components*. This transforms the optimization into a (regularized) regression problem that isolates the causal effect.

3.2 Robinson Decomposition

One way to formulate such nuisance components is via the *Robinson decomposition* [47]. Originally a reformulation of the CATE for binary treatments, it was used by the *R-learner* [42] to construct a plug-in estimator. The R-learner exploits the decomposition by partialling out the confounding of $\mathbf{X}$ on $\mathbf{T}$ and Y. It also isolates the CATE, thereby removing regularization bias.

Let the treatment variable be $T \in \{0, 1\}$ and the outcome model $p\left(y \mid \mathbf{x}, \mathbf{t}\right)$ parameterized as

$$Y = f(\mathbf{X}, T) + \varepsilon \equiv \mu_0(\mathbf{X}) + T \times \tau_b(\mathbf{X}) + \varepsilon, \tag{3}$$

where we define error term ε such that $\mathbb{E}\left[\varepsilon \mid \mathbf{x}, \mathbf{t}\right] = \mathbb{E}\left[\varepsilon \mid \mathbf{x}\right] = 0$, and $\tau_b(\mathbf{x}) \triangleq \tau(1, 0, \mathbf{x})$.

Define the *propensity score* $e(\mathbf{x}) \triangleq p\left(T = 1 \mid \mathbf{x}\right)$ and the *conditional mean outcome* as

$$m(\mathbf{x}) \triangleq \mathbb{E}\left[Y \mid \mathbf{x}\right] = \mu_0(\mathbf{x}) + e(\mathbf{x})\,\tau_{\mathrm{b}}(\mathbf{x}). \tag{4}$$

From model (3) and the previous definitions, it follows that

$$Y - m(\mathbf{X}) = \left(T - e(\mathbf{X})\right)\tau_{\mathrm{b}}(\mathbf{X}) + \varepsilon, \tag{5}$$

allowing us to define the estimator

$$\widehat{\tau}_{\mathrm{b}}(\cdot) = \underset{\tau_{\mathrm{b}}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{y}_i - \tilde{t}_i \times \tau_{\mathrm{b}}(\mathbf{x}_i)\right)^2 + \Lambda\left(\tau_{\mathrm{b}}(\cdot)\right)\right\}, \tag{6}$$

where $\tilde{y}_i \triangleq y_i - \widehat{m}(\mathbf{x}_i)$ and $\tilde{t}_i \triangleq t_i - \widehat{e}(\mathbf{x}_i)$ are pseudo-data points defined through estimated nuisance functions $\widehat{m}(\cdot), \widehat{e}(\cdot)$, which can be learned separately with any supervised learning algorithm.

4 The Generalized Robinson Decomposition

Our goal is to estimate the CATE $\tau(\mathbf{t}', \mathbf{t}, \mathbf{x})$ for structured interventions $\mathbf{t}', \mathbf{t}$ (e.g., graphs, images, or text embeddings) while accounting for the confounding of $\mathbf{X}$ on $\mathbf{T}$ and Y. Inspired by the Robinson decomposition, which has enabled flexible CATE estimation for binary treatments [6, 9, 33, 42], we propose the *Generalized Robinson Decomposition* from which we extract a pseudo-outcome that targets the causal effect. We demonstrate this decomposition's usefulness from both a theoretical (quasi-oracle convergence rate in Section 4.2) and practical point of view (*Structured Intervention Networks* in Section 5). For details on its motivation and derivation, we refer the reader to Appendix B.

4.1 Generalizing the Robinson Decomposition

To generalize the Robinson decomposition to structured treatments, we introduce two concepts: (a) the assumption that the causal effect is a *product effect*: the outcome function $f^*(\mathbf{X}, \mathbf{T})$ can be written as an inner product of two functionals over the covariates and the treatment, respectively, and (b) *propensity features*, which have a natural connection to the propensity score.

Assumption 3. *(Product effect) We consider the following partial parameterization of $p(y \mid \mathbf{x}, \mathbf{t})$,*

$$Y = \underbrace{g(\mathbf{X})^\top h(\mathbf{T})}_{=: f^*(\mathbf{X}, \mathbf{T})} + \varepsilon, \tag{7}$$

where $g : \mathcal{X} \to \mathbb{R}^d, h : \mathcal{T} \to \mathbb{R}^d$ and $\mathbb{E}[\varepsilon \mid \mathbf{x}, \mathbf{t}] = \mathbb{E}\left[\varepsilon \mid \mathbf{x}\right] = 0$, for all $(\mathbf{x}, \mathbf{t}) \in \mathcal{X} \times \mathcal{T}$.

We consider this assumption to be mild, as we can formally justify its universality. The following asserts that provided we allow the dimensionality of g and h to grow, we may approximate any arbitrary bounded continuous functions in $\mathcal{C}(\mathcal{X} \times \mathcal{T})$ where $\mathcal{X} \times \mathcal{T}$ is compact.

