Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design

Ahmed M. Alaa [1] **Mihaela van der Schaar** [1 2 3]

Abstract

Estimating heterogeneous treatment effects from observational data is a central problem in many domains. Because *counterfactual* data is inaccessible, the problem differs fundamentally from supervised learning, and entails a more complex set of modeling choices. Despite a variety of recently proposed algorithmic solutions, a principled guideline for building estimators of treatment effects using machine learning algorithms is still lacking. In this paper, we provide such a guideline by characterizing the *fundamental limits* of estimating heterogeneous treatment effects, and establishing conditions under which these limits can be achieved. Our analysis reveals that the relative importance of the different aspects of observational data vary with the sample size. For instance, we show that *selection bias* matters only in small-sample regimes, whereas with a large sample size, the way an algorithm models the *control* and *treated* outcomes is what bottlenecks its performance. Guided by our analysis, we build a practical algorithm for estimating treatment effects using a *non-stationary* Gaussian processes with *doubly-robust* hyperparameters. Using a standard semi-synthetic simulation setup, we show that our algorithm outperforms the state-of-the-art, and that the behavior of existing algorithms conforms with our analysis.

1. Introduction

The problem of estimating heterogeneous (individualized) causal effects of a treatment from observational data is central in many application domains, including public health and drug development (Foster et al., 2011), computational advertising (Bottou et al., 2013), and social sciences (Xie et al., 2012). The increasing availability of observational data in all these domains has encouraged the development of various machine learning algorithms tailored for inferring treatment effects using observational data (e.g. (Li & Fu, 2017; Wager & Athey, 2017; Shalit et al., 2017; Alaa & van der Schaar, 2017)). Due to the peculiarity of the treatment effect estimation problem, these algorithms needed to address various modeling aspects that are foreign to standard supervised learning setups; such aspects include ways to handle *sample selection bias* (Heckman, 1977), and ways to model *treated* and *untreated* data points. Despite a variety of recent algorithmic approaches, principled guidelines for model design are lacking.

In this paper, we provide guidelines for designing practical treatment effect estimation algorithms in the context of Bayesian nonparametric inference, and propose one possible instantiation of an algorithm that follows our guidelines. We set these guidelines by characterizing the fundamental limits of estimating treatment effects, and studying the impact of various common modeling choices on the achievability of those limits. In what follows, we provide a brief technical background for the treatment effect estimation problem, along with a summary of our contributions.

1.1. Background and Summary of Contributions

Our analysis hinges on the Rubin-Neyman potential outcomes model (Rubin, 2005). That is, we consider an observational dataset with a population of subjects, where each subject i is endowed with a d-dimensional feature $X_i \in \mathcal{X}$. We assume that $\mathcal{X} = [0,1]^d$, but most of our results hold for general compact metric spaces (bounded, closed sets in $\mathbb{R}^d$). A treatment assignment indicator $W_i \in \{0,1\}$ is associated with subject i; $W_i = 1$ if the treatment under study was applied to subject i, and $W_i = 0$ otherwise. Subject i's responses with and without the treatment (the potential outcomes) are denoted as $Y_i^{(1)}$ and $Y_i^{(0)}$, respectively. Treatments are assigned to subjects according to an underlying policy that depends on the subjects' features, i.e. $W_i \not\perp\!\!\!\perp X_i$. This dependence is quantified via the conditional distribution $p(x) = \mathbb{P}(W_i = 1 | X_i = x)$, also known as the *propensity score* of subject i (Rosenbaum & Rubin,

[1]University of California, Los Angeles, USA [2]University of Oxford, Oxford, UK [3]Alan Turing Institute, London, UK. Correspondence to: Ahmed M. Alaa <ahmedmalaa@ucla.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

1984). The response $Y_i^{(W_i)}$ is the "factual outcome" which we observe in the data, whereas $Y_i^{(1-W_i)}$ is the unrealized "counterfactual outcome" (Bottou et al., 2013). An observational dataset $\mathcal{D}_n$ comprises n samples of the form:

$$\boxed{\boldsymbol{\mathcal{D}_n = \{X_i, W_i, Y_i^{(W_i)}\}_{i=1}^n}} \tag{1}$$

The causal effect of the treatment on subject i with a feature $X_i = x$ is characterized through the *conditional average treatment effect* (CATE) function $T(x)$, which is defined as the expected difference between the two potential outcomes (Rubin, 2005), i.e.

$$\boxed{\boldsymbol{T(x) = \mathbb{E}[\, Y_i^{(1)} - Y_i^{(0)} \mid X_i = x \,]}} \tag{2}$$

