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 [52]. 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, 52, 56, 57, 60, 64, 67].

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).

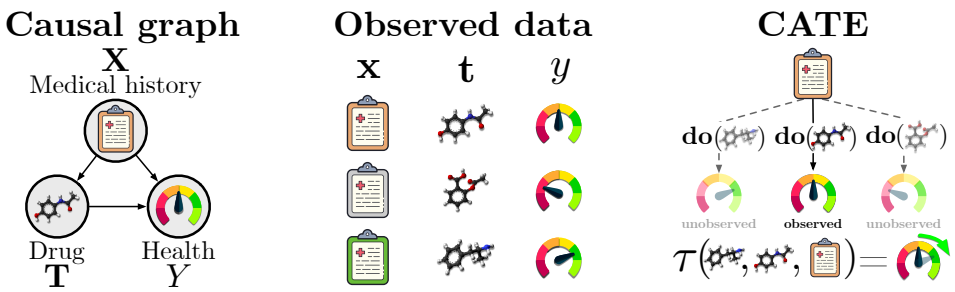


Figure 1: **Illustration of CATE estimation with structured treatments (e.g., molecular graphs).** *Left*: Problem setup with features $\mathbf{X}$, treatment $\mathbf{T}$, and outcome Y. *Center*: Observations the estimator has access to, typically containing only one outcome per individual. *Right*: 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* [48] $e(\mathbf{x}) \triangleq p\left(T = 1 \mid \mathbf{x}\right)$ 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, text) 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 the usefulness of this decomposition from both a theoretical view (quasi-oracle convergence rate in Section 4.2) and practical 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) we assume 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 separate functionals, one over the covariates and one over the treatment, and (b) *propensity features*, which partial out the effects from the covariates on the treatment features. Similar techniques have been previously shown to add to the robustness of estimation [9, 42].

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

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

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 $(\mathbf{x}, \mathbf{t}) \in \mathcal{X} \times \mathcal{T}$.

This assumption is 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 $\delta > 0$, 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 $\|f - g^\top h\|_{L_2(P_{\mathcal{X}\times\mathcal{T}})} \leq \delta$. (Proof in Appendix C)*

This assumption allows us to simplify the expression of the CATE for treatments $\mathbf{t}', \mathbf{t}$, given $\mathbf{x}$,

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

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

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

Given nuisance estimates $\widehat{m}(\cdot), \widehat{e}^h(\cdot)$, we can use this decomposition to derive an optimization problem for $h(\cdot), g(\cdot)$ (note $\widehat{e}^h(\cdot)$ implicitly depends on $h(\cdot)$, we address this dependence in Section 5).

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

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 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 $e(\cdot), m(\cdot)$ are unknown, and we only have access to the estimators $\widehat{e}(\cdot), \widehat{m}(\cdot)$, then as long as the estimates converge at $n^{-1/4}$ rate, the estimator $\widehat{\tau}_{\mathrm{b}}(\cdot)$ 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 CATE estimator will converge at an $\widetilde{O}(n^{-\frac{1}{2(1+p)}})$ rate for arbitrarily small $p > 0$, recovering the parametric convergence rate for when the true $m(\cdot)$ and $e^h(\cdot)$ 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 fixed basis functions. By Proposition 1, as long as we have enough basis functions, this representation is flexible enough to capture the true outcome function.

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, known 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 (unknown) matrix of coefficients $\boldsymbol{\Theta}^*$.*

Note that by setting $g = \boldsymbol{\alpha}^\top \Theta^*$ and $h = \boldsymbol{\beta}$, we recover eq. (7). Additionally, we will need overlap in the basis features $\boldsymbol{\alpha}(\mathcal{X}), \boldsymbol{\beta}(\mathcal{T})$.

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

Assumption 5 is typically weaker than requiring overlap in $\mathbf{X}$ and $\mathbf{T}$, i.e., 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 flexibility in learning complex causal relationships, and scalability to large data-sets. This implementation of GRD strikes a balance between theory and practice: while we assumed fixed basis-functions in Section 4.2, in practice, we often need to learn the feature maps 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].

[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 $\widehat{e}_i^h(\mathbf{X})$ 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.

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 $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})$, 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

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

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 $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})$.

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

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

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

note again the dependence 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.

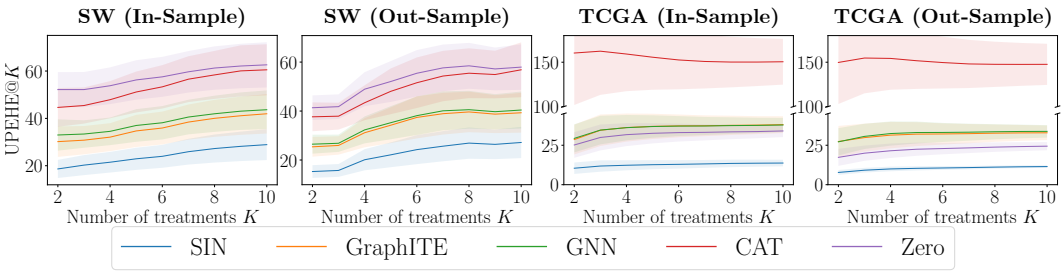


Figure 3: UPEHE@K for $K \in \{2, \ldots, 10\}$.

5.2 Advantages of SIN

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

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

Here we evaluate how CATE estimation with our proposed model 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 [61] as treatments. *The Cancer Genomic Atlas* (TCGA) simulation uses 9,659 gene expression measurements of cancer patients for covariates [62] 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 CATE 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* [50], where the nodes are atoms and each edge type corresponds to a specific bond type. We use the implementations of PyTorch Geometric [11].

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	$\mathbf{23.00 \pm 4.56}$	$\mathbf{23.19 \pm 5.56}$	$\mathbf{10.98 \pm 3.45}$	$\mathbf{8.15 \pm 1.46}$

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 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.

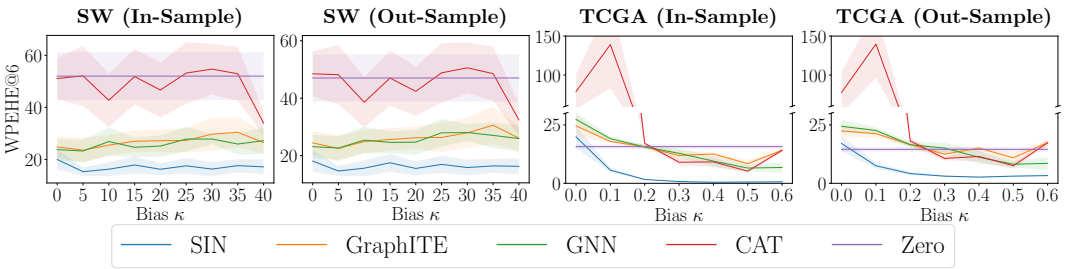


Figure 4: WPEHE@6 over increasing bias strength κ.

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, 58, 63] or *proxy variables* [35, 37, 59]. 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, 58].

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] Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] For clarity, we added a bullet-point for each of our contributions, roughly matching the structure of Section 4.
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide proofs in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] https://github.com/JeanKaddour/SIN.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Additional details are in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We will provide these details in the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] We will provide these details in the Appendix.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Other Related Work

Plug-in estimators. A recent line of work for CATE estimation derives *plug-in estimators* [8].[5] These work by decomposing CATE estimation into multiple sub-problems (so-called *nuisance components*), each solvable using any supervised learning method [10, 15, 26, 29, 42]. Currently, these approaches are limited to binary treatment setups. Our approach is inspired by these methods, extending plug-in estimation to structured treatment settings.

[5] These are also called *meta-learners*. To avoid confusion with *meta-learning*, we call these *plug-in estimators*.

CATE estimation with neural networks. Neural network CATE estimators typically use separate prediction heads for each treatment option [24, 32, 41, 51, 52, 54, 55]. This architectural design reduces one source of regularization bias: the influence of the treatment indicator variable might be lost in the high-dimensional network representations. Extending this idea directly to structured treatments would not only be computationally expensive, but would also not be able to make use of treatment features or learn treatment representations.

Multiple treatments. While Inverse Probability Weighting (IPW) [30, 31, 66] is a popular technique for estimating effects with multiple, categorical treatments, it requires estimating the propensity density which is infeasible in settings with hundreds or thousands of treatments; some of which may have not been seen during training. Nabi et al. [39] propose a framework for sufficient dimensionality reduction of high-dimensional treatments based on semiparametric inference theory. Besides relying on IPW, this approach is designed for average treatment effects (not CATEs).

B The Generalized Robinson Decomposition

B.1 Motivation

frequently the influence of $\mathbf{T}$ on Y is very different from the influence of $\mathbf{X}$ on Y. Specifically, $f(\mathbf{X}, \mathbf{T})$ often has different smoothness in $\mathbf{X}$ and $\mathbf{T}$. For instance, different health histories $\mathbf{X}$ for a fixed treatment $\mathbf{t}$ will have a much more variable effect on Y than different treatments $\mathbf{t}$ for a given history $\mathbf{X}$. This is why methods like the R-learner [42] have carefully separated estimation functions of $\mathbf{X}$ from functions of $\mathbf{T}$ [8].

A generic way to extend the Robinson decomposition to arbitrary treatments is to learn a model $\hat{f}(\mathbf{X}, \mathbf{T})$ defined over the entire outcome surface, via mean outcome $\hat{m}(\mathbf{X})$ and treatment conditional density $p(\mathbf{T} \mid \mathbf{X})$. In this case, we fit the relationship

$$Y - m(\mathbf{X}) = f(\mathbf{X}, \mathbf{T}) - e^p(\mathbf{X}) + \varepsilon, \quad \text{where} \quad e^p(\mathbf{x}) \triangleq \mathbb{E}\left[f(\mathbf{X}, \mathbf{T}) \mid \mathbf{x}\right]. \tag{18}$$

To learn $f(\cdot, \cdot)$ from a dataset $\mathcal{D} = \left\{(\mathbf{x}_i, \mathbf{t}_i, y_i)\right\}_{i=1}^{n}$ we need to solve,

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \left[\left\{y_i - \hat{m}(\mathbf{x}_i)\right\} - \left(f(\mathbf{x}_i, \mathbf{t}_i) - \hat{e}^p(\mathbf{x}_i)\right)\right]^2, \tag{19}$$

where $\mathcal{F}$ is some function space and $\hat{m}$ is a plug-in finite sample estimate of m. Because e^p contains f, we need to estimate it, which we denote $\hat{e}^p$.

One solution is to estimate the propensity $p(\mathbf{T} \mid \mathbf{X})$ and use it to compute $\hat{e}^p$. However, this approach requires conditional density estimation over potentially high-dimensional, structured treatments, which remains an open research question [65], and is prone to high variance [58]. Further, to compute $\hat{e}^p$ from it, one has to resort to Monte Carlo evaluation. By learning propensity features $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})$ instead of $p(\mathbf{T} \mid \mathbf{X})$, we avoid these issues.

Another option is to solve for f, fix it, then estimate $\hat{e}^p$ using regression from finite samples, and iterate to a fixed point. However, there is a fundamental issue with this approach: we are typically interested in regularizing the causal effect directly as opposed to the generic regression function. This is why, for instance, the R-learner parameterizes $\mu_1(\mathbf{x})$ as a (nuisance) baseline $\mu_0(\mathbf{x})$ plus the CATE $\tau_{\mathrm{b}}(\mathbf{x})$. The black-box $f(\mathbf{x}, \mathbf{t})$ does not capture the asymmetry between $\mathbf{x}$ and $\mathbf{t}$ in the implied CATE $f(\mathbf{x}, \mathbf{t}) - f(\mathbf{x}, \mathbf{t}')$. Further, unlike the binary case, in many applications, we do not have a baseline treatment $\mathbf{t}_0$ with respect to which we could parameterize f in terms of some $\tau(\mathbf{t}, \mathbf{t}_0, \mathbf{x})$. To regularize the causal effect more directly, we make the product effect assumption, which allows us to partial out confounding.