Proposition 1. *(Universality of product effect) Let $\mathcal{H}_{\mathcal{X} \times \mathcal{T}}$ be a Reproducing Kernel Hilbert Space (RKHS) on the set $\mathcal{X} \times \mathcal{T}$ with universal kernel k. For any $\epsilon > 0$, and any $f \in \mathcal{H}_{\mathcal{X} \times \mathcal{T}}$, there is a $d \in \mathbb{N}$ such that there exist two $d-$dimensional vector fields $g : \mathcal{X} \to \mathbb{R}^d$ and $h : \mathcal{T} \to \mathbb{R}^d$, where $\|f - g^\top h\|_{L_2(P_{\mathcal{X} \times \mathcal{T}})} \leq \epsilon$. (Proof in App. C)*

Our quantity of interest, the CATE of structured treatments $\mathbf{t}', \mathbf{t}$, given covariates $\mathbf{x}$, follows as

$$\tau\left(\mathbf{t}', \mathbf{t}, \mathbf{x}\right) = g(\mathbf{x})^\top \left(h\left(\mathbf{t}'\right) - h(\mathbf{t})\right). \tag{8}$$

Define *propensity features* $e^h(\mathbf{x}) \triangleq \mathbb{E}\left[h(\mathbf{T}) \mid \mathbf{x}\right]$ and $m(\mathbf{x}) \triangleq \mathbb{E}\left[Y \mid \mathbf{x}\right] = g(\mathbf{x})^\top e^h(\mathbf{x})$.

Following the same steps as in Section 3.2, the generalized Robinson decomposition for eq. (7) is

$$\boxed{Y - m(\mathbf{X}) = g(\mathbf{X})^\top \left(h(\mathbf{T}) - e^h(\mathbf{X})\right) + \varepsilon.} \tag{9}$$

For a fixed $h(\cdot)$, a direct generalization of the R-learner to structured treatments is the solution of

$$\widehat{g}(\cdot) = \arg\min_g \left\{ \frac{1}{n} \sum_{i=1}^n \left(Y_i - \widehat{m}(\mathbf{X}_i) - g(\mathbf{X}_i)^\top \left(h(\mathbf{T}_i) - \widehat{e}^h(\mathbf{X}_i)\right)\right)^2 + \Lambda\left(g(\cdot)\right) \right\}. \tag{10}$$

4.2 Quasi-oracle error bound of Generalized Robinson Decomposition

We establish the main theoretical result of our paper: a *quasi-oracle convergence guarantee* for our formulation of the generalized Robinson decomposition under a finite-basis representation of the outcome function. This result is analogous to the R-learner for binary CATEs [42]: when the true $e(\mathbf{x}), m(\mathbf{x})$ are unknown, and we only have access to the estimators $\widehat{e}(\mathbf{x}), \widehat{m}(\mathbf{x})$, then as long as the estimates converge at $n^{-1/4}$ rate, the estimator $\widehat{\tau}_{\mathrm{b}}(\mathbf{x})$ achieves the same error bounds as an *oracle* who has ground-truth knowledge of these two nuisance components.

More formally, provided the nuisance estimators $\widehat{m}(\cdot)$ and $\widehat{e}^h(\cdot)$ converge at an $O\left(n^{-1/4}\right)$ rate, our estimator $\widehat{f}$ will converge at $\widetilde{O}(n^{-\frac{1}{2(1+p)}})$ rate for arbitrarily small p, recovering the parametric convergence rate for when the true $m(\cdot)$ and $e(\cdot)$ are provided as oracle quantities.

Our analysis assumes that the outcome $\mathbb{E}\left[Y \mid \mathbf{X}=\mathbf{x}, \mathbf{T}=\mathbf{t}\right]$ can be written as a linear combination of orthonormal basis functions, and overlap in the features of $\mathbf{X}$ and $\mathbf{T}$.

Assumption 4. *Let $\boldsymbol{\alpha}(\mathbf{X}) \in \mathbb{R}^{d_{\boldsymbol{\alpha}}}$, $\boldsymbol{\beta}(\mathbf{T}) \in \mathbb{R}^{d_{\boldsymbol{\beta}}}$ be fixed orthonormal basis features on $\mathbf{X} \in \mathbb{R}^{d_{\mathbf{x}}}$, $\mathbf{T} \in \mathbb{R}^{d_{\mathbf{t}}}$, respectively. The true outcome function $f^*(\mathbf{x}, \mathbf{t}) = \mathbb{E}[Y \mid \mathbf{X}=\mathbf{x}, \mathbf{T}=\mathbf{t}]$ can be written as $f^*(\mathbf{x}, \mathbf{t}) = \boldsymbol{\alpha}^\top(\mathbf{x})\boldsymbol{\Theta}^*\boldsymbol{\beta}(\mathbf{t})$ for some matrix of coefficients $\boldsymbol{\Theta}^*$.*

Assumption 5 (Overlap in features)**.** *The marginal distribution of features $\mathcal{P}_{\boldsymbol{\alpha}(\mathcal{X})\boldsymbol{\beta}(\mathcal{T})}$ is positive, i.e. $\operatorname{supp}[\mathcal{P}_{\boldsymbol{\alpha}(\mathcal{X})\boldsymbol{\beta}(\mathcal{T})}] = \boldsymbol{\alpha}(\mathcal{X})\boldsymbol{\beta}(\mathcal{T})$.*

Assumption 5 is typically weaker than requiring overlap in $\mathbf{X}$ and $\mathbf{T}$, e.g., when $d_{\boldsymbol{\alpha}}, d_{\boldsymbol{\beta}} \ll d_{\mathbf{x}}, d_{\mathbf{t}}$.