Our goal is to identify a set of guiding principles for building estimators of the CATE $T(x)$ using samples from $\mathcal{D}_n$. Throughout the paper, we will assume that the joint density $d\mathbb{P}(X_i, W_i, Y_i^{(0)}, Y_i^{(1)})$ supports the assumptions of *unconfoundedness* and *overlap*, which are necessary for causal identifiability and consistency. Unconfoundedness requires that $(Y_i^{(0)}, Y_i^{(1)}) \perp\!\!\!\perp W_i \mid X_i$, whereas overlap requires that $0 < p(x) < 1$ (Rosenbaum & Rubin, 1984). **Selection bias** occurs in $\mathcal{D}_n$ since the distribution of the treated/control subjects does not match that of the overall population.

In order to come up with principled guidelines for building estimators of $T(x)$, we characterize the fundamental (information-theoretic) limits of estimating the CTAE using samples from $\mathcal{D}_n$, and identify the modeling choices that would allow achieving those limits. To this end, in **Section 3** we tackle the following question: **what are the fundamental limits of CATE estimation?** We answer this question by deriving the *optimal minimax rate* for estimating $T(x)$ using $\mathcal{D}_n$. Interestingly, it turns out that the optimal rate **does not** depend on **selection bias**, but rather on the **smoothness** and **sparsity** of the more "complex" of the functions $\mathbb{E}[\, Y_i^{(0)} \mid X_i = x \,]$ and $\mathbb{E}[\, Y_i^{(1)} \mid X_i = x \,]$. We focus our analysis on Bayesian nonparametric methods, since they have the appealing properties of being robust to misspecification and are accessible for theoretical analysis.

Our analysis reveals that the relative importance of the different modeling aspects vary with the sample size. In particular, in the **large-sample regime**, selection bias does not pose a serious problem, and the model's performance would be mainly determined by its **structure**, i.e. the way the outcomes $Y_i^{(0)}$ and $Y_i^{(1)}$ are modeled, and the impact of that on variable selection and hyperparameter tuning. On the contrary, selection bias can seriously harm a model's generalization performance in **small-sample regimes**. A good model should then be carefully designed so that it operates well in both regimes by possessing the right **model structure** that would allow learning at a fast rate, and the right **model selection** (hyperparameter optimization) scheme that would account for selection bias.

In Section 4, we build a practical CATE estimation algorithm guided by the results of the analyses in Section 3. We model the outcomes $Y_i^{(0)}$ and $Y_i^{(1)}$ using a Gaussian process with a *non-stationary* kernel that captures the different relevant variables and different levels of smoothness of the functions $\mathbb{E}[\, Y_i^{(0)} \mid X_i = x \,]$ and $\mathbb{E}[\, Y_i^{(1)} \mid X_i = x \,]$. We prove that this model structure can achieve the optimal rate of CATE estimation when tuned with the right hyperparameters. We also propose a *doubly-robust* hyperparameter optimization scheme that accounts for selection bias in small-sample regimes, without hindering the model's minimax-optimality in the large sample limit. We show that our algorithm outperforms state-of-the-art methods using a well-known semi-synthetic simulation setup.

1.2. Related Work

Very few works have attempted to characterize the limits of CATE estimation, or study the impact of different modeling choices on the CATE estimation performance in a principled manner. (Alaa & van der Schaar, 2018) characterized the asymptotic "information rates" for different CATE estimators, but provided no clear guidelines on practical model design or an analysis of the impact of sample selection bias. The study in (Künzel et al., 2017) was rather empirical in nature, comparing the performance of different regression structures for the potential outcomes while ignoring selection bias. A similar study, but focusing only on random forest models, was conducted in (Lu et al., 2017).

Most of the previous works have been algorithmic in nature, focusing mainly on devising algorithms that correct for selection bias (e.g. (Johansson et al., 2016; Shalit et al., 2017; Wager & Athey, 2017; Li & Fu, 2017)). Some of these works cast the selection bias problem as a problem of *covariate shift* (Sugiyama et al., 2007), and use techniques from *representation learning* to learn feature maps that balance the biased data (e.g. (Li & Fu, 2017; Shalit et al., 2017; Johansson et al., 2016)). However, those works report much bigger improvements in CATE estimation when changing their model structure (e.g. architecture of a neural network), as compared to the gains attained by only accounting for bias (see the comparisons between the TARnet and BNN models in (Shalit et al., 2017)). Similar observations are reported in (Alaa & van der Schaar, 2017; Atan et al., 2018), where the selection of the model structure seemed to influence the achieved CATE estimation performance even when selection bias is not accounted for. Despite of that, none of these works offer a discussion on whether selection bias is actually the main challenge in CATE estimation, or whether the outcomes' model structure may have a bigger influence on performance.