B.2 Derivation in detail

We consider the product effect parameterization of $p(y \mid \mathbf{x}, \mathbf{t})$,

$$Y = \underbrace{g(\mathbf{X})^\top h(\mathbf{T})}_{=: f(\mathbf{X}, \mathbf{T})} + \varepsilon, \tag{20}$$

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 $(\mathbf{x}, \mathbf{t}) \in \mathcal{X} \times \mathcal{T}$. Rearranging eq. (20) yields the Robinson residual

$$\varepsilon = Y - g\left(\mathbf{X}\right)^\top h\left(\mathbf{T}\right), \tag{21}$$

which we aim to rewrite in terms of $m(\mathbf{X})$. To this end, we define *propensity features* $e^h\left(\mathbf{X}\right)$ as

$$e^h\left(\mathbf{X}\right) \triangleq \mathbb{E}\left[h\left(\mathbf{T}\right) \mid \mathbf{X}\right], \quad \text{such that} \quad m\left(\mathbf{X}\right) = \mathbb{E}\left[Y \mid \mathbf{X}\right] = g\left(\mathbf{X}\right)^\top e^h\left(\mathbf{X}\right). \tag{22}$$

To obtain the generalized Robinson decomposition, one rewrites eq. (21) as

$$\varepsilon = Y - \left(g\left(\mathbf{X}\right)^\top \left[h\left(\mathbf{T}\right) + \cancel{e^h\left(\mathbf{X}\right)} - \cancel{e^h\left(\mathbf{X}\right)}\right]\right) \tag{23}$$

$$= Y - \left(g\left(\mathbf{X}\right)^\top e^h\left(\mathbf{X}\right) + g\left(\mathbf{X}\right)^\top \left(h\left(\mathbf{T}\right) - e^h\left(\mathbf{X}\right)\right)\right) \tag{24}$$

$$= Y - \underbrace{\left(g\left(\mathbf{X}\right)^\top e^h\left(\mathbf{X}\right)\right)}_{m(\mathbf{X})} - g\left(\mathbf{X}\right)^\top \left(h\left(\mathbf{T}\right) - e^h\left(\mathbf{X}\right)\right). \tag{25}$$

Hence, the generalized Robinson decomposition is

$$Y - m(\mathbf{X}) = g\left(\mathbf{X}\right)^\top \left(h\left(\mathbf{T}\right) - e^h\left(\mathbf{X}\right)\right) + \varepsilon. \tag{26}$$

C Universality of Product Decomposition

Proof of Proposition 1.

Proof. Define $\mathcal{H}_{0\mathcal{X}\times\mathcal{T}}$ as $\left\{f\left(\mathbf{x}, \mathbf{t}\right) = \sum_{i=1}^n \alpha_i k\left(\left(\mathbf{x}_i, \mathbf{t}_i\right), \left(\mathbf{x}, \mathbf{t}\right)\right) | n \in \mathbb{N}, \alpha_{i=1,\cdots,n} \in \mathbb{R}\right\}$. By definition, the RKHS $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ is the set of pointwise limits of Cauchy sequences $(f_n)_n \in \mathcal{H}_{0\mathcal{X}\times\mathcal{T}}$. By Lemma 41 of [53], the Cauchy sequences also converges in the $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ norm.

For any $f \in \mathcal{H}_{\mathcal{X}\times\mathcal{T}}$, pick its Cauchy sequence $(f_n) in \mathcal{H}_{0\mathcal{X}\times\mathcal{T}}$. Since $\sum_{i=1}^\infty \alpha_i k\left(\left(\mathbf{x}_i, \mathbf{t}_i\right), \left(\mathbf{x}, \mathbf{t}\right)\right)$ converges in $\|\cdot\|_{\mathcal{H}_{\mathcal{X}\times\mathcal{T}}}$, for any $\tilde{\epsilon}$ there exist a $\tilde{d}$ such that let $f_{\tilde{d}} = \sum_{i=1}^{\tilde{d}} \alpha_i k\left(\left(\mathbf{x}_i, \mathbf{t}_i\right), \left(\mathbf{x}, \mathbf{t}\right)\right)$, then

$$\|f_{\tilde{d}} - f\|_{\mathcal{H}_{\mathcal{X}\times\mathcal{T}}} \leq \tilde{\epsilon} \tag{27}$$

Since for any RKHS with kernel k, the RKHS norm is always an upper bound on the L_2 norm up to scaling by a constant C_k,

$$\|f_{\tilde{d}} - f\|_{L_2(P_{(\mathcal{X}\times\mathcal{T})})} \leq C_k \|f_{\tilde{d}} - f\|_{\mathcal{H}_{\mathcal{X}\times\mathcal{T}}} \leq C_k \tilde{\epsilon} \tag{28}$$

Then for any ϵ, we can choose $d \in \mathbb{N}$ such that $\|f_d - f\|_{L_2(P_{\mathcal{X}\times\mathcal{T}})} \leq C_k \cdot \frac{\epsilon}{C_k} = \epsilon$.

It remains to show that f_d can be written as $g^\top h$ as required. $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ is isometrically isomorphic to $\mathcal{H}_{\mathcal{X}} \times \mathcal{H}_{\mathcal{T}}$; we can decompose k into the product kernel

$$k\left(\left(\mathbf{x}, \mathbf{t}\right), \left(\mathbf{x}', \mathbf{t}'\right)\right) = k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}'\right) k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}'\right). \tag{29}$$

Thus $f_d\left(\mathbf{x}, \mathbf{t}\right) = \sum_{i=1}^d \alpha_i k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}_i) k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}_i)$. Set $g(\mathbf{x}) = \left(\alpha_1 k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}_1\right), \cdots, \alpha_d k_{\mathcal{X}}\left(\mathbf{x}, \mathbf{x}_d\right)\right)^\top$, $h(\mathbf{t}) = \left(k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}_1\right), \cdots, k_{\mathcal{T}}\left(\mathbf{t}, \mathbf{t}_d\right)\right)^\top$, we obtain $f_d = g^\top$. □

D Experimental Details

D.1 Simulations

Baseline effect Similarly as in [7, 10, 41], for each run of the experiment, we randomly sample a vector $\mathbf{u}_0 \sim \mathcal{U}(\mathbf{0}, \mathbf{1})$, and set $\mathbf{v}_0 = \mathbf{u}_0 / \|\mathbf{u}_0\|$ where $\|\cdot\|$ is the Euclidean norm. We then model the baseline effect as

$$\mu_0\left(\mathbf{x}\right) = \mathbf{v}_0^\top \mathbf{x}. \tag{30}$$

D.1.1 Small-World Networks

Covariates We uniformly sample 20-dimensional multivariate covariates $\mathbf{X} \sim \mathcal{U}(-\mathbf{1}, \mathbf{1})$. The in-sample dataset consists of 1,000 units, and the out-sample one of 500. For the treatment assignment, we square the covariates element-wise; i.e., we sample treatment assignments according to $p\left(\mathbf{T} \mid \mathbf{x}^2\right)$.

Graph interventions For each graph intervention, we uniformly sample a number of nodes between 10 and 120, number of neighbors for each node between 3 and 8, and the probability of rewiring each edge between 0.1 and 1 Then, we repeatedly generate Watts–Strogatz small-world graphs until we get a connected one. Each vertex has one feature, which is its degree centrality. We denote a graph's node connectivity as $\nu(\mathcal{G})$ and its average shortest path length as $l(\mathcal{G})$.

Outcomes Analogously as for the baseline effect, we generate two randomly sampled vectors $\mathbf{v}_\nu$ and $\mathbf{v}_l$. Then, given an assigned graph treatment $\mathcal{G}$ and a covariate vector $\mathbf{x}$, we generate the outcome as

$$Y = 100\mu_0(\mathbf{x}) + 0.2\nu(\mathcal{G})^2 \cdot \mathbf{v}_\nu^\top \mathbf{x} + l(\mathcal{G}) \cdot \mathbf{v}_l^\top \mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \tag{31}$$

D.1.2 TCGA

Covariates The *The Cancer Genomic Atlas* (TCGA) simulation uses 4,000-dimensional 9,659 gene expression measurements of cancer patients for covariates [62], i.e., each unit is a covariate vector $\mathbf{X} \in \mathbb{R}^{4000}$. The in-sample and out-sample datasets consist of 5,000 and 4,659 units, respectively. In each run, the units are split randomly into in- and out-sample datasets. We used the same version of the TCGA dataset as used by Bica et al. [7] and Schwab et al. [52].

Graph interventions In each run, we randomly sample 10,000 molecules from the Quantum Machine 9 (QM9) dataset [46, 49] (with 133k molecules in total). For each molecule, we create a relational graph, where each node corresponds to an atom and consist of 78 atom features. An edge corresponds to the chemical bond type, where we label each edge correspondingly, considering *single*, *double*, *triple* and *aromatic* bonds. Furthermore, for each molecule, we obtain 8 of its properties *mu, alpha, homo, lumo, gap, r2, zpve, u0*, which we collect in the vector $\mathbf{z} \in \mathbb{R}^8$.

Outcomes For each covariate vector $\mathbf{x}$, we compute its 8-dimensional PCA components, denoted by $\mathbf{x}^{(\text{PCA})} \in \mathbb{R}^8$. Then, given the molecular properties of the assigned molecule treatment $\mathbf{z}$, we generate outcomes by

$$Y = 10\mu_0(\mathbf{x}) + 0.01\mathbf{z}^\top \mathbf{x}^{(\text{PCA})} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \tag{32}$$

D.2 Hyper-parameters

To ensure a fair comparison between all models, we perform hyper-parameter optimization with random search for all models on held-out data and select the best hyper-parameters over 10 runs. While conceptually, choosing hyper-parameters based on predictive metrics may not necessarily lead to good CATE estimation performance, Neal et al. [40] provide empirical evidence that doing so indeed often does in practice.

Table 2 and Table 4 include the hyper-parameter search ranges we set in the SW and TCGA experiments, respectively. Table 3 and Table 5 include the fixed hyper-parameter values across all SW and TCGA experiments, respectively. We restricted the number of hyper-parameter optimization trials to 10 in all experiments. We observed that all models' performances are rather insensitive to hyper-parameter values in the considered search ranges, i.e., the performances across trials have not varied much. The search ranges for the HSIC penalty λ are taken from the experimental section of the GraphITE paper [16], where the authors also argue that their model's performance is insensitive to this weight. In consultation with Harada & Kashima [16], we use Ma et al. [34]'s implementation of the normalized HSIC. We use early stopping for all models based on their training loss. We noticed that a patience value below 10 often leads to pre-convergence stopping with subsequent sub-optimal performance for all models but GIN.

D.2.1 SW

Hyper-parameter	Search range
Num. of layers for covariates representations	2-4
Num. of layers for treatment representations	3-6
Num. of layers for $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})^*$	3-6
Num. of layers for $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	3-6
Num. of layer for final feed-forward network †	2-6
Dim. of hidden layers for covariates representations	50-300
Dim. of hidden layers for treatment representations	50-300
Dim. of hidden layers for $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})^*$	200-300
Dim. of hidden layers for $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	50-150
Dim. of $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})^*$	50-250
Dim. of final covariates/treatment layer	2-200
Dim. of hidden layers for final feed-forward network	50-300
Num. update steps K^*	10-20
Early stopping patience for $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})^*$	{5, 10}
Early stopping patience for $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T}), \widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	{1, 5}
Learning rates $\lambda_{\boldsymbol{\psi}}, \lambda_{\boldsymbol{\phi}}{}^*$	{5e-4, 1e-3}
Learning rate †	{5e-4, 1e-3}
Dropout for $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})^*$	{0, 0.2}
Dropout for $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	{0, 0.2 }
Weight of HSIC penalty $\lambda^\ddagger$	{0.001, 0.01, 1, 10, 100, 1000}

Table 2: Hyper-parameter search ranges for SW experiments. * denotes hyper-parameter only applicable for GIN; † applicable for all models but GIN, ‡ applicable only for GraphITE.

Hyper-parameter	Value
Optimizer	Adam [28]
Batch size	500
Weight decay (all optims.)	0
$\lambda_{\boldsymbol{\theta}}, \lambda_{\boldsymbol{\eta}}$	1e-3
Early stopping patience †	10
GNN Batch Norm	True
MLP Batch Norm (all MLPs)	False
Activation functions (all layers)	ReLU
Validation set size (in %)	20%

Table 3: Fixed hyper-parameter values across all SW experiments. * denotes hyper-parameter only applicable for GIN; † applicable for all models but GIN, ‡ applicable only for GraphITE.

D.2.2 TCGA

Hyper-parameter	Search range
Num. of layers for covariates representations	2-5
Num. of layers for treatment representations	3-6
Num. of layers for $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})^*$	2-4
Num. of layers for $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	1-6
Num. of layer for final feed-forward network †	1-5
Dim. of hidden layers for covariates representations	100-400
Dim. of hidden layers for treatment representations	100-400
Dim. of hidden layers for $\widehat{m}_{\boldsymbol{\theta}}(\mathbf{X})^*$	100-300
Dim. of hidden layers for $\widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	10-50
Dim. of $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T})^*$	200-600
Dim. of final covariates/treatment layer	2-800
Dim. of hidden layers for final feed-forward network	100-400
Num. update steps K^*	10-20
Early stopping patience for $\widehat{g}_{\boldsymbol{\psi}}(\mathbf{X}), \widehat{h}_{\boldsymbol{\phi}}(\mathbf{T}), \widehat{e}^h_{\boldsymbol{\eta}}(\mathbf{X})^*$	{5, 10}
Learning rates $\lambda_{\boldsymbol{\psi}}, \lambda_{\boldsymbol{\phi}}{}^*$	{5e-4, 1e-3}
Learning rate †	{5e-4, 1e-3}
Weight of HSIC penalty $\lambda^\ddagger$	{0.001, 0.01, 1, 10, 100, 1000}

Table 4: Hyper-parameter search ranges for TCGA experiments. * denotes hyper-parameter only applicable for GIN; † applicable for all models but GIN, ‡ applicable only for GraphITE.