With further technical assumptions specified in Appendix F, we establish the following theorem.

Theorem 2. *Let $\boldsymbol{\Theta}^*$ denote the representer of the true outcome function. Suppose Assumptions 5, 6, and 4 hold. Moreover, suppose that the propensity estimate $\widehat{e}^h$ is uniformly consistent,*

$$\sup_{\mathbf{x}\in\mathcal{X}} \|\widehat{e}^h(\mathbf{x}) - e^h(\mathbf{x})\| \to_p 0 \tag{11}$$

and the L_2 errors converge at rate

$$\mathbb{E}\left[\left\{\widehat{m}(\mathbf{X}) - m^*(\mathbf{X})\right\}^2\right], \mathbb{E}\left[\|\widehat{e}^h(\mathbf{X}) - e^h(\mathbf{X})\|^2\right] = \mathcal{O}(a_n^2) \tag{12}$$

for some sequence $a_n \to 0$, where (a_n) is such that $a_n = O(n^{-\kappa})$ with $\kappa > \frac{1}{4}$. Further, we define the regret as the excess risk

$$R\left(\widehat{\boldsymbol{\Theta}}_n\right) \triangleq L\left(\widehat{\boldsymbol{\Theta}}_n\right) - L\left(\boldsymbol{\Theta}^*\right), \; L\left(\boldsymbol{\Theta}\right) \triangleq \mathbb{E}\left[\left\{\left(Y - m^*\left(\mathbf{X}\right)\right) - \boldsymbol{\alpha}\left(\mathbf{X}\right)\boldsymbol{\Theta}\left(\boldsymbol{\beta}\left(\mathbf{T}\right) - e^h\left(\mathbf{X}\right)\right)\right\}^2\right]. \tag{13}$$

Suppose that we obtain $\widehat{\boldsymbol{\Theta}}_n$ via a penalized basis function regression variant of the generalized Robinson decomposition, with a properly chosen penalty $\Lambda_n\left(\|\widehat{\boldsymbol{\Theta}}_n\|_2\right)$ (specified in the proof). Then, $\widehat{\boldsymbol{\Theta}}_n$ satisfies the regret bound: $R\left(\widehat{\boldsymbol{\Theta}}_n\right) = \widetilde{O}(r_n^2)$ with $r_n = n^{-\frac{1}{2(1+p)}}$ for arbitrarily small $p > 0$.

5 Structured Intervention Networks

We introduce *Structured Intervention Networks* (SIN), a two-stage training algorithm for neural networks, which enables both flexibility in learning complex causal relationships as well as scalability to large data-sets. This implementation of GRD strikes a balance between theory and practice: while we assumed a fixed $h(\cdot)$ in eq. (10), in practice, we often need to learn the feature map $\widehat{h}\left(\mathbf{T}\right)$ from data. We leave the convergence analysis of this representation learning setting for future work.

5.1 Training Algorithm

We propose to simultaneously learn feature maps $\widehat{g}\left(\mathbf{X}\right), \widehat{h}\left(\mathbf{T}\right)$ using alternating gradient descent, so that they can adapt to each other. A remaining challenge is that learning $\widehat{e}^h(\mathbf{X})$ is now entangled with learning $\widehat{h}\left(\mathbf{T}\right)$. While the R-learner is based on the idea of *cross-fitting*, where at each data point i we pick estimates of the nuisances that do not use that data point, we introduce a pragmatic representation learning approach for $(\widehat{g}, \widehat{h})$ that does not use cross-fitting[3].

We learn surrogate models for the mean outcome and propensity features $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})$ and $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})$ with parameters $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}, \boldsymbol{\eta} \in \mathbb{R}^{d_{\boldsymbol{\eta}}}$, as well as feature maps for covariates and treatments $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})$,

[3]We could in principle use cross-fitting for $\widehat{e}^h$, although the loop between fitting $\widehat{h}$ alternating with $\widehat{e}^h$ would break the overall independence between $\widehat{e}^h_i(\mathbf{X})$ and data point i. While it is possible that cross-fitting for $\widehat{e}^h$ is still beneficial in this case, for simplicity and for computational savings, we did not implement it.

a SIN Training.

Input: Stage 1 data $\mathcal{D}_1 := \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, Stage 2 data $\mathcal{D}_2 := \{(\mathbf{x}_i, \mathbf{t}_i, y_i)\}_{i=1}^n$ Step sizes $\lambda_{\boldsymbol{\theta}}, \lambda_{\boldsymbol{\eta}}, \lambda_{\boldsymbol{\psi}}, \lambda_{\boldsymbol{\phi}}$. Number of update steps K. Mini-batch sizes B_1, B_2.