In contrast to the works above, this paper does not attempt to develop a model by presupposing that particular model-

ing aspects are of greater importance than others, but rather provides a framework for understanding the limits on the achievable performance, and how different modeling aspects influence a model's chance of achieving those limits. We use our analyses to both reflect on the modeling choices made in the works above, and also devise a novel, principled CATE estimation algorithms that achieves the fundamental performance limits.

2. Estimating CATE: Problem Setup

2.1. Potential Outcomes & Propensity Score

We consider the following *random design* regression model for the potential outcomes:

$$Y_i^{(w)} = f_w(X_i) + \varepsilon_{i,w},\ w \in \{0,1\}, \tag{3}$$

where $\varepsilon_{i,w} \sim \mathcal{N}(0, \sigma_w^2)$ is a Gaussian noise variable. It follows from (2) that the CATE is $T(x) = f_1(x) - f_0(x)$. The *response surfaces* $f_1(x)$ and $f_0(x)$ correspond to the subjects' responses with and without the treatment.

We assume that $f_w(.) : \mathcal{X} \to \mathbb{R}$, $w \in \{0,1\}$, is a totally bounded function that lives in a space of "smooth" or "regular" functions, with an unknown smoothness parameter α_w. We use Hölder balls for concreteness, although our results extend to other function spaces. A function $f_w(.)$ lies in the Hölder ball H^{α_w}, with a Hölder exponent $\alpha_w > 0$, if and only if it is bounded in sup-norm by a constant $C > 0$, all its partial derivatives up to order $\lfloor \alpha_w \rfloor$ exist, and all its partial derivatives of order $\lfloor \alpha_w \rfloor$ are Lipschitz with exponent $(\alpha_w - \lfloor \alpha_w \rfloor)$ and constant C. The Hölder exponents quantify the complexities of f_0 and f_1, and hence the hardness of estimating $T(x)$ would depend on α_0 and α_1.

2.2. Bayesian Nonparametric Inference

Nonparametric inference is immune to misspecification of the outcomes' and propensity models (Kennedy, 2018), and hence we focus on Bayesian nonparametric methods for inferring $T(.)$ on the basis of $\mathcal{D}_n$. Bayesian inference entails specifying a prior distribution $\mathit{\Pi}$ over $f_1(.)$ and $f_0(.)$, i.e.

$$f_0, f_1 \sim \mathit{\Pi}(\bar{\varphi}_{\beta_0}, \bar{\varphi}_{\beta_1}), \tag{4}$$

where $\bar{\varphi}_{\beta_w} = \{\varphi^k_{\beta_w}\}_{k=1}^{\infty}, w \in \{0,1\}$, are complete orthonormal bases (indexed by a parameter $\beta_w > 0$) with respect to Lebesgue measure in $\mathcal{X}$, $f_w = \sum_k \bar{f}^k_w \cdot \varphi^k_{\beta_w}$, and $\bar{f}^k_w = \langle f_w, \varphi^k_{\beta_w} \rangle$. Thus, for given bases $\bar{\varphi}_{\beta_0}$ and $\bar{\varphi}_{\beta_1}$, $\mathit{\Pi}$ places a probability distribution on the projections $\{\bar{f}^k_w\}_k$. Potential choices for the basis $\bar{\varphi}_{\beta_w}$ that would give rise to implementable Bayesian inference algorithms include regular wavelet basis (Zhang, 1997), radial basis for a reproducing kernel Hilbert space (RKHS) (van der Vaart et al., 2008), etc. In general, the parameter β_w would determine the smoothness of the function space spanned by $\bar{\varphi}_{\beta_w}$.

2.3. Towards Principled CATE Estimation

To evaluate the predictive accuracy of the Bayesian inference procedure, we analyze the "frequentist" loss of point estimators $\hat{T}(x)$ induced by the Bayesian posterior $d\mathit{\Pi}_n(T(x) \,|\, \mathcal{D}_n)$, assuming that $\mathcal{D}_n$ is generated based on fixed, *true* response surfaces $f_1(x)$ and $f_0(x)$. (This type of analysis is sometimes referred to as the "Frequentist-Bayes" analysis (Sniekers et al., 2015).) In particular, we quantify the performance of a point estimator $\hat{T}(x) = \delta(d\mathit{\Pi}_n(T(x) \,|\, \mathcal{D}_n))$ by its squared-$L^2(\mathbb{P})$ error, which was dubbed the *precision of estimating heterogeneous effects* **(PEHE)** in (Hill, 2011), and is formally defined as:

$$\boldsymbol{\psi}(\hat{T}) \triangleq \mathbb{E} \,\|\, \hat{T} - T \,\|^2_{L^2(\mathbb{P})}, \tag{5}$$

where $L^2(\mathbb{P})$ is the L^2 norm with respect to the feature distribution, i.e. $\|f(x)\|^2_{L^2(\mathbb{P})} = \int f^2(x)\, d\mathbb{P}(X = x)$.