Hyper-parameter	Value
Optimizer	Adam [28]
Batch size	1000
Weight decay (all optims.)	0
$\lambda_{\boldsymbol{\theta}}, \lambda_{\boldsymbol{\eta}}$	1e-3
Early stopping patience †	10
GNN Batch Norm	True
MLP Batch Norm (all MLPs)	False
Activation functions (all layers)	ReLU
Validation set size (in %)	20%

Table 5: Fixed hyper-parameter values across all TCGA experiments. * denotes hyper-parameter only applicable for GIN; † applicable for all models but GIN, ‡ applicable only for GraphITE.

D.2.3 Hardware details

All experiments were run on Microsoft Azure Virtual Machines with 12 Intel Xeon E5-2690 v4 CPUs and 2 NVIDIA Tesla K80 GPUs. No single trial took longer than ~ 30 minutes to run.

E Additional Results

E.1 Comparison of Performances on different K Treatments

We present additional WPEHE@K results for the experiments in Section 6.2 with varying K.

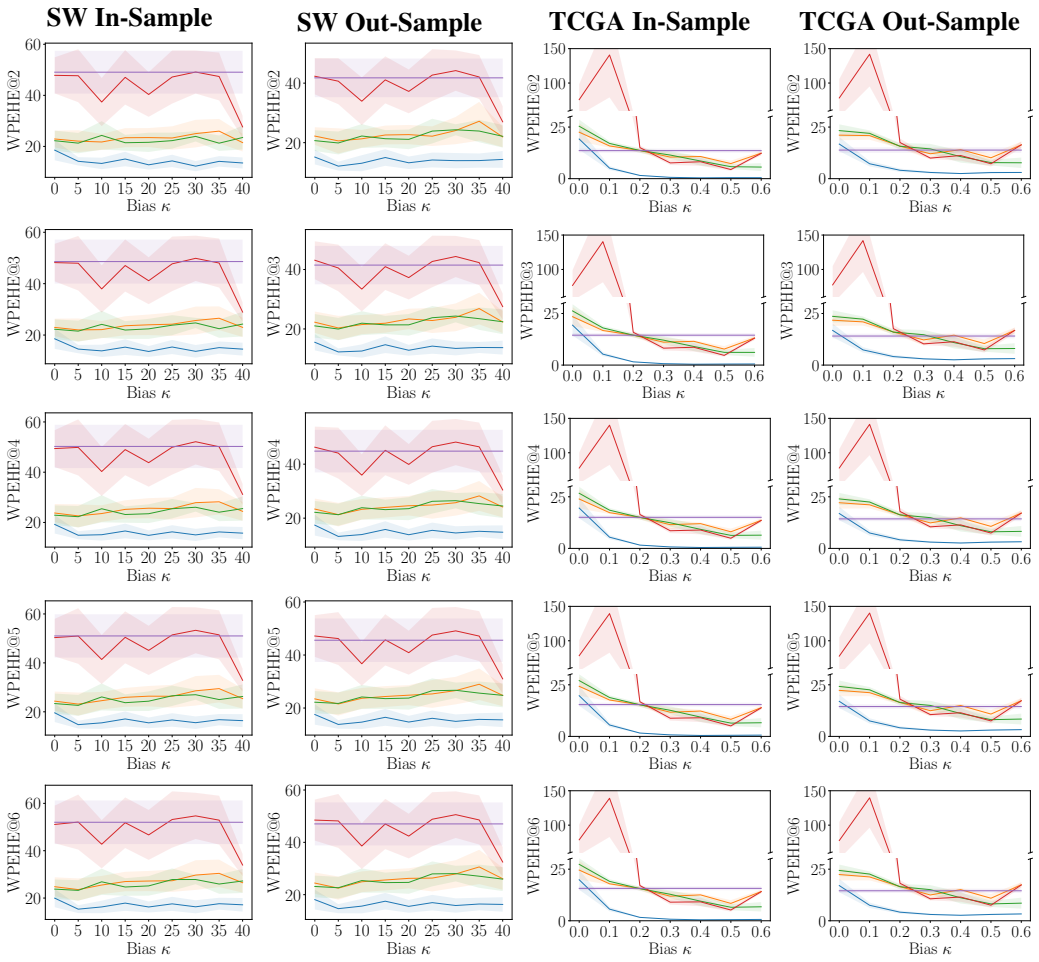
Method	SW		TCGA	
	In-sample	Out-sample	In-sample	Out-sample
WPEHE@2				
Zero	52.17 ± 7.37	41.36 ± 5.04	25.17 ± 8.12	17.33 ± 5.41
CAT	44.63 ± 8.18	37.65 ± 5.90	160.35 ± 58.56	149.75 ± 46.86
GNN	32.98 ± 6.63	26.47 ± 3.87	29.35 ± 8.90	27.17 ± 8.67
GraphITE	30.18 ± 6.45	25.39 ± 4.04	28.60 ± 9.44	27.37 ± 9.87
GIN	$\mathbf{18.00 \pm 3.83}$	$\mathbf{15.30 \pm 2.60}$	$\mathbf{10.44 \pm 3.62}$	$\mathbf{7.76 \pm 1.56}$
WPEHE@3				
Zero	51.61 ± 7.24	41.53 ± 4.96	25.97 ± 7.96	17.50 ± 5.11
CAT	44.87 ± 7.53	37.59 ± 5.46	159.48 ± 56.46	148.80 ± 44.87
GNN	32.97 ± 5.75	26.60 ± 3.70	30.22 ± 8.77	27.29 ± 8.30
GraphITE	30.39 ± 5.89	25.70 ± 3.70	29.71 ± 9.43	27.27 ± 9.38
GIN	$\mathbf{19.79 \pm 4.06}$	$\mathbf{15.54 \pm 2.56}$	$\mathbf{10.62 \pm 3.56}$	$\mathbf{7.94 \pm 1.51}$
WPEHE@4				
Zero	52.92 ± 7.47	47.93 ± 6.68	26.35 ± 7.79	17.76 ± 5.05
CAT	46.95 ± 7.65	42.47 ± 6.91	158.02 ± 54.76	148.08 ± 43.71
GNN	33.89 ± 5.73	31.51 ± 5.27	30.51 ± 8.57	27.53 ± 8.23
GraphITE	31.43 ± 5.75	30.39 ± 5.71	30.07 ± 9.22	27.48 ± 9.28
GIN	$\mathbf{20.78 \pm 4.11}$	$\mathbf{19.50 \pm 4.12}$	$\mathbf{10.76 \pm 3.51}$	$\mathbf{8.08 \pm 1.51}$
WPEHE@5				
Zero	55.02 ± 8.00	50.75 ± 7.92	26.53 ± 7.66	17.91 ± 4.96
CAT	49.78 ± 8.37	46.65 ± 8.86	156.77 ± 53.58	147.20 ± 42.86
GNN	36.06 ± 6.69	34.16 ± 6.41	30.61 ± 8.41	27.61 ± 8.10
GraphITE	33.69 ± 6.56	33.13 ± 6.92	30.22 ± 9.08	27.53 ± 9.12
GIN	$\mathbf{22.06 \pm 4.40}$	$\mathbf{21.19 \pm 4.80}$	$\mathbf{10.90 \pm 3.47}$	$\mathbf{8.13 \pm 1.49}$
WPEHE@6				
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
GIN	$\mathbf{23.00 \pm 4.56}$	$\mathbf{23.19 \pm 5.56}$	$\mathbf{10.98 \pm 3.45}$	$\mathbf{8.15 \pm 1.46}$
WPEHE@7				
Zero	58.16 ± 8.38	55.73 ± 9.01	26.66 ± 7.48	17.97 ± 4.81
CAT	54.62 ± 9.27	52.21 ± 9.74	155.24 ± 52.25	146.15 ± 41.90
GNN	39.21 ± 7.05	38.51 ± 7.50	30.67 ± 8.21	27.56 ± 7.86
GraphITE	37.00 ± 7.10	37.34 ± 8.05	30.33 ± 8.88	27.47 ± 8.86
GIN	$\mathbf{24.71 \pm 5.07}$	$\mathbf{24.46 \pm 5.79}$	$\mathbf{11.02 \pm 3.43}$	$\mathbf{8.17 \pm 1.45}$
WPEHE@8				
Zero	59.57 ± 8.74	56.61 ± 8.94	26.73 ± 7.43	18.03 ± 4.76
CAT	56.24 ± 9.71	53.33 ± 9.71	154.86 ± 51.85	145.94 ± 41.61
GNN	40.44 ± 7.36	39.04 ± 7.33	30.72 ± 8.16	27.49 ± 8.78
GraphITE	38.42 ± 7.46	38.06 ± 7.89	30.39 ± 8.82	27.49 ± 8.78
GIN	$\mathbf{25.90 \pm 5.51}$	$\mathbf{25.63 \pm 6.03}$	$\mathbf{11.10 \pm 3.43}$	$\mathbf{8.20 \pm 1.44}$
WPEHE@9				

Zero	60.39 ± 8.94	55.72 ± 8.44	26.75 ± 7.40	18.06 ± 4.73
CAT	57.78 ± 10.27	53.06 ± 9.36	154.60 ± 51.57	145.73 ± 41.37
GNN	41.45 ± 7.60	38.47 ± 6.92	30.72 ± 8.11	27.60 ± 7.74
GraphITE	39.43 ± 7.69	37.43 ± 7.48	30.39 ± 8.78	27.50 ± 8.72
GIN	$\mathbf{26.76 \pm 5.80}$	$\mathbf{25.30 \pm 5.75}$	$\mathbf{11.12 \pm 3.42}$	$\mathbf{8.22 \pm 1.43}$
WPEHE@10				
Zero	60.92 ± 9.10	56.44 ± 8.91	26.78 ± 7.35	18.09 ± 4.71
CAT	58.32 ± 10.29	54.76 ± 10.56	154.39 ± 51.32	145.57 ± 41.21
GNN	42.08 ± 7.82	39.11 ± 7.24	30.73 ± 8.07	27.61 ± 7.70
GraphITE	40.26 ± 7.94	37.99 ± 7.80	30.41 ± 8.74	27.51 ± 8.69
GIN	$\mathbf{27.47 \pm 6.07}$	$\mathbf{26.01 \pm 6.06}$	$\mathbf{11.13 \pm 3.41}$	$\mathbf{8.23 \pm 1.43}$

Table 6: Error of CATE estimation for all methods, measured by WPEHE@$1-10$. Results are averaged over 10 trials, $\pm$ denotes std. error.

E.2 Comparison of Robustness to different Bias Strengths κ

We present additional WPEHE@K results for the experiments in Section 6.3 over increasing bias strength κ and varying K.