Initialize parameters: $\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\psi}, \boldsymbol{\phi}$
while not converged **do** ▷ *Stage 1*
  Sample mini-batch $\{(\mathbf{x}_b, y_b)\}_{b=1}^{m_{B_1}}$
  Evaluate $J_m(\boldsymbol{\theta})$
  Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda_{\boldsymbol{\theta}} \widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
end while
while not converged **do** ▷ *Stage 2*
  Sample mini-batch $\{(\mathbf{x}_b, \mathbf{t}_b, y_b)\}_{b=1}^{n_{B_2}}$
  Evaluate $J_{g,h}(\boldsymbol{\psi}, \boldsymbol{\phi}), J_{e^h}(\boldsymbol{\eta})$
  for $k = 1$ to K **do**
    Update $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \lambda_{\boldsymbol{\phi}} \widehat{\nabla}_{\boldsymbol{\phi}} J_{g,h}(\boldsymbol{\psi}, \boldsymbol{\phi})$
    Update $\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} - \lambda_{\boldsymbol{\psi}} \widehat{\nabla}_{\boldsymbol{\psi}} J_{g,h}(\boldsymbol{\psi}, \boldsymbol{\phi})$
  end for
  Update $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} - \lambda_{\boldsymbol{\eta}} \widehat{\nabla}_{\boldsymbol{\eta}} J_{e^h}(\boldsymbol{\eta})$
end while

b Pseudocode in a PyTorch-like style.

```
# Initialize submodels and optimizers
m, e, g, h = MLP(...), MLP(...), MLP(...),
     GNN(...)
m_opt, e_opt, g_opt, h_opt = Adam(m.params(),
      m_lr), Adam(e.params(), e_lr), ...

# Stage 1
for batch in train_loader:
    X, Y = batch.X, batch.Y
    m_opt.zero_grad()
    F.mse_loss(m(X), Y).backward()
    m_opt.step()

# Stage 2
for batch in train_loader:
    X, T, Y = batch.X, batch.T, batch.Y
    for _ in range(num_update_steps):
        g_opt.zero_grad()
        h_opt.zero_grad()
        F.mse_loss((g(X)*(h(T) - e(X))).sum
             (-1), (Y-m(X))).backward()
        g_opt.step()
        h_opt.step()
    e_opt.zero_grad()
    F.mse_loss(e(X), h(T)).backward()
    e_opt.step()
```

Figure 2: The two-stage algorithm for training SIN.

parameterized by $\boldsymbol{\psi} \in \mathbb{R}^{d_{\boldsymbol{\psi}}}, \boldsymbol{\phi} \in \mathbb{R}^{d_{\boldsymbol{\phi}}}$. We denote regularizers by $\Lambda(\cdot)$. Figure 2 summarizes the algorithm. As the mean outcome model $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})$ does not depend on the other components, we learn it separately in Stage 1. In Stage 2, we alternate between learning $\boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\eta}$.

Stage 1: Learn parameters $\boldsymbol{\theta}$ of the mean outcome model $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})$ based on the objective

$$J_m(\boldsymbol{\theta}) = \sum_{i=1}^m \left(y_i - \widehat{m}_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)^2 + \Lambda(\boldsymbol{\theta}), \tag{14}$$

which relies only on covariates and outcome data $\mathcal{D}_1 := \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^m$.

Stage 2: Learn parameters $\boldsymbol{\psi}, \boldsymbol{\phi}$ for the covariates and treatments feature maps $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})$, as well as parameters $\boldsymbol{\eta}$ for the propensity features $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})$.

$$J_{g,h}(\boldsymbol{\phi}, \boldsymbol{\psi}) = \sum_{i=1}^n \left(y_i - \left\{ \widehat{m}_{\boldsymbol{\theta}}(\mathbf{x}_i) + \widehat{g}_{\boldsymbol{\psi}}(\mathbf{x}_i)^\top \left(\widehat{h}_{\boldsymbol{\phi}}(\mathbf{t}_i) - \widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{x}_i) \right) \right\} \right)^2 + \Lambda(\boldsymbol{\psi}) + \Lambda(\boldsymbol{\phi}). \tag{15}$$

This loss hinges on $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})$, which needs to be learned by

$$J_{e^h}(\boldsymbol{\eta}) = \sum_{i=1}^n \left\| \widehat{h}_{\boldsymbol{\phi}}(\mathbf{t}_i) - \widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{x}_i) \right\|_2^2 + \Lambda(\boldsymbol{\eta}), \tag{16}$$

and in turn depends on $\widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})$. While it may be tempting to learn $\boldsymbol{\psi}, \boldsymbol{\phi}$ and $\boldsymbol{\eta}$ jointly, they have fundamentally different objectives ($\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})$ is defined as an estimate of the expectation $\mathbb{E}\left[h(\mathbf{T}) \mid \mathbf{x}\right]$). Therefore, we employ an alternating optimization procedure, where we take $k \in \{1, \ldots, K\}$ optimization steps for $\boldsymbol{\psi}, \boldsymbol{\phi}$ towards $J_{g,h}(\boldsymbol{\psi}, \boldsymbol{\phi})$ and one step for learning $\boldsymbol{\eta}$. We observe that setting $K > 1$, i.e. updating $\boldsymbol{\psi}, \boldsymbol{\phi}$ more frequently than $\boldsymbol{\eta}$, stabilizes the training process.