Not a standard supervised learning problem...

The "fundamental problem of causal inference" is that for every subject i in $\mathcal{D}_n$, we only observe the **factual** outcome $Y_i^{(W_i)}$, whereas the **counterfactual** $Y_i^{(1-W_i)}$ remains unknown, which renders empirical evaluation of the PEHE in (5) impossible. Moreover, $\mathcal{D}_n$ would generally exhibit sample **selection bias** (Heckman, 1977), because the treatment assignment mechanism (decided by $p(x)$) creates a discrepancy between the feature distributions of the treated/control population and the overall population. Thus, standard **supervised learning** approaches based on empirical risk minimization cannot be used to learn a generalizable model for the CATE from samples in $\mathcal{D}_n$. This gives rise to the following fundamental modeling questions that are peculiar to the CATE estimation problem:

- **[Q1]:** How should the treatment assignment indicator W_i be incorporated into the learning model?
- **[Q2]:** How should selection bias be handled?

Adequate answers to **[Q1]** and **[Q2]** would provide guidelines for selecting the prior $\mathit{\Pi}(\bar{\varphi}_{\beta_0}, \bar{\varphi}_{\beta_1})$. Addressing the modeling questions above requires a profound understanding of the **fundamental limits** of CATE estimation, in addition to an understanding of the impact of different modeling choices on the **achievability** of such limits. The next Sections provide principled answers to **[Q1]** and **[Q2]** by addressing the following, more fundamental questions:

Section 3: What are the *limits* on the performance that can be achieved by *any* estimator of the CATE?

Section 4: How can we build *practical algorithms* that can achieve the performance limits?

3. Fundamental Limits of CATE Estimation

In this Section, we establish an information-theoretic limit on the performance of *any* CATE estimator. In what follows, we use the standard Bachmann-Landau order notation, and write $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$. The notation $a \lesssim b$ means that $a \leq Cb$ for a universal constant C, and $\asymp$ denotes asymptotic equivalence.

3.1. Optimal Minimax Rates

The "hardness" of a nonparametric estimation problem is typically characterized by its *minimax* risk (Stone, 1982), i.e. the minimum worst case risk achieved by *any* estimator when the estimand is known to live in a given function space (Yang et al., 2015). In the following Theorem, we establish the optimal minimax rate for the PEHE risk in terms of the complexity of the response surfaces f_0 and f_1.

Theorem 1. *Suppose that $\mathcal{X} = [0,1]^d$, and that f_w depends on a subset of d_w features with $d_w \leq \min\{n, d\}$ for $w \in \{0,1\}$. If $f_0 \in H^{\alpha_0}$ and $f_1 \in H^{\alpha_1}$, then the optimal minimax rate is:*

$$\inf_{\hat{T}} \sup_{f_0, f_1} \boldsymbol{\psi}(\hat{T}) \asymp \underbrace{n^{-\left(1+\frac{1}{2}\left(\frac{d_0}{\alpha_0} \vee \frac{d_1}{\alpha_1}\right)\right)^{-1}}}_{\textbf{CATE estimation}} \vee \underbrace{\log\left(\frac{d^{d_0+d_1}}{d_0^{d_0}\, d_1^{d_1}}\right)^{\frac{1}{n}}}_{\textbf{Variable selection}}.$$

The above holds for any $p(.) \in H^{\alpha_p}$, $\alpha_p > 0$. □

In Theorem 1, the supremum is taken over α_w-Hölder balls ($w \in \{0,1\}$), whereas the infimum is taken over all possible Bayesian estimators. The minimax rate in Theorem 1 corresponds to the **fastest rate** by which **any** (Bayesian) estimator $\hat{T}(.)$ can approximate the CATE function $T(.)$. The proof of Theorem 1 (provided in the supplement) uses information-theoretic techniques based on Fano's method to derive algorithm-independent estimation rates (Yang & Barron, 1999). In the following set of remarks, we revisit **[Q1]** and **[Q2]** in the light of the results of Theorem 1.

How can Theorem 1 help us address [Q1] & [Q2]?