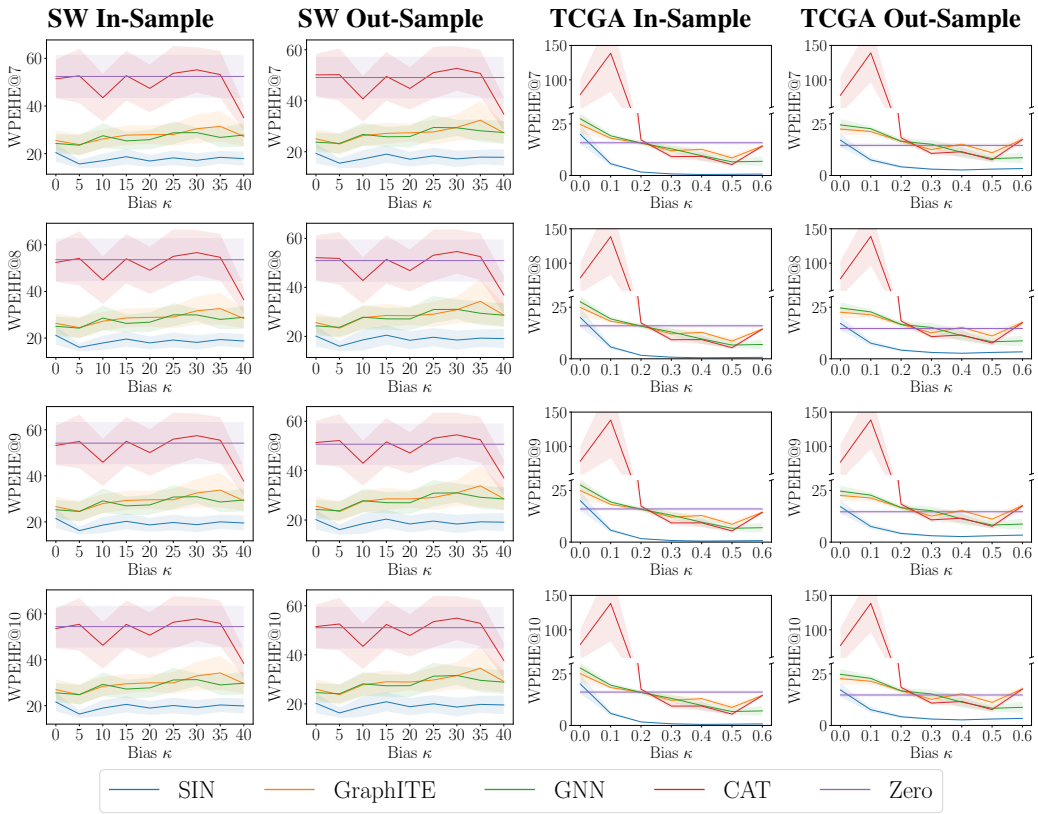


Figure 5: WPEHE@K over increasing bias strength κ and varying K.

F Quasi-oracle rates for generalized R-Learner

The goal of this section is to establish error bounds for learning conditional average treatment effects (CATEs) when treatments are *continuous*. To do so, we will assume that the response function, $\mathbb{E}[Y|\mathbf{X}, \mathbf{T}]$ can be written as follows,

$$\mathbb{E}\left[Y \mid \mathbf{X}, \mathbf{T}\right] = \boldsymbol{\alpha}\left(\mathbf{X}\right)^{\top} \Theta^{*} \boldsymbol{\beta}\left(\mathbf{T}\right), \tag{33}$$

where $\boldsymbol{\alpha}(\mathbf{X}) \in \mathbb{R}^{d_{\mathbf{X}}}, \boldsymbol{\beta}(\mathbf{T}) \in \mathbb{R}^{d_{\mathbf{T}}}$ are fixed, known basis functions[6] (where $d_{\mathbf{X}}, d_{\mathbf{T}} < \infty$) and $\Theta^{*} \in \mathbb{R}^{d_{\mathbf{X}} \times d_{\mathbf{T}}}$ is unknown. We will show that we can learn Θ^{*} using the generalized Robinson decomposition in eq. (10) (i.e., the minimization is now over Θ) with the same error rate as if we had known the true *oracle* nuisance functions m^{*} and e^{p}, provided our estimates of m^{*} and e^{P} converge to the ground truths at $O(n^{-1/4})$ rate.

The reason we consider the above fixed basis setting instead of the more generic setup in the paper is because there are many things that make the analysis of a more general setup difficult:

- There is a non-trivial dependence between estimators $m(\cdot), e(\cdot), g(\cdot), h(\cdot)$ created by fitting using the entire dataset (as opposed to using cross-fitting).
- Representation learning of the features typically involves non-convex loss functions; the convergence analysis of such is largely still an untackled question.
- In the infinite-basis setting the problem becomes ill-posed (our current work provides insight into fixing this, in particular in Lemma 8).

Addressing these issues is an interesting area of future work. Meanwhile, in this work, we focus on the scenario where the features (i.e. basis functions) are fixed. We first sketch our result without technical jargon as follows.

[6] In deep learning jargon, each dimension of the basis functions, α_i, β_j, is simply called a *feature*.

Theorem (Sketch). *Write $m(\mathbf{x}) := \mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}\right]$ and $e^P(\mathbf{x}) := \mathbb{E}\left[\boldsymbol{\beta}(\mathbf{T}) \mid \mathbf{X} = \mathbf{x}\right]$. When the ground truths m and e^P are unavailable, we can still estimate $\mathbb{E}\left[Y \mid \mathbf{X}, \mathbf{T}\right]$ almost with rate $O\left(n^{-1/2}\right)$ using only estimates of m and e^P, provided the estimates themselves converge at rate $O\left(n^{-1/4}\right)$.*

F.1 Preliminaries

To specify the above formally, we follow e.g. [53] to construct an RKHS for the hypothesis space of the response function f as follows. Let $\mathcal{X}$ and $\mathcal{T}$ be compact metric spaces, endowed with finite Borel measures $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{T}$. Let $\{\alpha_i\}_{i=1}^{d_\alpha} \subset \mathcal{L}_2(\mathcal{X}, \mathcal{P}_\mathcal{X})$ and $\{\beta_i\}_{i=1}^{d_\beta} \subset \mathcal{L}_2(\mathcal{T}, \mathcal{P}_\mathcal{T})$ denote subsets of orthonormal functions in $\mathcal{L}_2(\mathcal{X}, \mathcal{P}_\mathcal{X})$ and $\mathcal{L}_2(\mathcal{T}, \mathcal{P}_\mathcal{T})$ which are feature maps for $\mathbf{X}$ and $\mathbf{T}$, respectively. Write $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{d_\alpha}, \mathbb{R}^{d_\beta}$ as the vectors of features on $\mathbf{X}$ and $\mathbf{T}$, with $\boldsymbol{\alpha}_i(\mathbf{x}) := \alpha_i(\mathbf{x})$ and $\boldsymbol{\beta}_j(\mathbf{t}) := \beta_j(\mathbf{t})$. Then define $k_\mathbf{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as $k_\mathbf{X}(\mathbf{x}_1, \mathbf{x}_2) = \langle \boldsymbol{\alpha}(\mathbf{x}_1), \boldsymbol{\alpha}(\mathbf{x}_2)\rangle_2$ where $\langle \cdot, \cdot \rangle_2$ is the standard Euclidean dot product in $\mathbb{R}^d$, and define similarly $k_\mathbf{T} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ as $k_\mathbf{T}(\mathbf{t}_1, \mathbf{t}_2) = \langle \boldsymbol{\beta}(\mathbf{t}_1), \boldsymbol{\beta}(\mathbf{t}_2)\rangle_2$. Then clearly $k_\mathbf{X}$ and $k_\mathbf{T}$ are positive definite functions and by Moore-Aronsajn [53, Section 4] there exist unique RKHSes $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{T}$ with kernels $k_\mathbf{X}$ and $k_\mathbf{T}$.

For the readers familiar with [42], we can connect the setup to that of [42] as follows: following e.g. [36], an element g in $\mathcal{H}_\mathcal{X}$ can be represented by $g(\mathbf{x}) = \langle \theta, \boldsymbol{\alpha}(\mathbf{x})\rangle_2 = \langle g, \boldsymbol{\alpha}(\mathbf{x})\rangle_{\mathcal{H}_\mathcal{X}}$. Following [53], we can define an integral operator based on the kernel $k_\mathbf{X}$:

$$S_{k_\mathbf{X}} : \mathcal{L}_2(\mathcal{X}; \mathcal{P}_\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \quad \text{where } \mathcal{C}(\mathcal{X}) \text{ are the continuous functions on } \mathcal{X}. \tag{34}$$

$$(S_{k_\mathbf{X}} f)(\mathbf{X}) = \int k_\mathbf{X}(\mathbf{x}_1, \mathbf{x}_2) f(\mathbf{x}_2) d\mathcal{P}_\mathcal{X}(\mathbf{x}_2), \qquad f \in \mathcal{L}_2(\mathcal{X}; \mathcal{P}_\mathcal{X}) \tag{35}$$

$$T_{k_\mathbf{X}} = I_{k_\mathbf{X}} \circ S_{k_\mathbf{X}} \tag{36}$$

$$\text{with the inclusion } I_{k_\mathbf{X}} : \mathcal{C}(\mathcal{X}) \hookrightarrow \mathcal{L}_2(\mathcal{X}; \mathcal{P}_\mathcal{X}) \tag{37}$$

Clearly the eigenfunctions of $T_{k_\mathbf{X}}$ are the orthonormal functions $\{\alpha_i\}_{i=1}^{d_\alpha}$ and the non-zero eigenvalues are $\{\sigma_i = 1\}_{i=1}^{d_\alpha}$.

$\mathcal{H}_\mathcal{T}$ can be dealt with similarly to $\mathcal{H}_\mathcal{X}$. Since $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ is isometrically isomorphic to $\mathcal{H}_\mathcal{X} \times \mathcal{H}_\mathcal{T}$, we can identify the basis functions on $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ as $\{\alpha_i\beta_j\}_{i,j=1}^{d_\alpha,d_\beta}$, the eigenvalues as $\{\sigma_{ij} = 1\}_{i,j=1}^{d_\alpha,d_\beta}$, and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{X}\times\mathcal{T}}} = \langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{X}} \langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{T}}$. By construction, the RKHS norm and the L_2 norm of $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ are both equal to the matrix $2-$norm of the function representer, that is, for $f \in \mathcal{H}_{\mathcal{X}\times\mathcal{T}}, f(\mathbf{x}, \mathbf{t}) = \langle \boldsymbol{\Theta}, \boldsymbol{\alpha}(\mathbf{x}) \otimes \boldsymbol{\beta}(\mathbf{t})\rangle_2$,

$$\|f\|_{\mathcal{H}_{\mathcal{X}\times\mathcal{T}}} = \|f\|_{L_2} = \|\boldsymbol{\Theta}\|_2 \tag{38}$$

Trivially, for all $0 < p < 1$, the eigenvalues σ_{ij} satisfy $G = \sup_{i,j\geq 1}(i + d_\mathbf{X}(j-1))^{1/p}\sigma_{ij}$ for some constant $G < \infty$, which was posed as an assumption in [42].

Remark 1. *We did not need to require $\mathcal{X}$ and $\mathcal{T}$ as compact metric spaces. Requiring them to be measurable spaces on which we can define $\mathcal{L}_2$ functions should be enough. But compact metric spaces also include most spaces of practical concern, including graph spaces, so we choose it since it satisfies the conditions of Mercer's theorem.*

F.2 Problem set-up

We assume that the true response function lie in $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$:

Assumption 4. *The true response 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 $\boldsymbol{\Theta}^*$.*

First we write down the population and empirical loss functions we consider. In order to assert that every element of $\mathcal{H}_{\mathcal{X}\times\mathcal{T}}$ can be uniquely represented by some $\boldsymbol{\Theta}$, we use $f_\boldsymbol{\Theta}$ to denote $f_\boldsymbol{\Theta} := \boldsymbol{\alpha}(\mathbf{x})^T\boldsymbol{\Theta}\boldsymbol{\beta}(\mathbf{t}) \in \mathcal{H}_{\mathcal{X}\times\mathcal{T}}$.

The expected loss of $f_\boldsymbol{\Theta}$ is defined by:

$$L(f_\boldsymbol{\Theta}) = L(\boldsymbol{\Theta}) = \mathbb{E}\left[\left\{(Y - m^*(\mathbf{X})) - \boldsymbol{\alpha}(\mathbf{X})^T\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}) - e^P(\mathbf{X}))\right\}^2\right] \tag{39}$$

The oracle (empirical) loss is defined by:

$$\tilde{L}_n(f_{\boldsymbol{\Theta}}) = \tilde{L}_n(\boldsymbol{\Theta}) = \sum_{l=1}^{n} \left[\left\{ (Y - m^*(\mathbf{X}_l)) - \boldsymbol{\alpha}(\mathbf{X}_l)^T \boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^P(\mathbf{X}_l)) \right\}^2 \right] \tag{40}$$

The feasible (empirical) loss is defined by:

$$\hat{L}_n(f_{\boldsymbol{\Theta}}) = \hat{L}_n(\boldsymbol{\Theta}) = \sum_{l=1}^{n} \left[\left\{ (Y - \hat{m}(\mathbf{X}_l)) - \boldsymbol{\alpha}(\mathbf{X}_l)^T \boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - \hat{e}^P(\mathbf{X}_l)) \right\}^2 \right] \tag{41}$$

Note that we use $L(f_{\boldsymbol{\Theta}})$ and $L(\boldsymbol{\Theta})$ interchangeably due to the bijection between $\boldsymbol{\Theta} \in \mathbb{R}^{d_{\mathbf{X}} \times d_{\mathbf{T}}}$ and $\mathcal{H}_{\mathcal{X} \times \mathcal{T}}$.