5.2 Advantages of SIN

We conclude by describing the beneficial properties of SIN, particularly in finite-sample regimes:

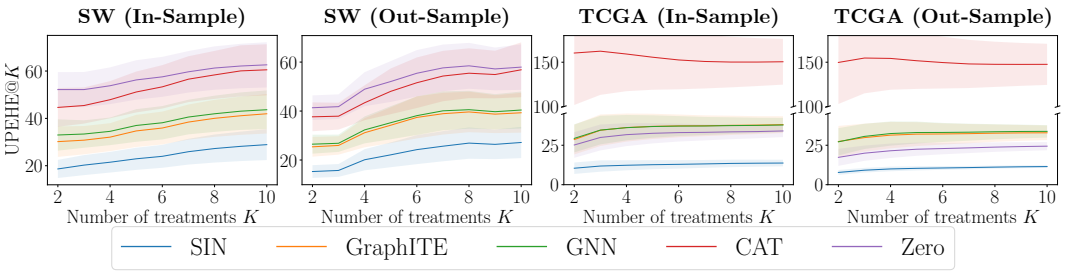

Figure 3: UPEHE@K for $K \in \{2, \ldots, 10\}$.

1. **Targeted regularization**: Regularizing $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})$ in eq. (15) after partialing out confounding is a type of targeted regularization of the isolated causal effect. In contrast, outcome estimation methods can suffer from regularization-induced confounding, e.g., regularizing the effect estimate away from zero in the service of trying to improve predictive performance [29].
2. **Propensity features**: Learning propensity features can help us to (i) partial out parts of $\mathbf{X}$ that cause the treatment but not the outcome, and (ii) dispose unnecessary components of $\mathbf{T}$.
3. **Data-efficiency**: In contrast to methods that split the data into disjoint models for each treatment group (known as *T-learners* for binary treatments [8, 10]), sharing causal effect parameters between all covariates regardless of their assigned treatment increases data-efficiency.
4. **Partial data**: In settings without access to both the treatment assignment and the outcome but only access to one of them, one can leverage that data to improve the (nuisance) estimator further, e.g., when a patient's recovery is observed one year after a drug was administered [33].

6 Experiments

The goal is to evaluate how deconfounding through SIN compares with prior methods.

6.1 Experimental Setup

Datasets. To be able to compute CATE estimation error w.r.t. a ground truth, we design two causal models: a simpler synthetic model with small-world graph treatments and a more complex model with real-world molecular graph treatments and gene expression covariates. The Small-World (SW) simulation contains 1,000 uniformly sampled covariates and 200 randomly generated Watts–Strogatz small-world graphs [60] as treatments. *The Cancer Genomic Atlas* (TCGA) simulation uses 9,659 gene expression measurements of cancer patients for covariates [61] and 10,000 sampled molecules from the QM9 dataset [46] as treatments. Appendix D details the data-generating schemes.

Baselines. We compare our method to (1) **Zero**, a sanity-check baseline that consistently predicts zero treatment effect and equals the mean squared treatment effect (poorly regularized models may perform worse than that due to confounding), (2) **CAT**, a categorical treatment variable model using one-hot encoded treatment indicator vectors, (3) **GNN**, a model that first encodes treatments with a GNN and then concatenates treatment and individual features for regression, (4) **GraphITE** [16], a CATE estimation method designed for graph treatments (more details in Section 2). GNN and CAT reflect the performance of standard regression models. The contrast between these two provides insight into whether the additional graph structure of the treatment improves CATE estimation. To deal with unseen treatments during CAT's evaluation, we map such to the most similar ones seen during training based on their Euclidean distance in the embedding space of the GNN baseline.

Graph models. For small-world networks, we use *k-dimensional GNNs* [38], as to distinguish graphs they take higher-order structures into account. To model molecular graphs, we use *Relational Graph Convolutional Networks* [49], where the nodes are atoms and each edge type corresponds to a specific bond type. We use the implementations of PyTorch Geometric [11].

Evaluation metrics. We extend the *expected Precision in Estimation of Heterogeneous Effect* (PEHE) commonly used in binary treatment settings [19] to arbitrary pairs of treatments $(\mathbf{t}, \mathbf{t}')$ as

Table 1: Error of CATE estimation for all methods, measured by WPEHE@6. Results are averaged over 10 trials, $\pm$ denotes std. error (each trial samples treatment assignment matrix $\mathbf{W}$).

Method	**SW**		**TCGA**	
	In-sample	Out-sample	In-sample	Out-sample
Zero	56.26 ± 8.12	53.77 ± 8.93	26.63 ± 7.55	17.94 ± 4.86
CAT	51.75 ± 8.85	49.76 ± 9.73	155.88 ± 52.82	146.62 ± 42.32
GNN	37.10 ± 6.84	36.74 ± 7.42	30.67 ± 8.29	27.57 ± 7.95
GraphITE	34.81 ± 6.70	35.94 ± 8.07	30.31 ± 8.96	27.48 ± 8.95
SIN	$\mathbf{23.00 \pm 4.56}$	$\mathbf{23.19 \pm 5.56}$	$\mathbf{10.98 \pm 3.45}$	$\mathbf{8.15 \pm 1.46}$