▷ **Remark 1 (Smoothness & sparsity)**

Theorem 1 says that estimating CATE is as hard as nonparametric regression for functions with additive sparsity (Raskutti et al., 2009; Yang et al., 2015). The minimax rate in Theorem 1 decomposes into a term reflecting the complexity of CATE estimation under correct variable selection for f_0 and f_1, and a term reflecting the complexity of variable selection. Variable selection complexity remains small as long as $\log(d) = \Theta(n^{\zeta})$, for some $\zeta \in (0,1)$, and approaches the parametric rates as $\zeta \to 0$. The minimax rate will generally be dominated by the complexity of CATE estimation, and will approach the parametric rates only for very smooth response surfaces with small number of relevant dimensions, i.e. $\frac{d_0}{\alpha_0} \vee \frac{d_1}{\alpha_1} \to 0$.

The main takeaway from Theorem 1 is that the CATE learning rate is determined by the more "complex" of the surfaces f_0 and f_1, where complexity is quantified by the sparsity-to-smoothness ratio d_w/α_w for $w \in \{0,1\}$. Thus, a model would achieve the optimal CATE learning rate only if it selects the correct relevant variables for f_0 and f_1, and tunes its "hyperparameters" (i.e. smoothness of the prior) to cope with a complexity of $\frac{d_0}{\alpha_0} \vee \frac{d_1}{\alpha_1}$. When $\frac{d_0}{\alpha_0}$ and $\frac{d_1}{\alpha_1}$ are very different (e.g. f_0 and f_1 have different relevant features), rate-optimal estimation is possible only if the model incorporates such differences in $\Pi(\bar{\varphi}_{\beta_0}, \bar{\varphi}_{\beta_1})$.

The discussion above **provides a concrete answer to [Q1]:** the treatment assignment variable w should be incorporated into the model in such a way that it **encodes the different relevant dimensions and smoothness levels of f_0 and f_1 in the bases $\bar{\varphi}_{\beta_0}$ and $\bar{\varphi}_{\beta_1}$.** (The simplest way to achieve this is to use two separate models for f_0 and f_1.) This is not fulfilled by many of the previous models that built a single regression function of the from $f : \mathcal{X} \times \{0,1\} \to \mathbb{R}$, and estimated the CATE as $\hat{T}(x) = f(x,1) - f(x,0)$ (Hill, 2011; Johansson et al., 2016; Powers et al., 2017). This is because such models enforced the smoothness of the prior along all features to be the same for $w = 0$ and $w = 1$.

▷ **Remark 2 (Selection bias)**

Theorem 1 gives a rather surprising answer to **[Q2]**: the **optimal learning rate** is **oblivious to selection bias**. Such a finding is consistent with previous results on nonparametric kernel density estimation under selection bias (Borrajo et al., 2017), and parametric Bayesian inference under *covariate shift* (Shimodaira, 2000; Sugiyama & Storkey, 2007). It shows that many of the recent works have missed the target; the works in (Johansson et al., 2016; Shalit et al., 2017; Alaa & van der Schaar, 2017) cast the problem of CATE estimation as one of **covariate shift** that results from selection bias. However, Theorem 1 says that selection bias is not a problem when we have a sufficiently large amount of data. This is because selection bias is inherently a misspecification problem, and hence its impact on nonparametric inference is washed away in large-sample regimes.

Remarks 1 and 2 posit an explanation for various recurrent (empirical) findings reported in previous literature. For instance, (Hahn et al., 2017) found that separate modeling of f_0 and f_1 via Bayesian additive regression trees (BART) outperforms the well-known single-surface BART model developed in (Hill, 2011). Similar findings were reported for models based on Gaussian processes (Alaa & van der Schaar, 2017), and models based on deep neural networks (Shalit et al., 2017). All such findings can be explained in the light of Remark 1. On the other hand, Remark 2 may provide an explanation as to why the "TARnet" model in (Shalit et al., 2017), which models f_0 and f_1 using separate neural networks and does not account for selection

bias, outperformed the "BNN" model in (Johansson et al., 2016), which regularizes for selection bias but fits a single-output network for f_0 and f_1.