The corresponding regret functions are defined by

$$R(\boldsymbol{\Theta}) = L(\boldsymbol{\Theta}) - L(\boldsymbol{\Theta}^*) \tag{42}$$

$$\tilde{R}_n(\boldsymbol{\Theta}) = \tilde{L}_n(\boldsymbol{\Theta}) - \tilde{L}_n(\boldsymbol{\Theta}^*) \tag{43}$$

$$\hat{R}_n(\boldsymbol{\Theta}) = \hat{L}_n(\boldsymbol{\Theta}) - \hat{L}_n(\boldsymbol{\Theta}^*) \tag{44}$$

We now formally state the assumptions we need to derive the result in Theorem 2.

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

Assumption 6 (Boundedness)**.** *Without loss of generality, we assume that for all $\mathbf{X} \in \mathcal{X}, \mathbf{T} \in \mathcal{T}$,* $\sup_i \|\alpha_i(\mathbf{X})\|_\infty, \sup_j \|\beta_j(\mathbf{T})\|_\infty \leq A < \infty$. *We also assume that the outcome Y are almost surely bounded, i.e.* $\mathbb{P}(|Y| < B < \infty) = 1$.

For clarity, we list all notations we use here.

Notation.

- $\mathcal{H}$: A Product Reproducing Kernel Hilbert Space with finite number of basis functions, with $\boldsymbol{\alpha}$ the features of X and $\boldsymbol{\beta}$ the features of T.
- Θ: The matrix of coefficients for a given function in $\mathcal{H}$.
- f_Θ: $f_\Theta(X, T) := \boldsymbol{\alpha}(X)^\top \Theta \boldsymbol{\beta}(T)$.
- $\mathcal{H}_c$: The subset of $\mathcal{H}$ which is the ball of radius c.
- Θ_c: f_{Θ_c} is a minimiser of the loss in $\mathcal{H}_c$.
- $R(f_\Theta)$: $L(f_\Theta) - L(f^*)$.
- $R(f_\Theta; c)$: $L(f_\Theta) - L(f_{\Theta_c})$

Convention. Throughout, we will use capital letters $A, B, C, ...$, possibly with subscripts and superscripts, e.g. $A_1, B^{(2)}$, etc. to denote constants. We may overload notation and use the same letter to denote different constants.

F.3 Proof strategy

Here we lay forward the detailed proof for the quasi-oracle convergence rate for a featurized continuous heterogeneous treatment effect estimation algorithm with Robinson decomposition. Our proof extends the structure of Nie & Wager [42]. To make the proof self-contained while simultaneously highlighting the differences with Nie & Wager [42], we present a complete version of the proof, where we will pause to describe any difference and its significance where it appears.

The high-level idea of showing 'quasi-oracle' error rate is as follows. First, we show that both the feasible loss and the oracle loss satisfy the same (quasi-)isomorphism with the true loss, where the tightness of the quasi-isomorphism increases as sample size increases. The quasi-isomorphism with the true loss then leads us to bound the feasible and oracle losses by the same quantity, which decreases to 0 as sample size grows indefinitely. To show the (quasi-)isomorphism for the oracle

learner can be done by leveraging on the standard least-squares regression ideas [36]; to achieve the same for the feasible learner relies on the fact that the feasible loss differs from the oracle loss by only a small amount relative to the true loss, which constitutes the bulk of the proof.

We start with stating the formal lemma which connects quasi-isomorphism with loss bounds.

F.4 From quasi-isomorphism to regret bound

Definition 3 (loss function)**.** *A function is a **loss function** if it maps from a hypothesis class $\mathcal{H}$, to the real numbers $\mathbb{R}$.*

Lemma 4. *Let $\check{L}(f_\Theta \in \mathcal{H}_c)$ be a loss function, and $\check{R}(f_\Theta; c) = \check{L}(f_\Theta) - \check{L}(f_{\Theta_c})$ be the associated c-regret. Suppose $\rho(r)$ is a positive, continuous, increasing function. If, $\forall\ 1 \leq c \leq C$ and some $k > 1$, the following inequality holds for all $f_\Theta \in \mathcal{H}_c$:*

$$\frac{1}{k}\check{R}(f_\Theta; c) - \rho(c) \leq R(f_\Theta; c) \leq k\check{R}(f_\Theta; c) + \rho(c) \tag{45}$$

Then, writing $\kappa_1 = 2k + \frac{1}{k}$ and $\kappa_2 = 2k^2 + 3$, any solution to the regularized minimization problem with $\Lambda(c) \geq \rho(c)$,

$$f_{\check{\boldsymbol{\Theta}}} \in \underset{f_\Theta \in \mathcal{H}_C}{\arg\min}\{\check{L}(f_\Theta) + \kappa_1 \Lambda(f_\Theta)_{\mathcal{H}}\} \tag{46}$$

also satisfied the following risk bound:

$$L(f_{\check{\boldsymbol{\Theta}}}) \leq \inf_{f_\Theta \in \mathcal{H}_C}\{L(f_\Theta) + \kappa_2 \Lambda(f_\Theta)_{\mathcal{H}} \tag{47}$$

Proof. Notice that $\{\mathcal{H}_c; c \geq 1\}$ is an ordered set. Thus the same argument as [42] applies. □

Lemma 4 tells us that if we have a quasi-isomorphism of the regrets in the form of 45, we immediately can bound the expected risk of the (regularized) minimizer of the corresponding loss, $\check{L}$ as in 47.

F.5 A concrete instance of $\rho(c)$ satisfying 45

By setting $\check{R}$ to $\tilde{R}_n$, 4 gives us a way to bound the oracle regret, but we still need a concrete formulation of $\rho(c)$ to derive the oracle convergence rate. To this end, we may use the result of Mendelson & Neeman [36], but first we must show that their results can be applied to our setting.

Mendelson & Neeman [36] consider the optimization over a space of RKHS functions with the least-squares loss. Our oracle case can be thought of in the same way as follows: since m^* is an oracle quantity, $Y - m^*(\mathbf{X})$ can be thought of as the labels, the space $\overline{\mathcal{H}} = \{f_{\boldsymbol{\Theta}} : \mathcal{X} \times \mathcal{T} \to R;$ for some $\boldsymbol{\Theta} \in \mathbb{R}^{d_{\mathbf{X}} \times d_{\mathbf{T}}}, f_{\boldsymbol{\Theta}}(\mathbf{x}, \mathbf{t}) = \boldsymbol{\alpha}(\mathbf{x})^\top \boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{t}) - e^P(\mathbf{x}))\}$ is an RKHS with features $\boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^P(\mathbf{X}))$. Thus, our setting can be thought of as a least-squares optimization over the RKHS $\overline{\mathcal{H}}$ and the results from [36] applies. To use the results of [36], we still need the following technical result which decomposes $\overline{\mathcal{H}}$ into an *ordered, parameterized hierarchy*.

Definition 5 (Ordered, parameterized hierarchy)**.** *As defined in [36], let $\mathcal{F}$ be a class of functions and suppose that there is a collection of subsets $\{\mathcal{F}_r; r \geq 1\}$ with the following properties:*

1. *$\{\mathcal{F}_r : r \geq 1\}$ is monotone (i.e. whenever $r \leq s, \mathcal{F}_r \subseteq \mathcal{F}_s$);*
2. *for every $r \geq 1$, there exists a unique element $f_r^* \in \mathcal{F}_r$ such that $L(f_r^*) = \inf_{f \in \mathcal{F}_r} L(f)$;*
3. *the map $r \to L(f_r^*)$ is continuous;*
4. *for every $r_0 \geq 1$, $\bigcap_{r \leq r_0} \mathcal{F}_r = \mathcal{F}_{r_0}$;*
5. *$\bigcup_{r \leq 1} \mathcal{F}_r = \mathcal{F}$.*

*Given a class of functions $\mathcal{F}$, we say that $\{\mathcal{F}_r; r \geq 1\}$ is an **ordered, parameterized hierarchy of** $\mathcal{F}$ if the above conditions 1-5 are satisfied.*

Lemma 6. *Define*

$$\overline{\mathcal{H}_c} := \{f_\Theta : \mathcal{X} \times \mathcal{T} \to \mathbb{R} : \exists \boldsymbol{\Theta}, \|\boldsymbol{\Theta}\|_2 \leq c, \tag{48}$$

$$\text{s.t.} \quad f_\Theta(\mathbf{X}, \mathbf{T}) = \boldsymbol{\alpha}(\mathbf{X})^\top \boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\}, \tag{49}$$

then $\left\{\overline{\mathcal{H}_c}\right\}_{1 \leq c \leq C}$ *is an ordered parameterized hierarchy.*

Proof. The first, fourth and fifth properties follow immediately. $\overline{\mathcal{H}_c}$ is clearly convex. It is compact because every sequence $\{f_{\boldsymbol{\Theta}_i}\}_i \subset \overline{\mathcal{H}_c}$ is induced by $\{\boldsymbol{\Theta}_i\}_i \subset \mathbb{R}^n$, $\|\boldsymbol{\Theta}_i\|_2 \leq c$, and by Bolzano-Weierstrass theorem in $\mathbb{R}^n$, every bounded sequence has a convergent subsequence $\{\boldsymbol{\Theta}_k\}_k \subset \{\boldsymbol{\Theta}_i\}_i$ (w.r.t. the Euclidean norm). Thus pick the N such that for all $k \geq N$ where $\|\boldsymbol{\Theta}_k - \boldsymbol{\Theta}_N\|_2 \leq \epsilon$, and then

$$\left\|f_{\boldsymbol{\Theta}_k} - f_{\boldsymbol{\Theta}_N}\right\|_{L_2(P(\mathcal{X},\mathcal{T}))} = \left\|f_{\boldsymbol{\Theta}_k - \boldsymbol{\Theta}_N}\right\|_{L_2(P(\mathcal{X},\mathcal{T}))} = \mathbb{E}\left[\langle \boldsymbol{\Theta}_k - \boldsymbol{\Theta}_N, \boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\rangle^2\right]^{1/2} \tag{50}$$

$$\leq \mathbb{E}\left[\|\boldsymbol{\Theta}_k - \boldsymbol{\Theta}_N\|_2 \left\|\boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\right\|_2\right]^{1/2} \tag{51}$$

$$\leq \epsilon \mathbb{E}\left[\left\|\boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\right\|_2\right]^{1/2} \leq \epsilon B, \tag{52}$$

where $\left\|\boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\right\|_2 \leq B$ by Assumption 6 for some constant B. The second property now follows from the fact that $\overline{\mathcal{H}_c}$ is convex and compact. The third property follows by the same argument as [36]. □

Mendelson & Neeman [36] thus provides a formulation of ρ which, with some constant $U(\epsilon)$, for large enough n and probability at least $1 - \epsilon$, satisfies 120 for the oracle loss function $\tilde{R}_n$ with $k = 2$:

$$\rho_n(c) = U(\epsilon)\left\{1 + \log(n) + \log\left(\log(c + e)\right)\right\}\left(\frac{(c+1)^p \log(n)}{\sqrt{n}}\right)^{2/(1+p)} \tag{53}$$

Thus, we may now realize the convergence rate for the oracle learner as follows.

F.6 Oracle convergence rate.

With 53, Lemma 4 immediately implies that penalized regression over $\mathcal{H}_C$ with the oracle loss function $\tilde{L}_n(\cdot)$ and regularizer $\kappa_1 \rho_n(c)$ satisfies the bound below with high probability:

$$R(\tilde{\boldsymbol{\Theta}}_n) = L(\tilde{\boldsymbol{\Theta}}_n) - L(\boldsymbol{\Theta}^*) \leq \inf_{\boldsymbol{\Theta} \in \mathcal{H}_C} \{L((\boldsymbol{\Theta}) + \kappa_2 \rho_n(\|\boldsymbol{\Theta}\|_{\mathcal{H}})\} - L(\boldsymbol{\Theta}^*) \tag{54}$$

Furthermore, Corollary 2.7 in [36] gives that for any $1 < c < C$,

$$\inf_{\boldsymbol{\Theta} \in \mathcal{H}_C} \{L(\boldsymbol{\Theta}) + \kappa_2 \rho_n(\|\boldsymbol{\Theta}\|_{\mathcal{H}})\} \leq L(\boldsymbol{\Theta}^*) + \{L(\boldsymbol{\Theta}_c^*) - L(\boldsymbol{\Theta}^*)\} + \kappa_2 \rho_n(c) \tag{55}$$

Finally, note that for large enough c,

$$\left\{L(\boldsymbol{\Theta}_c^*) - L(\boldsymbol{\Theta}^*)\right\} = 0, \tag{56}$$

so the error is dominated by $\rho_n(c)$, at

$$R\left(\tilde{\boldsymbol{\Theta}}_n\right) = \mathcal{O}\left(\left(\log(n)\right)^{\frac{3+p}{1+p}} n^{-\frac{1}{1+p}}\right) = \tilde{\mathcal{O}}(n^{-\frac{1}{1+p}}), \tag{57}$$

where $\tilde{\mathcal{O}}$ notation ignores the logarithmic factors.