follows. We denote the *Unweighted PEHE* (UPEHE) and the *Weighted PEHE* (WPEHE) as

$$\epsilon_{\text{UPEHE(WPEHE)}} \triangleq \int_{\mathcal{X}} \left(\widehat{\tau}\left(\mathbf{t}', \mathbf{t}, \mathbf{x}\right) - \tau\left(\mathbf{t}', \mathbf{t}, \mathbf{x}\right)\right)^2 p\left(\mathbf{t} \mid \mathbf{x}\right) p\left(\mathbf{t}' \mid \mathbf{x}\right) p\left(\mathbf{x}\right) \mathrm{d}\mathbf{x}, \tag{17}$$

where the weighted version gives less importance to treatment pairs that are less likely; to account for the fact that such pairs will have higher estimation errors. In fact, as the reliability of estimated effects decreases by how likely they are in the observational study, we evaluate all methods on U/WPEHE truncated to the top K treatments, which we call U/WPEHE@K. To compute this, for each $\mathbf{x}$, we rank all treatments by their propensity $p\left(\mathbf{t} \mid \mathbf{x}\right)$ (given by the causal model) in descending order. We take the top K treatments and compute the U/WPEHE for all $\binom{K}{2}$ treatment pairs.

In-sample vs. out-sample. A common benchmark for causal inference methods is the *in-sample* task, which we include here for completeness: estimating CATEs for covariate values $\mathbf{x}$ found in the training set. This task is still non-trivial, as the outcome of only one treatment is observed during training [4]. In contrast, and arguably of more relevance to decision making, the goal of the *out-sample* task is to estimate CATEs for completely unseen covariate realizations $\mathbf{x}'$.

Hyper-parameter tuning. To ensure a fair comparison, we perform hyper-parameter optimization with random search for all models on held-out data and select the best hyper-parameters over 10 runs.

Propensity. We define the propensity (or *treatment selection bias*) as $p\left(\mathbf{T} \mid \mathbf{x}\right) = \mathrm{softmax}\left(\kappa \mathbf{W}^\top \mathbf{X}\right)$, where $\mathbf{W} \in \mathbb{R}^{|\mathcal{T}| \times d}, \forall i, j : W_{ij} \sim \mathcal{U}\left[0, 1\right]$ is a random matrix (sampled then fixed for each run). Recall $|\mathcal{T}|$ is the number of available treatments and let d be the dimensionality of the covariates. Here the *bias strength* κ is a temperature parameter that determines the flatness of the propensity (the lower the flatter, i.e., $\kappa = 0$ corresponds to the uniform distribution).

6.2 Comparison of Performances on different K Treatments

Figure 3 shows the UPEHE@K of all methods for $K \in \{2, \ldots, 10\}$. We also report the WPEHE@6 of all methods in Table 1. Unless stated otherwise, we report results for bias strengths $\kappa = 10$ and $\kappa = 0.1$ in the SW and TCGA datasets, respectively across 10 random trials.

The results indicate that the relative performance of each method, for both the in-sample and out-sample estimation tasks, is consistent. Further, they suggest that, overall, the performance of SIN is best due to a better isolation of the causal effect from the observed data compared to other methods. The performance difference between CAT and GNN across all results indicate that accounting for graph information significantly improves the estimates. We observe from the SW experiments that GraphITE [16] performs slightly better than GNN, while it is nearly the same as GNN on TCGA.

Surprisingly, the results of the TCGA experiments with low bias strength $\kappa = 0.1$ expose that all models but SIN fail to isolate causal effects better than the Zero baseline. These results confirm that confounding effects of $\mathbf{X}$ on Y combined with moderate causal effects can cause severe regularization bias for black-box regression models, while SIN partials these out from the outcome by $\widehat{m}_{\boldsymbol{\theta}}\left(\mathbf{X}\right)$. We include additional results on convergence and larger values of K in Appendix E.1.

[4] The original motivation comes from Fisherian designs where the only source of randomness is on the treatment assignment [20]. Our motivation is simpler: rule out the extra variability from different covariates, highlighting the difference between methods due to different loss functions and less due to smoothing abilities.

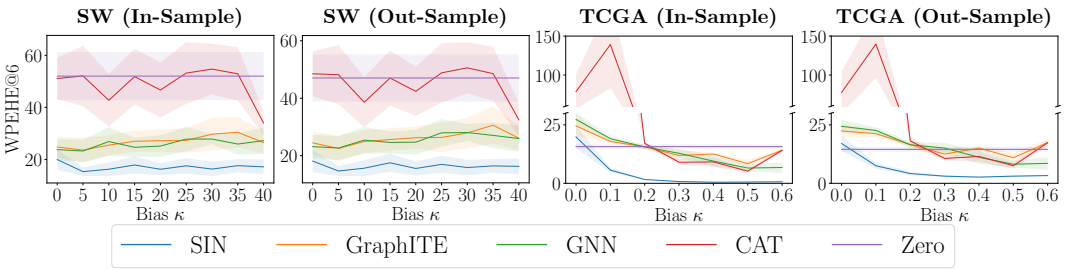

Figure 4: WPEHE@6 over increasing bias strength κ.