3.2. Backing off from "Asymptopia"

Theorem 1 shows that selection bias does not hinder the optimal minimax rates, and that it is only the structural properties of the prior $\Pi(\bar{\varphi}_{\beta_0}, \bar{\varphi}_{\beta_1})$ that determine a model's rate of learning. But does the achieved learning rate suffice as a sole criterion for addressing the modeling questions **[Q1]** and **[Q2]**? The answer is "yes" only if $\mathcal{D}_n$ comes from a large observational dataset, in which case the learning rate suffices as a descriptor for the large-sample performance. However, if $\mathcal{D}_n$ is small, which is typical in post-hoc analyses of clinical trials (Foster et al., 2011), then one should make the design choices that would optimize the small-sample performance. In order to give a more complete picture of the performance in large and small-sample regimes, we derive the following bound on the PEHE:

$$\boldsymbol{\psi}(\hat{T}) \leq \bar{C} \cdot \exp(D_2(Q_0 \,\|\, Q)) \cdot \|f_0 - \hat{f}_0\|^2_{L^2(\mathbb{P}_0)} + \bar{C} \cdot \exp(\underbrace{D_2(Q_1 \,\|\, Q)}_{\substack{\textbf{Réyni} \\ \textbf{Divergence}}}) \cdot \underbrace{\|f_1 - \hat{f}_1\|^2_{L^2(\mathbb{P}_1)}}_{\substack{\textbf{Supervised} \\ \textbf{learning loss}}}, \quad (6)$$

for some $\bar{C} > 0$, where $L^2(\mathbb{P}_w)$, for $w \in \{0, 1\}$, is the L^2 norm with respect to $d\mathbb{P}(X = x \,|\, W = w)$, $Q = d\mathbb{P}(X = x)$, $Q_w = d\mathbb{P}(X = x \,|\, W = w)$, and $D_m(p \,\|\, q)$ is the m^{th} order Réyni divergence. The bound in (6) holds for all $n > 0$, and is tight (refer to the supplement); it shows that the PEHE is a weighted linear combination of the mean squared losses for the two underlying supervised problems of learning f_0 and f_1 with **no covariate shift**, where the weights are determined by the extent of the mismatch between the distributions of the treated and control populations, quantified by the Réyni divergence measure. If $\mathcal{D}_n$ is a dataset obtained from a randomized controlled trial ($Q = Q_0 = Q_1$), then we have $D_2(Q_0 \,\|\, Q) = D_2(Q_1 \,\|\, Q) = 0$, and the bound boils down to a sum of two supervised learning losses, i.e. $\boldsymbol{\psi}(\hat{T}) \leq \bar{C} \cdot \|f_0 - \hat{f}_0\|^2_{L^2(\mathbb{P})} + \bar{C} \cdot \|f_1 - \hat{f}_1\|^2_{L^2(\mathbb{P})}$.

Since the minimax rate for standard nonparametric regression is $\|f_w - \hat{f}_w\|_2^2 \asymp C_w \cdot n^{\frac{-2\alpha_w}{2\alpha_w + d_w}}$ (Stone, 1982), when $d_0/\alpha_0 >> d_1/\alpha_1$, the first-order Taylor approximation for the logarithm of the PEHE in (6) is given by:

$$\log(\boldsymbol{\psi}(\hat{T})) \approx \underbrace{D_2(Q_0 \| Q)}_{\substack{\textbf{Selection} \\ \textbf{bias}}} + \underbrace{\log(C_0)}_{\substack{\textbf{Bias} \\ \textbf{correction}}} - \underbrace{\frac{2\alpha_0}{2\alpha_0 + d_0}}_{\textbf{Learning rate}} \log(n) + O\left(n^{\frac{-2\alpha_1}{2\alpha_1 + d_1} + \frac{2\alpha_0}{2\alpha_0 + d_0}}\right). \quad (7)$$

That is, when viewed on a log-log scale, the behavior of the PEHE versus the number of samples can be described as follows. $\log(\text{PEHE})$ is a linear function of $\log(n)$. Selection bias adds a constant offset to $\log(\text{PEHE})$, but does not affect its slope, which harms the performance only in the small-sample regime. In the large-sample regime, the slope of $\log(\text{PEHE})$, which depends solely on the smoothness and sparsity of the response surfaces, dominates the performance, and selection bias becomes less of a problem. Figure 1 depicts the PEHE in (7) on a log-log scale.

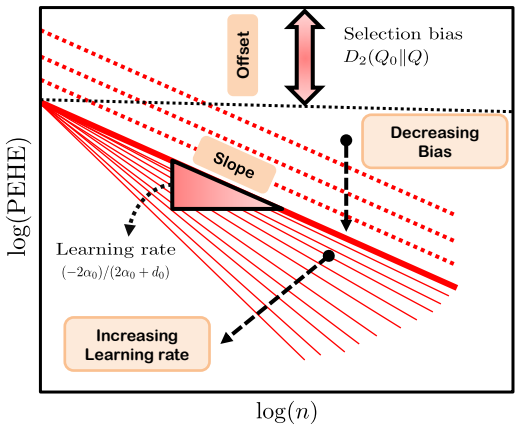


Figure 1. The PEHE in (7) plotted on a log-log scale.