F.7 Bridging $\hat{R}_n$ and $\tilde{R}_n$

Now that we have the oracle convergence rate, we show a bridging result which will let us conclude that 45 holds for $\hat{R}_n$ as well, and thus the oracle rate also holds for $\hat{R}_n$.

To yield that bridging result, we first need to leverage the assumption of overlap to relate the L_2 difference between $f_{\boldsymbol{\Theta}}$ and f_{Θ_c}, i.e. $\mathbb{E}\left[\left(f_{\boldsymbol{\Theta}}(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T})\right)^2\right]$, with the $c-$regret $R(\boldsymbol{\Theta};c)$. We first show that the L_2 difference is always upper bounded by the regret up to a constant.

Lemma 7. *$\exists \epsilon > 0$ s.t. for all $f \in \mathcal{H}_c$, $\mathbb{E}_\alpha[\langle f, \alpha\rangle^2] \geq \epsilon \|f\|_{L_2}$ where α is a r.v. taking values in $\mathcal{H}_c$ and the support of α is of Lebesgue-measure non-zero in $\mathcal{H}_c$.*

Proof. Let $S = \{f \in \mathcal{H}_c : \|f\|_{\mathcal{H}_c} = 1\}$, and define $g : S \to \mathbb{R}^+$ as $g(f) = \mathbb{E}_{\boldsymbol{\alpha}}[\langle f, \alpha\rangle^2]$. By Jensen's inequality, $\mathbb{E}_{\boldsymbol{\alpha}}[\langle f, \alpha\rangle^2] \geq 0$ since $\langle f, \cdot\rangle^2 : \alpha \mapsto \langle f, \alpha\rangle^2$ is a convex function in α. Moreover, whenever $supp[\mathcal{P}_{\boldsymbol{\alpha}}]$ is Lebesgue-measure non-zero in $\mathcal{H}_c$, $\langle f, \cdot\rangle^2$ is non-linear on $supp[\mathcal{P}_{\boldsymbol{\alpha}}]$, so the inequality is strict:

$$\mathbb{E}_{\boldsymbol{\alpha}}[\langle f, \alpha\rangle^2] > 0. \tag{58}$$

Now since $\mathcal{H}_c$ is finite-dimensional, S is compact. Since g is continuous in f, and the continuous image of a compact set is compact, we have that $g(S)$ is compact, and therefore closed.

Note, at this point, that $g(S)$ is the set of values achieved by $\mathbb{E}_{\boldsymbol{\alpha}}[\langle f, \alpha\rangle^2]$ at various values of f. By equation 58, $g(S) \not\ni 0$. Since $g(S)$ is compact, its complement thus contains 0. Moreover, since $\mathbb{R}^+ \setminus g(S) \ni 0$, $\exists$ a ball around 0 of radius $\tilde{\epsilon} > 0$ s.t. $[0, \tilde{\epsilon}) \subset \mathbb{R}^+ \setminus g(S)$. Therefore, $g(S) \subset \mathbb{R}^+$ is lower bounded by $\tilde{\epsilon} > 0$.

Therefore,

$$\forall f \in \mathcal{H}_c, \mathbb{E}_{\boldsymbol{\alpha}}\left[\langle f, \alpha\rangle^2\right] = \|f\|^2_{\mathcal{H}_c}\, \mathbb{E}_{\boldsymbol{\alpha}}\left[\left\langle \frac{f}{\|f\|^2_{\mathcal{H}_c}}, \alpha\right\rangle^2\right] \geq \epsilon \|f\|^2_{\mathcal{H}_c} = \epsilon \|f\|^2_{L_2}, \tag{59}$$

for some $\epsilon > 0$. The last inequality is due to 38. □

Lemma 8 (Usage of the overlap condition in the multiple treatment setting)**.** *Under Assumption 5, i.e. we have overlap on the features, that is $supp[\mathcal{P}_{\boldsymbol{\alpha}(\mathcal{X})\times\boldsymbol{\beta}(\mathcal{T})}] = \boldsymbol{\alpha}(\mathcal{X}) \times \boldsymbol{\beta}(\mathcal{T})$, then $\exists A \in \mathbb{R}$ s.t.*

$$\mathbb{E}[\left(f_\Theta(X,T) - f_{\Theta_c}(X,T)\right)^2] < AR(\Theta;c) \tag{60}$$

Proof. Within $\mathcal{H}_c$, we seek to upper bound excess L_2 risk of f_Θ by its c-regret $R(\Theta;c)$; $R(\Theta;c) = L(\Theta) - L(\Theta_c)$.

First we write down the expected loss functional again:

$$L(\boldsymbol{\Theta}) = \mathbb{E}[(\{Y - m^*(\mathbf{X})\} - \{f_\Theta(\mathbf{X}.\mathbf{T}) - \mathbb{E}[f_\Theta(\mathbf{X},\mathbf{T})|\mathbf{X}]\})^2] \tag{61}$$

$$= \mathbb{E}[\mathbb{V}\{Y - m^*(\mathbf{X})|\mathbf{X},\mathbf{T}\}] + \mathbb{E}[\{(f^*(\mathbf{X},\mathbf{T}) - f_\Theta(\mathbf{X},\mathbf{T})) - \mathbb{E}[f^*(\mathbf{X},\mathbf{T}) - f_\Theta(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2] \tag{62}$$

Thus the regret of $\boldsymbol{\Theta}$, which is defined as $L(\boldsymbol{\Theta}) - L(f^*)$, is:

$$R(\boldsymbol{\Theta}) = \mathbb{E}\left[\left\{(f^*(\mathbf{X},\mathbf{T}) - f_\Theta(\mathbf{X},\mathbf{T}) - \mathbb{E}[f^*(\mathbf{X},\mathbf{T}) - f_\Theta(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\right\}^2\right] \tag{63}$$

$$\begin{aligned} &= \mathbb{E}[\{(f_\Theta(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T})) - \mathbb{E}[f_\Theta(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}] \\ &\quad + (f_{\Theta_c}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T})) - \mathbb{E}[f_{\Theta_c}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2] \end{aligned} \tag{64}$$

$$\begin{aligned} &= \mathbb{E}[\{f_\Theta(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T})) - \mathbb{E}[f_\Theta(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2] \\ &\quad + \mathbb{E}[\{f_{\Theta_c}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) - \mathbb{E}[f_{\Theta_c}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2] \\ &\quad + 2\mathbb{E}[\{(f_\Theta(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T})) - \mathbb{E}[f_\Theta(\mathbf{X},\mathbf{T}) - f_{\Theta_c}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\} \\ &\quad \cdot \{(f_{\Theta_c}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T})) - \mathbb{E}[f_{\Theta_c}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}] \end{aligned} \tag{65}$$

Note that, by definition the c-regret of $\boldsymbol{\Theta}$ is just the difference between the regret of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_c$. And the regret of $\boldsymbol{\Theta}_c$ is the second term in equation 65. Thus, the c-regret of $\boldsymbol{\Theta}$ is the first and third term of equation 65.

Now, note that the third term is non-negative because $\mathcal{H}_c$ is convex. To see this, note that it is equal to

$$\frac{\partial}{\partial \epsilon} R\left(\mathbf{\Theta}_c + \epsilon\left(\mathbf{\Theta} - \mathbf{\Theta}_c\right)\right)|_{\epsilon=0}, \tag{66}$$

which must be non-negative for any $\mathbf{\Theta} \in \mathcal{H}_c$ since otherwise there will be another point in $\mathcal{H}_c$ which has a smaller regret than $\mathbf{\Theta}_c$.

Therefore,

$$R(\mathbf{\Theta}; c) \geq \mathbb{E}\left[\left\{f_\Theta(\mathbf{X}, \mathbf{T}) - f_{\Theta_c}(\mathbf{X}, \mathbf{T})) - \mathbb{E}\left[f_\Theta(\mathbf{X}, \mathbf{T}) - f_{\Theta_c}(\mathbf{X}, \mathbf{T}) \mid \mathbf{X}\right]\right\}^2\right] \tag{67}$$

$$= \mathbb{E}\left[\left\{\boldsymbol{\alpha}(\mathbf{X})^\top(\mathbf{\Theta} - \mathbf{\Theta}_c)(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\right\}^2\right] \tag{68}$$

$$= \mathbb{E}\left[\left\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\right\rangle^2\right] \tag{69}$$

Now, we would like to show that $\mathbb{E}\left[\left\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \boldsymbol{\alpha}(\mathbf{X}) \otimes \left(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})\right)\right\rangle^2\right]$ is bounded below by the norm of $\mathbf{\Theta} - \mathbf{\Theta}_c$ up to some multiplicative constant. We do so using Lemma 7. Under the context of Lemma 7, set $f := \mathbf{\Theta} - \mathbf{\Theta}_c$, and $\alpha := \boldsymbol{\alpha}(\mathbf{X}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))$. To check that the support of α is not of measure 0, we first note that the support of $\boldsymbol{\alpha}(\mathbf{X})$ is not measure 0 by assumption; secondly, the support of $\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})$ is not measure 0 provided that $P(\boldsymbol{\beta}(T) \mid X)$ is a positive measure for any X. Then by Lemma 7, we have that $\exists \epsilon > 0$

$$R(\mathbf{\Theta}; c) \geq \epsilon \|f_{\mathbf{\Theta}} - f_{\mathbf{\Theta}_c}\|_{L_2} \tag{70}$$

□

Immediately after Lemma 8, we derive a bound on the infinity norm using the regret function which we will repeatedly use later.

Corollary 9. *Following from 38 and Lemma 8,*

$$\|\mathbf{\Theta} - \mathbf{\Theta}_c\|_\infty \leq const(p)\|f_{\mathbf{\Theta}} - f_{\mathbf{\Theta}_c}\|_{\mathcal{H}}^p \|f_{\mathbf{\Theta}} - f_{\mathbf{\Theta}_c}\|_{L_2}^{1-p} \leq const(p) c^p R(\mathbf{\Theta}; c)^{\frac{1-p}{2}} \tag{71}$$

where we note that the second inequality follows from combining Lemma 8 with the fact that for $f_{\mathbf{\Theta}} \in \mathcal{H}_c, \|f_{\mathbf{\Theta}} - f_{\mathbf{\Theta}_c}\| \leq 2c$ by the triangle inequality.

Proof. Immediate from 38 and Lemma 8. □

Using Lemma 8, we can further show that the L_2 difference between two constrained optima only depends on the L_2 norm of the one with the weaker constraint.

Corollary 10. *Suppose we have overlap, i.e. Assumption 5. Then with a positive constant $const. > 0$, the following holds for $1 < c < c'$.*

$$\|f_{\Theta_c} - f_{\mathbf{\Theta}_{c'}}\|_{L_2} \leq const. \|f_{\mathbf{\Theta}_{c'}}\|_{L_2} \tag{72}$$

Proof. We have shown that

$$R(\mathbf{\Theta}; c) \geq \epsilon \|f_{\mathbf{\Theta}} - f_{\mathbf{\Theta}_c}\|_{L_2}^2 \tag{73}$$

Then following [42], we check that

$$\|\mathbf{\Theta}_c - \frac{c}{c'}\mathbf{\Theta}_{c'}\|_{L_2}^2 \leq \epsilon R(\frac{c}{c'}\mathbf{\Theta}_{c'}; c) \tag{74}$$

$$= \epsilon\left(L(\frac{c}{c'}\mathbf{\Theta}_{c'}) - L(\mathbf{\Theta}_c)\right) \tag{75}$$

$$\leq \epsilon\left(L(\frac{c}{c'}\mathbf{\Theta}_{c'}) - L(\mathbf{\Theta}_{c'})\right) \tag{76}$$

$$= \epsilon\left(R(\frac{c}{c'}\mathbf{\Theta}_{c'}) - R(\mathbf{\Theta}_{c'})\right) \tag{77}$$