6.3 Comparison of Robustness to different Bias Strengths κ

A strong selection bias (i.e. large κ) in the observed data makes CATE estimation more difficult, as it becomes unlikely to see certain treatments $\mathbf{t} \in \mathcal{T}$ for particular covariates $\mathbf{x}$. Here, we assess each model's robustness to varying levels of selection bias, determined by κ, across 5 random seeds. In Figure 4, we see that SIN outperforms the baselines across the entire range of considered biases. Interestingly, SIN performs competitively even in a case with no selection bias ($\kappa = 0$, which corresponds to a randomized experiment). Importantly, all performances seem to either stagnate (SW) or to increase (TCGA) with increasing biases. Notably, the poor performance of CAT suddenly improves on datasets with high bias. We believe this is because, in high bias regimes, we see fewer distinct treatments overall, which allows the CAT model to approach the performance of GNN.

7 Limitations, Future Work and Potential Negative Societal Impacts

Limitations and future work. Firstly, in some real-life domains, Assumption 1 (Unconfoundedness) can be too strong, as there may exist *hidden confounders*. There are two common strategies to deal with them: utilizing *instrumental variables* [17, 57, 62] or *proxy variables* [35, 37, 58]. Developing new approaches for structured interventions in such settings is a promising future direction. Secondly, SIN is based on neural networks; however, neural network initialization can impact final estimates. To obtain consistency guarantees, GRD can be combined with kernel methods [35, 57].

Potential negative societal impacts. Because causal inference methods make recommendations about interventions to apply in real-world settings, misapplying them can have a negative real-world impact. It is crucial to thoroughly test these methods on realistic simulations and alter aspects of them to understand how violations of assumptions impact estimation. We have aimed to provide a comprehensive evaluation of structured treatment methods by showing how estimation degrades as less likely treatments are considered (Figure 3) and as treatment bias increases (Figure 4).

8 Conclusion

The main contributions of this paper are two-fold: (i) the generalized Robinson decomposition that yields a pseudo-outcome targeting the causal effect while possessing a quasi-oracle convergence guarantee under mild assumptions, and (ii) Structured Intervention Networks, a practical algorithm using representation learning that outperforms prior approaches in experiments with graph treatments.

Acknowledgements

We thank Antonin Schrab, David Watson, Jakob Zeitler, Limor Gultchin, Marc Deisenroth and Shonosuke Harada for useful discussions and constructive feedback on the paper. JK and YZ acknowledge support by the Engineering and Physical Sciences Research Council with grant number EP/S021566/1. This work was partially supported by an ONR grant number N62909-19-1-2096 to RS. We thank the Alan Turing Institute for the provision of Azure cloud computing resources.

References

[1] Alaa, A. and van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th*

International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pp. 129–138, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[2] Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Arbour, D., Dimmery, D., and Sondhi, A. Permutation weighting. *arXiv preprint arXiv:1901.01230*, 2020.

[4] Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[5] Athey, S. and Wager, S. Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*, 2019.

[6] Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47 (2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709.

[7] Bica, I., Jordon, J., and van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[8] Caron, A., Manolopoulou, I., and Baio, G. Estimating individual treatment effects using non-parametric regression models: a review. *arXiv preprint arXiv:2009.06472*, 2020.

[9] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221.

[10] Curth, A. and van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 2021.

[11] Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[12] Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

[13] for Health Statistics, N. C. et al. 2007–2008 national health and nutrition examination survey (nhanes). *US Department of Health and Human Services, Centers for Disease Control and Prevention: Hyattsville, MD, USA*, 2008.

[14] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. A kernel statistical test of independence. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 585–592. Curran Associates, Inc., 2007.

[15] Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[16] Harada, S. and Kashima, H. Graphite: Estimating individual effects of graph-structured treatments. *arXiv preprint arXiv:2009.14061*, 2020.

[17] Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1414–1423. PMLR, 06–11 Aug 2017.

[18] Hatt, T. and Feuerriegel, S. Estimating average treatment effects via orthogonal regularization. *arXiv preprint arXiv:2101.08490*, 2021.

[19] Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

[20] Imbens, G. and Rubin, D. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[21] Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 2004. doi: 10.1162/003465304323023651.

[22] Jesson, A., Mindermann, S., Shalit, U., and Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

[23] Jesson, A., Mindermann, S., Gal, Y., and Shalit, U. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*. PMLR, 2021.

[24] Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

[25] Kallus, N. DeepMatch: Balancing deep covariate representations for causal inference using adversarial training. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5067–5077. PMLR, 13–18 Jul 2020.

[26] Kennedy, E. H. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[27] Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(4):1229, 2017.

[28] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[29] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1804597116.

[30] Li, F. et al. Propensity score weighting for causal inference with multiple treatments. *Annals of Applied Statistics*, 13(4):2389–2415, 2019.

[31] Lopez, M. J. and Gutman, R. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pp. 432–454, 2017.

[32] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[33] Lu, D., Tao, C., Chen, J., Li, F., Guo, F., and Carin, L. Reconsidering generative objectives for counterfactual reasoning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21539–21553. Curran Associates, Inc., 2020.