4. CATE Estimation using Non-Stationary Gaussian Process Regression

In this Section, we build on the analyses conducted in Section 3 to design a practical algorithm for CATE estimation.

4.1. Non-Stationary Gaussian Process Priors

We specify the prior $\Pi(\bar{\varphi}_{\beta_0}, \bar{\varphi}_{\beta_1})$ as a Gaussian process (GP) over functions of the form $g : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$, with a kernel $\mathbf{K}_\beta$, and a hyperparameter set β as follows:

$$g \sim \mathcal{GP}\left(0, \mathbf{K}_\beta(z, z')\right), \quad (8)$$

where $z = (x, w) \in \mathcal{X} \times \{0, 1\}$, and $f_w(x) = g(x, w)$. The kernel $\mathbf{K}_\beta$ specifies the bases $\bar{\varphi}_{\beta_0}$ and $\bar{\varphi}_{\beta_1}$ through its induced canonical feature map $\mathbf{K}_\beta(., z)$ (Rasmussen & Williams, 2006; Alvarez et al., 2012). As pointed out in **remark 1**, the treatment assignment variable w should encode the different relevant dimensions and smoothness levels of f_0 and f_1. Thus, we model $\mathbf{K}_\beta$ as a *non-stationary* kernel that depends explicitly on w as follows:

$$\begin{aligned}
\mathbf{K}_\beta(z, z') &= \mathbf{\Gamma}(w, w') \cdot \mathbf{k}^T_\beta(x, x'), \\
\mathbf{k}_\beta(x, x') &= [k_{\beta_0}(x, x'), k_{\beta_1}(x, x'), k_{\beta_0}(x, x') + k_{\beta_1}(x, x')], \\
\mathbf{\Gamma}(w, w') &= [\Gamma_0(w, w'), \Gamma_1(w, w'), 1 - \Gamma_0(w, w') - \Gamma_1(w, w')],
\end{aligned}$$

where $\Gamma_0(w, w') = (1 - w)(1 - w')$, $\Gamma_1(w, w') = w \cdot w'$, and $k_{\beta_w}(x, x')$ is a Matérn kernel with a length-scale parameter

β_w, for $w \in \{0,1\}$. The kernel defined above ensures that any covariance matrix induced by points in $\mathcal{X} \times \{0,1\}$ is positive definite. Variable selection is implemented by using the *automatic relevance determination* version of the Matérn kernel (Rasmussen & Williams, 2006). The non-stationarity of $\mathbf{K}_\beta$ allows setting **different** length-scales and relevant variables for the marginal priors on f_0 and f_1 while sharing data between the two surfaces, i.e.

$$\mathbf{K}_\beta((x,w),(x',w)) = k_{\beta_w}(x,x'),\ w \in \{0,1\},$$
$$\mathbf{K}_\beta((x,w),(x',w')) = k_{\beta_0}(x,x') + k_{\beta_1}(x,x'),\ w \neq w'. \quad (9)$$

That is, all draws from the prior give Matérn sample paths with different smoothness levels (β_0 and β_1) for f_0 and f_1, respectively, and the correlations between the paths are captured via the kernel mixture $k_{\beta_0}(x,x') + k_{\beta_1}(x,x')$. Note that draws from a Matérn prior with length-scale β are almost surely $\bar{\beta}$-Hölder for all $\bar{\beta} \leq \beta$ (Vaart & Zanten, 2011). Thus, $\mathcal{GP}(0, \mathbf{K}_\beta)$ specifies a β_w-Hölder ball as an a priori regularity class for response surface f_w, $w \in \{0,1\}$.

In the following Theorem, we show that point estimators induced by the prior $\mathcal{GP}(0, \mathbf{K}_\beta)$ can achieve the optimal minimax rate in Theorem 1.