To bound $R\left(\frac{c}{c'}\mathbf{\Theta}_{c'}\right) - R\left(\mathbf{\Theta}_{c'}\right)$, note

$$
\begin{aligned}
R(\mathbf{\Theta}) = \mathbb{E}\Big[&\{(f_{\Theta}(\mathbf{X},\mathbf{T}) - f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T})) - \mathbb{E}[f_{\Theta}(\mathbf{X},\mathbf{T}) - f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2\Big] \\
&+ \mathbb{E}\Big[\{f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) - \mathbb{E}[f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2\Big] \\
&+ 2\mathbb{E}\Bigg[\Big\{(f_{\Theta}(\mathbf{X},\mathbf{T}) - f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T})) - \mathbb{E}\left[f_{\Theta}(\mathbf{X},\mathbf{T}) - f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}\right]\Big\} \\
&\quad \cdot \Big\{(f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T})) - \mathbb{E}\left[f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}\right]\Big\}\Bigg]
\end{aligned} \tag{78}
$$

so $R(\mathbf{\Theta}_{c'})$ is just the second term of equation 78, which we drop when considering $R(\frac{c}{c'}\mathbf{\Theta}_{c'}) - R(\mathbf{\Theta}_{c'})$

$$
\begin{aligned}
R\left(\frac{c}{c'}\mathbf{\Theta}_{c'}\right) - R(\mathbf{\Theta}_{c'}) = \mathbb{E}\Big[&\{(\frac{c}{c'}-1)f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - \mathbb{E}[(\frac{c}{c'}-1)f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}]\}^2\Big] \\
&+ 2\mathbb{E}\Bigg[\Big\{(\frac{c}{c'}-1)f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - \mathbb{E}\left[(\frac{c}{c'}-1)f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) \mid \mathbf{X}\right]\Big\} \\
&\quad \cdot \Big\{(f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T})) - \mathbb{E}\left[f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}\right]\Big\}\Bigg]
\end{aligned} \tag{79}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left\{\boldsymbol{\alpha}(\mathbf{X})^\top\left(\frac{c}{c'}-1\right)\mathbf{\Theta}_{c'}\left(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})\right)\right\}^2\right] \\
&\quad + 2\mathbb{E}\Bigg[\Bigg\{\boldsymbol{\alpha}(\mathbf{X})^\top\left(\frac{c}{c'}-1\right)\mathbf{\Theta}_{c'}\left(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})\right)\Bigg\} \\
&\quad\quad \cdot \Big\{(f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T})) - \mathbb{E}\left[f_{\mathbf{\Theta}_{c'}}(\mathbf{X},\mathbf{T}) - f^*(\mathbf{X},\mathbf{T}) \mid \mathbf{X}\right]\Big\}\Bigg]
\end{aligned} \tag{80}
$$

Denote the two terms E_1 and E_2. By the same argument as Lemma 7, where the Lebesgue-measure-non-zero condition is satisfied by Assumption 5, there exist a constant $const. > 0$ such that $E_1 \geq \left(\frac{c}{c'}-1\right)^2 const.\|f_{\mathbf{\Theta}_{c'}}\|_{L_2} \to const.\|f^*\|_{L_2}$ as $c' \to \infty$. But for E_2, note that $\|f_{\mathbf{\Theta}_{c'}} - f^*\|_{L_2} \to 0$ as $c' \to \infty$. So $E_2 = o(E_1)$, and under mild conditions there exists a constant $F > 0$ such that for all c, c',

$$
R\left(\frac{c}{c'}\mathbf{\Theta}_{c'}\right) - R(\mathbf{\Theta}_{c'}) \leq F\mathbb{E}\left[\left\{\boldsymbol{\alpha}(\mathbf{X})^\top\mathbf{\Theta}_{c'}\left(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})\right)\right\}^2\right] \tag{81}
$$

. Then note:

$$\mathbb{E}\left[\left\{\boldsymbol{\alpha}(\mathbf{X})^\top \boldsymbol{\Theta}_{c'}\left(\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})\right)\right\}^2\right]$$

$$= \mathbb{E}\left[\langle \boldsymbol{\Theta}_{c'}, \boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))\rangle^2\right] \tag{82}$$

$$= \mathbb{E}\left[\langle \boldsymbol{\Theta}_{c'} \otimes \boldsymbol{\Theta}_{c'}, (\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))) \otimes (\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})))\rangle\right] \tag{83}$$

$$= \langle \boldsymbol{\Theta}_{c'} \otimes \boldsymbol{\Theta}_{c'}, \mathbb{E}\left[(\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))) \otimes (\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})))\right]\rangle \tag{84}$$

$$\leq \|\boldsymbol{\Theta}_{c'} \otimes \boldsymbol{\Theta}_{c'}\| \left\|\mathbb{E}\left[(\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))) \otimes (\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})))\right]\right\| \tag{85}$$

$$= \|\boldsymbol{\Theta}_{c'}\|^2 \left\|\underbrace{\mathbb{E}\left[(\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X}))) \otimes (\boldsymbol{\alpha}(\mathbf{T}) \otimes (\boldsymbol{\beta}(\mathbf{T}) - e^p(\mathbf{X})))\right]}_{constant}\right\| \tag{86}$$

$$= const. \|f_{\boldsymbol{\Theta}_{c'}}\|^2_{\mathcal{H}_{c'}} \tag{87}$$

$$= const. \|f_{\boldsymbol{\Theta}_{c'}}\|^2_{L_2} \tag{88}$$

where Eq. 85 is by Cauchy-Schwarz and the equation 86 uses the fact that under Euclidean norms for finite dimensional real vectors $\mathbf{a}, \mathbf{b}$, $\|\mathbf{a} \otimes \mathbf{b}\| = \|\mathbf{a}\|\|\mathbf{b}\|$. equation 87 is due to the vector 2-norm of $\boldsymbol{\Theta}$ is equal to the RKHS norm of $f_{\boldsymbol{\Theta}}$, and equation 88 is due to the fact that in finite dimensions all norms are Lipschitz equivalent. Note that the constant factors in 87 and 88 may be different but that both positive.

Then finally by the triangle inequality,

$$\|f_{\Theta_c} - f_{\boldsymbol{\Theta}_{c'}}\|_{L_2} \leq \|f_{\boldsymbol{\Theta}_{c'}} - \frac{c}{c'} f_{\boldsymbol{\Theta}_{c'}}\|_{L_2} + \|f_{\Theta_c} - \frac{c}{c'} f_{\boldsymbol{\Theta}_{c'}}\|_{L_2} \tag{89}$$

$$\leq \left(1 - \frac{c}{c'}\right) \|f_{\boldsymbol{\Theta}_{c'}}\|_{L_2} + constant. \|f_{\boldsymbol{\Theta}_{c'}}\|_{L_2} \tag{90}$$

$$\leq const. \|f_{\boldsymbol{\Theta}_{c'}}\|_{L_2} \tag{91}$$

again for a positive constant factor in the last equality. □

Now we have arrived at the position to bound the difference between the oracle and feasible regrets by functions of the true regret. We first present Lemma 11 which bounds the difference between $\hat{R}_n$ and $\tilde{R}_n$ in terms of R. Then, we leverage the result by [42] to linearize the dependence on R.

Lemma 11. *Suppose that the propensity estimate $e^p(\mathbf{x})$ is uniformly consistent,*

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\widehat{e}^p(\mathbf{x}) - e^p(\mathbf{x})\| \to_p 0 \tag{92}$$

and the L_2 errors converge at rate

$$\mathbb{E}\left[\{\widehat{m}(\mathbf{X}) - m^*(\mathbf{X})\}^2\right], \mathbb{E}\left[\|\widehat{e}^p(\mathbf{X}) - e^p(\mathbf{X})\|^2\right] = \mathcal{O}(a_n^2) \tag{93}$$

for some sequence $a_n \to 0$. Suppose, moreover, Assumptions 5, 6 and 4 hold. Then, for any $\epsilon > 0$, there exists a constant $U(\epsilon)$ such that the regret functions induced by the oracle learner and the feasible learner are coupled with probability at least $1 - \epsilon$ as

$$\begin{aligned}\left|\widehat{R}_n(\boldsymbol{\Theta}; c) - \widetilde{R}_n(\boldsymbol{\Theta}; c)\right| \leq U(\varepsilon) \Big\{ & c^p R(\boldsymbol{\Theta}; c)^{(1-p)/2} a_n^2 + c^{2p} R(\boldsymbol{\Theta}; c)^{1-p} \tfrac{1}{\sqrt{n}} \log(n) \\ & + c^{2p} R(\boldsymbol{\Theta}; c)^{1-p} \tfrac{1}{n} \log\left(\tfrac{cn^{1/(1-p)}}{R(\boldsymbol{\Theta}; c)}\right) + c^p R(\boldsymbol{\Theta}; c)^{1-\frac{p}{2}} \tfrac{1}{\sqrt{n}} \sqrt{\log\left(\tfrac{cn^{1/(1-p)}}{R(\boldsymbol{\Theta}; c)}\right)} \\ & + c^p R(\boldsymbol{\Theta}; c)^{(1-p)/2} a_n \tfrac{1}{\sqrt{n}} \sqrt{\log\left(\tfrac{cn^{1/(1-p)}}{R(\boldsymbol{\Theta}; c)}\right)} + \xi_n R(\boldsymbol{\Theta}; c) \Big\}\end{aligned} \tag{94}$$

simultaneously for all $1 \leq c \leq \log(n)$.

Proof. Following [42], we start by decomposing the feasible loss function $\hat{L}_n(\boldsymbol{\Theta})$ into the oracle loss together with additional terms as follows:

$$\hat{L}_n(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{l=1}^{n}\Big((Y_l - \widehat{m}_{(-q(l))}(\mathbf{X}_l)) - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\Big)^2 \tag{95}$$

$$= \frac{1}{n}\sum_{l=1}^{n}\Big[(Y_l - m^*(\mathbf{X}_l)) + \{m^*(\mathbf{X}_l) - \widehat{m}(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l)) - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\Big]^2 \tag{96}$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{l=1}^{n}\Big[\{Y_l - m^*(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\Big]^2 \\
&\quad + \frac{1}{n}\sum_{l=1}^{n}[\{m^*(\mathbf{X}_l) - \widehat{m}(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))})]^2 \\
&\quad + \frac{2}{n}\sum_{l=1}^{n}\Big[\{Y_l - m^*(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\Big] \\
&\cdot \Big[\{m^*(\mathbf{X}_l) - \widehat{m}_{(-q(l))}(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\Big]
\end{aligned} \tag{97}$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{l=1}^{n}\Big[\{Y_l - m^*(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\Big]^2 \\
&\quad + \frac{1}{n}\sum_{l=1}^{n}\Big[\{m^*(\mathbf{X}_l) - \widehat{m}_{(-q(l))}(\mathbf{X}_l)\} - \boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\Big]^2 \\
&\quad - \frac{2}{n}\sum_{l=1}^{n}\{Y_l - m^*(\mathbf{X}_l)\}\boldsymbol{\alpha}(\mathbf{X}_l)\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l)) \\
&\quad - \frac{2}{n}\sum_{l=1}^{n}\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\{m^*(\mathbf{X}_l) - \widehat{m}_{(-q(l))}(\mathbf{X}_l)\} \\
&\quad + \frac{2}{n}\sum_{l=1}^{n}\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))
\end{aligned} \tag{98}$$

Furthermore, we may verify that some terms cancel out when we restrict our attention to the main objective of interest

$$\widehat{R}(\boldsymbol{\Theta}; c) - \tilde{R}(\boldsymbol{\Theta}; c) = \hat{L}_n(\boldsymbol{\Theta}) - \hat{L}_n(\boldsymbol{\Theta}_c) - \tilde{L}_n(\boldsymbol{\Theta}) + \tilde{L}_n(\boldsymbol{\Theta}_c) \tag{99}$$

In particular, note that the first term in the decomposition above is exactly $\tilde{L}_n(\boldsymbol{\Theta})$. Thus

$$
\begin{aligned}
&\widehat{R}(\boldsymbol{\Theta}; c) - \tilde{R}(\boldsymbol{\Theta}; c) \\
&= -\frac{2}{n}\sum_{l=1}^{n}\{m^*(\mathbf{X}_l) - \widehat{m}_{(-q(l))}(\mathbf{X}_l)\}\boldsymbol{\alpha}(\mathbf{X}_l)^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}_c)(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l)) \\
&\quad + \frac{1}{n}\sum_{l=1}^{n}\{\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\}^2 - \{\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}_c(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\}^2 \\
&\quad - \frac{2}{n}\sum_{l=1}^{n}\{Y_l - m^*(\mathbf{X}_l)\}\boldsymbol{\alpha}(\mathbf{X}_l)^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}_c)(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l)) \\
&\quad - \frac{2}{n}\sum_{l=1}^{n}\boldsymbol{\alpha}(\mathbf{X}_l)^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}_c)(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\{m^*(\mathbf{X}_l) - \widehat{m}_{(-q(l))}(\mathbf{X}_l)\} \\
&\quad + \frac{2}{n}\sum_{l=1}^{n}\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l)) \\
&\quad - \frac{2}{n}\sum_{l=1}^{n}\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}_c(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\boldsymbol{\alpha}(\mathbf{X}_l)^\top\boldsymbol{\Theta}_c(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))
\end{aligned}
\tag{100}
$$