[34] Ma, K. W., Lewis, J. P., and Kleijn, W. B. The HSIC bottleneck: Deep learning without back-propagation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 5085–5092. AAAI Press, 2020.

[35] Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M. J., Gretton, A., and Muandet, K. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*. PMLR, 2021.

[36] Mendelson, S. and Neeman, J. Regularization in kernel learning. *The Annals of Statistics*, 38(1): 526 – 565, 2010. doi: 10.1214/09-AOS728. URL https://doi.org/10.1214/09-AOS728.

[37] Miao, W., Geng, Z., and Tchetgen, E. T. Identifying causal effects with proxy variables of an unmeasured confounder. *arXiv preprint arXiv:1609.08816*, 2018.

[38] Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4602–4609, Jul. 2019. doi: 10.1609/aaai.v33i01.33014602.

[39] Nabi, R., McNutt, T., and Shpitser, I. Semiparametric causal sufficient dimension reduction of high dimensional treatments. *arXiv preprint arXiv:1710.06727*, 2020.

[40] Neal, B., Huang, C.-W., and Raghupathi, S. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2021.

[41] Nie, L., Ye, M., qiang liu, and Nicolae, D. {VCN}et and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021.

[42] Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 09 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa076.

[43] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[44] Pearl, J. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[45] Pollmann, M. Causal inference for spatial treatments. *arXiv preprint arXiv:2011.00373*, 2020.

[46] Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

[47] Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 00129682, 14680262.

[48] Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d. PMID: 23088335.

[49] Schlichtkrull, M. S., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., and Alam, M. (eds.), *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pp. 593–607. Springer, 2018.

[50] Schwab, P., Linhardt, L., and Karlen, W. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2019.

[51] Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., and Karlen, W. Learning Counterfactual Representations for Estimating Individual Dose-Response Curves. In *AAAI Conference on Artificial Intelligence*, 2020.

[52] Sejdinovic, D. and Gretton, A. What is an rkhs?, 2014. URL http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_2014.pdf.

[53] Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

[54] Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[55] Shi, C., Veitch, V., and Blei, D. Invariant representation learning for treatment effect estimation. *arXiv preprint arXiv:2011.12379*, 2020.

[56] Silva, R. Observational-interventional priors for dose-response learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[57] Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4595–4607, 2019.

[58] Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

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

[60] Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world'networks. *nature*, 393 (6684):440–442, 1998.

[61] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[62] Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

[63] Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[64] Zhang, D. W., Burghouts, G. J., and Snoek, C. G. M. Set prediction without imposing structure as conditional density estimation. In *International Conference on Learning Representations*, 2021.

[65] Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.

[66] Zhu, Y., Coffman, D. L., and Ghosh, D. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1):25–40, 2015. doi: doi:10.1515/jci-2014-0022.

Proof. We have established that when we set ρ_n as

$$\rho_n(c) = U(\epsilon)\{1 + \log(n) + \log\log(c+e)\}\left(\frac{(c+1)^p \log(n)}{\sqrt{n}}\right)^{2/(1+p)},$$

we have that for every ϵ there exist a constant $U(\epsilon)$ such that for large enough n the following is satisfied with probability at least $1-\epsilon$:

$$\frac{1}{2}\tilde{R}_n(f_\Theta; c) - \rho_n(c) \leq R_n(f_\Theta; c) \leq 2\tilde{R}_n(f_\Theta; c) + \rho_n(c) \tag{120}$$

Subsection F.6 argued that this leads to a rate of $\tilde{\mathcal{O}}(n^{-\frac{1}{1+p}})$ for $R(\tilde{\mathbf{\Theta}})$.

Now to show that feasible learner matches the rate of the oracle learner,

Eq. 119 implies that

$$R(\mathbf{\Theta}; c) \leq 2\tilde{R}_n(\mathbf{\Theta}; c) + \rho_n(c) \tag{121}$$

$$\leq 2\hat{R}_n(\tau; c) + 0.25kR(\tau; c) + k\rho_n(c) \tag{122}$$

Rearranging the inequality implies that

$$R(\mathbf{\Theta}; c) \leq \frac{8}{3}\hat{R}_n(\mathbf{\Theta}; c) + 2\rho_n(c) \tag{123}$$

for large n for all $1 < c < \log(n)$, with probability at least $1 - 2\epsilon$. It can then be checked following a symmetrical argument, that

$$\frac{3}{8}\hat{R}_n(\mathbf{\Theta}; c) - 2\rho_n(c) \leq R(\mathbf{\Theta}; c) \leq \frac{8}{3}\hat{R}_n(\mathbf{\Theta}; c) + 2\rho_n(c) \tag{124}$$

for n large enough for all $1 \leq c \leq \log(n)$ with probability at least $1 - 4\epsilon$.

Then, following the same argument as [42], we find that the feasible minimizer has the same regret bound as the oracle minimizer: $R(\hat{\mathbf{\Theta}}_n) = \tilde{\mathcal{O}}\left(n^{-\frac{1}{1+p}}\right)$.

This is to say:

$$\boxed{R(\hat{\mathbf{\Theta}}_n) = O(r_n^2),\ r_n = n^{-\frac{1}{2(1+p)}}} \tag{125}$$

□