Letting $A_1^c(\boldsymbol{\Theta})$, $A_2^c(\boldsymbol{\Theta})$, $B_1^c(\boldsymbol{\Theta})$ and $B_3^c(\boldsymbol{\Theta})$ denote these 5 summands respectively, we seek to bound each of the terms in terms of $R(\boldsymbol{\Theta}; c)$. Starting with $A_1^c(\boldsymbol{\Theta})$, we extract $\boldsymbol{\Theta} - \boldsymbol{\Theta}_c$ by its infinity norm and by Cauchy-Schwarz,

$$
|A_1^c(\boldsymbol{\Theta})| \leq 2\sqrt{\frac{1}{n}\sum_{l=1}^{n}\left\{m^*(\mathbf{X}) - \widehat{m}_{(-q(l))}(\mathbf{X}_l)\right\}^2}
\tag{101}
$$

$$
\cdot\sqrt{\frac{1}{n}\sum_{l=1}^{n}\left\|\boldsymbol{\alpha}(\mathbf{X}_l)\otimes\left(e^p(\mathbf{X}) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l)\right)\right\|^2}\cdot\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_c\|_\infty
\tag{102}
$$

Using the fact that $\|\mathbf{a}\otimes\mathbf{b}\| = \|\mathbf{a}\|\|\mathbf{b}\|$ for $\mathbf{a}$ and $\mathbf{b}$ some (finite dimensional) vector, we may separate out the norm of $\boldsymbol{\alpha}(\mathbf{X})$ and we know $\|\boldsymbol{\alpha}(\mathbf{X})\|^2$ is uniformly bounded by Assumption 6. By equation 93 and Markov's inequality, the mean squared errors of the $m-$ and $e-$models decay at rate $O_P(a_n)$. Therefore, applying 71 to bound the infinity-norm discrepancy $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_c\|_\infty$, we find that simultaneously for all $c \geq 1$,

$$
\sup\{c^{-p}R(\boldsymbol{\Theta}; c)^{-\frac{1-p}{2}}|A_1^c(\boldsymbol{\Theta})| : f_\Theta \in \mathcal{H}_c, c \geq 1\} = O_P(a_n^2)
\tag{103}
$$

Following [42] and using a similar argument to extract $\|\boldsymbol{\alpha}(\mathbf{X})\|$ and bound $\boldsymbol{\Theta} - \boldsymbol{\Theta}_c$ by the c-regret $\|R(\boldsymbol{\Theta}; c)\|$, we get that

$$
|A_2^c| = \mathcal{O}_P\left(\left(c^pR(\boldsymbol{\Theta}; c)^{\frac{1-p}{2}} + c^{2p}R(\boldsymbol{\Theta}; c)^{1-p}\right)a_n^2\right)
\tag{104}
$$

In order to bound $B_1^c(\boldsymbol{\Theta})$, decomposing it with respect to the cross fitting structure, we consider

$$
B_{1,q}^c(\boldsymbol{\Theta}) = \frac{\sum_{\{l:q(l)\}}2\{Y - m^*(\mathbf{X})\}\boldsymbol{\alpha}(\mathbf{X}_l)^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}_c)(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))}{|\{l : q(l) = q\}|},
\tag{105}
$$

noting that $|B_1^c(\boldsymbol{\Theta})| \leq \sigma_{q=1}^Q|B_{1,q}^c(\boldsymbol{\Theta})|$. In particular, we bound its supremum $\sup B_{1,q}^c(\boldsymbol{\Theta})$. To proceed, we bound this quantity over sets indexed by c and δ such that $\|f_{\boldsymbol{\Theta}} - f_{\boldsymbol{\Theta}_c}\|_{L^2} \leq \delta$:

$$
\sup_{\boldsymbol{\Theta}\in\mathcal{H}_c}\left\{B_{1,q}^c(\boldsymbol{\Theta}) : \|f_{\boldsymbol{\Theta}} - f_{\boldsymbol{\Theta}_c}\|_{L^2} \leq \delta\right\}.
\tag{106}
$$

Letting $\mathcal{I}^{(-q)} = \{\mathbf{X}_l, \mathbf{T}_l, Y_l : q(l) \neq q\}$ denote the set of data points excluded in the $q-$fold, using a similar procedure to [42], we can check that the conditional expectation $\mathbb{E}\left[B^c_{1,q} \mid \mathcal{I}^{(-q)}\right] = 0$. By conditioning on $\mathcal{I}^{(-q)}$, the summands in $B^c_{1,q}(\mathbf{\Theta})$ become independent, as $\widehat{e}^p(\mathbf{X})(\mathbf{X}_l)$ is now only random in $\mathbf{X}$.

Now, the next step in [42] is to bound the expectation of the supremum of $B^c_{1,q}$ using [42, Lemma 5] and [42, Eq. (36)]. Since we work with a vector of propensity features instead of a single propensity score unlike in [42], we need to apply [42, Lemma 5] d times where d is the dimension of $e^p(\mathbf{X})$:

$$B^c_{1,q}(\mathbf{\Theta}) = \frac{\langle(\mathbf{\Theta} - \mathbf{\Theta}_c), \sum_{\{l:q(l)\}} 2\{Y - m^*(\mathbf{X})\}\boldsymbol{\alpha}(\mathbf{X}_l) \otimes (e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\rangle}{|\{l : q(l) = q\}|} \tag{107}$$

$$= \frac{\sum_{ij}(\mathbf{\Theta} - \mathbf{\Theta}_c)_{ij}, \sum_{\{l:q(l)\}} 2\{Y - m^*(\mathbf{X})\}\alpha_i(\mathbf{X}_l)(e^p_j(\mathbf{X}_l) - \hat{e}^p_{(-q(l)),j}(\mathbf{X}_l))}{|\{l : q(l) = q\}|}, \tag{108}$$

so

$$\sup_{f_{\mathbf{\Theta}} \in \mathcal{H}_c} \{B^c_{1,q}(\mathbf{\Theta})\} \leq \frac{\sum_{ij} \sup_{f_{\mathbf{\Theta}} \in \mathcal{H}_c} \sum_{\{l:q(l)\}} 2\{Y - m^*(\mathbf{X})\}\alpha_i(\mathbf{X}_l)(e^p_j(\mathbf{X}_l) - \hat{e}^p_{(-q(l)),j}(\mathbf{X}_l))(\mathbf{\Theta} - \mathbf{\Theta}_c)_{ij}}{|\{l : q(l) = q\}|} \tag{109}$$

So bounding each term indexed by ij using Lemma 5 of [42] and equation 93, we will get the same bound as in [42] because the sum over ij is finite and $\boldsymbol{\alpha}$ is bounded.

Then, using a similar argument to [42], we may obtain that for any fixed $c, \delta, \epsilon > 0$, there exists a different constant B such that with probability at least $1 - \epsilon$,

$$\sup_{\tau \in \mathcal{H}_c} \left\{B^c_{1,q}(\tau) \mid \mathcal{I}^{(-q)} : \|f_{\mathbf{\Theta}} - f_{\mathbf{\Theta}_c}\|_{L_2} \leq \delta\right\}$$
$$< B\left\{c^p\delta^{1-p}a_n\frac{\log(n)}{\sqrt{n}} + \frac{c^p\delta^{1-p}a_n}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\varepsilon}\right)} + \frac{1}{n}c^p\delta^{1-p}\log\left(\frac{1}{\varepsilon}\right)\right\}, \tag{110}$$

which holds unconditionally of $\mathcal{I}^{(-q)}$. In order to establish the bound for all values of c and δ simultaneously, we may proceed with the same argument as [42]; instead of [42, Lemma 6], we replace with our Lemma 10, which is our extension to the multidimensional setting. $B^c_2(\mathbf{\Theta})$ may be bounded similarly.

To bound $B_3(\mathbf{\Theta})$, the argument of [42] is easily extended as well, using the decomposition which we detail below.

To simplify notation, write

$$\mathbf{a}_l = \boldsymbol{\alpha}(\mathbf{X}_l) \otimes \left(\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l)\right) \tag{111}$$

$$\mathbf{b}_l = \boldsymbol{\alpha}(\mathbf{X}_l) \otimes \left(e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l)\right) \tag{112}$$

Note:

$$B_3^c = \frac{2}{n}\sum_{l=1}^{n}\langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle - \frac{2}{n}\sum_{l=1}^{n}\langle \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta}_c, \mathbf{b}_l\rangle \tag{113}$$

$$= \frac{2}{n}\sum_{l=1}^{n}\Bigg\{2\langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle - \langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle - \langle \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle + \langle \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle - \langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta}_c, \mathbf{b}_l\rangle + \langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta}_c, \mathbf{b}_l\rangle - \langle \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta}_c, \mathbf{b}_l\rangle\Bigg\} \tag{114}$$

$$= \frac{2}{n}\sum_{l=1}^{n}\Bigg\{\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle + \langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \mathbf{b}_l\rangle \tag{115}$$

$$- \langle \mathbf{\Theta} - \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \mathbf{b}_l\rangle\Bigg\} \tag{116}$$

$$\leq \left|\frac{2}{n}\sum_{l=1}^{n}\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \mathbf{a}_l\rangle\langle \mathbf{\Theta}, \mathbf{b}_l\rangle\right| + \left|\frac{2}{n}\sum_{l=1}^{n}\langle \mathbf{\Theta}, \mathbf{a}_l\rangle\langle \mathbf{\Theta} - \mathbf{\Theta}_c, \mathbf{b}_l\rangle\right| + \frac{2}{n}\sum_{l=1}^{n}\|\mathbf{\Theta} - \mathbf{\Theta}_c\|_2^2\|\mathbf{a}_l\|_2\|\mathbf{b}_l\|_2 \tag{117}$$

where the last term of the last inequality follows by Cauchy-Schwarz.

The first two terms can be bounded similarly to the argument used for bounding $B_1^c(\mathbf{\Theta})$. For the last term, we note that $\|\mathbf{\Theta} - \mathbf{\Theta}_c\|_2 = \|f_\Theta - f_{\Theta_c}\|_{L_2}$ since by construction the RKHS norm and the L_2 norms are equal. Therefore, the last term is bounded by $\xi_n\|f_\Theta - f_{\Theta_c}\|_{L_2}$ where

$$\xi_n = \|\boldsymbol{\alpha}(\mathbf{X}_l) \otimes (\boldsymbol{\beta}(\mathbf{T}_l) - e^p(\mathbf{X}_l))\|_\infty\|\boldsymbol{\alpha}(\mathbf{X}_l) \otimes (e^p(\mathbf{X}_l) - \widehat{e}^p_{(-q(l))}(\mathbf{X}_l))\|_\infty = o(1). \tag{118}$$

Note that we do not need the lower order terms present in [42] which followed from [42, Lemma 7]. Thus the desired result follows. □

By [42, Lemma 2], Lemma 11 implies that under Assumptions 6 to 4, and the conditions in Lemma 11 and Lemma 4, where the (a_n) in Lemma 11 is such that $a_n = O(n^{-\kappa})$ with $\kappa > \frac{1}{4}$, then

$$\left|\hat{R}_n(\mathbf{\Theta}; c) - \tilde{R}_n(\mathbf{\Theta}; c)\right| \leq 0.125R(\mathbf{\Theta}; c) + o(\rho_n(c)) \tag{119}$$

with probability at least $1 - \epsilon$, for all $\mathbf{\Theta} \in \mathcal{H}_c$, $1 \leq c \leq \log(n)$ for large enough n.

Thus we have finally bridged $\hat{R}_n$ and $\tilde{R}_n$ with respect to the expected regret R. We are ready to prove our main theorem which concerns the regret bound of $\hat{R}_n$.

F.8 Using the bridge result to derive feasible regret bound

Theorem 2. *Under Assumptions 5, 6, 4 and the conditions in Lemma 11 and Lemma 4, where the (a_n) in Lemma 11 is such that $a_n = O(n^{-\kappa})$ with $\kappa > \frac{1}{4}$, and suppose that we obtain $\hat{\mathbf{\Theta}}$ via a penalized basis function regression variant of the generalized R-learner, with a properly chosen penalty of the form $\Lambda_n(\|\hat{\mathbf{\Theta}}\|_2)$ that grows faster than $\rho_n(\|\hat{\mathbf{\Theta}}\|_2)$ in 53 . Then $\hat{\mathbf{\Theta}}$ satisfies the same regret bound as $\tilde{\mathbf{\Theta}}$, $R(\hat{\mathbf{\Theta}}_n) = \tilde{O}(n^{\frac{1}{1+p}})$.*

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:

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

□