Theorem 2. *Suppose that the d_w relevant features for f_w are known a priori for $w \in \{0,1\}$. If $f_0 \in H^{\alpha_0}$, $f_1 \in H^{\alpha_1}$, $\Pi = \mathcal{GP}(0, \mathbf{K}_\beta)$, and $\hat{T} = \mathbb{E}_\Pi[T \mid \mathcal{D}_n]$, then we have that*

$$\boldsymbol{\psi}(\hat{T}) \lesssim n^{-\frac{2(\alpha_0 \wedge \beta_0)}{2\beta_0 + d_0}} \vee n^{-\frac{2(\alpha_1 \wedge \beta_1)}{2\beta_1 + d_1}}$$

whenever $\min\{\alpha_0, \alpha_1, \beta_0, \beta_1\} \geq d/2$. $\square$

Note that posterior consistency holds for all combinations of $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ since the support of the Matérn prior is the space of bounded continuous functions[1]. The bound in Theorem 2 can be shown to be tight using the results in (Castillo, 2008). Theorem 2 says that the posterior induced by the prior $\mathcal{GP}(0, \mathbf{K}_\beta)$ contracts around the true CATE function at the optimal rate given in Theorem 1 provided that the following **matching condition** is met:

$$\beta_v = \alpha_v$$
$$\alpha_v \tfrac{d_{1-v}}{d_v} \leq \beta_{1-v} \leq \alpha_{1-v} + \tfrac{\alpha_{1-v} \cdot d_v}{2\alpha_v} - \tfrac{d_{1-v}}{2}, \quad (10)$$

where $v = 1$ if $d_1/\alpha_1 > d_0/\alpha_0$, and $v = 0$ otherwise. The condition in (10) implies that achieving the optimal rate (steepest slope in Figure 1) via the non-stationary GP prior in Section 4.1 is only a matter of hyperparameter tuning: the smoothness of the prior needs to match the smoothness of the "more complex" of the two response surfaces. Note that Theorem 2 implies that we do not need to handle selection bias in order to achieve the optimal rate, which is consistent with the earlier discussion in **remark 2**.

[1] This is because the RKHS associated with the prior lies dense in the space of bounded continuous functions (van der Vaart & van Zanten, 2008; van der Vaart et al., 2008).

4.2. Doubly-Robust Hyperparameters

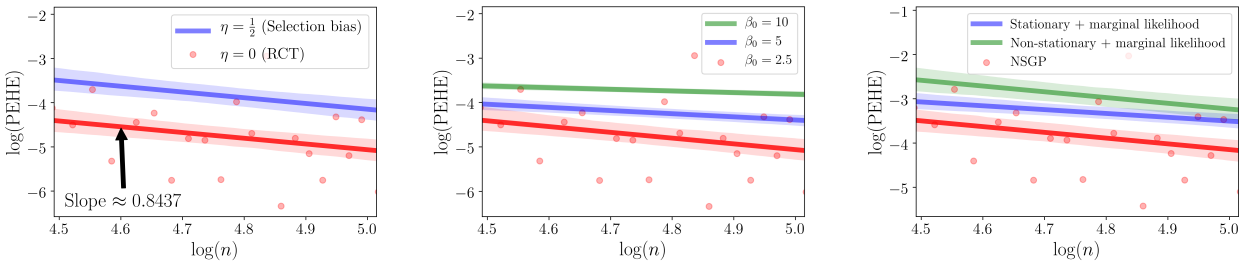
Theorem 2 says that the optimal minimax rate for CATE estimation can be achieved by satisfying the smoothness matching condition in (10). However, in practice, the smoothness levels of the true response functions are unknown and need to be learned from the data. Moreover, since selection bias is impactful in small-sample regimes, ignoring it may lead to a poor generalization performance when the size of $\mathcal{D}_n$ is small. In this Section, we propose a hyperparameter optimization algorithm that accounts for selection bias while ensuring minimax-optimality in the large-sample limit.

Previous works tend to adjust for selection bias "mechanically" using variants of importance sampling approaches based on inverse-propensity-weighting (IPW) (Sugiyama et al., 2007; Shimodaira, 2000), and kernel mean matching (Huang et al., 2007), or by learning a "balanced representation" of treated and control populations (Li & Fu, 2017). We do not attempt to explicitly adjust for selection bias using ad-hoc approaches, and rather seek the "informationally optimal" estimator of the PEHE. That is, we seek the **most efficient** (unbiased) estimator $\hat{\psi}^*(\hat{T})$ of $\psi(\hat{T})$, which satisfies an analog of the Cramér-Rao bound (information-inequality) in parametric estimation, i.e. $\text{Var}[\hat{\psi}^*(\hat{T})] \leq \text{Var}[\hat{\psi}(\hat{T})]$, for any estimator $\hat{\psi}(\hat{T})$.

Classical Cramér-Rao bounds do not apply to estimators of the form $\hat{\psi}^*(\hat{T})$, since such estimators are functionals of nonparametric objects. There are, however, analogous information inequalities for nonparametric estimation, including Bhattacharyya's variance bound (Bhattacharyya, 1946), and its generalization due to Bickel (Bickel et al., 1998). We proceed by realizing that the PEHE $\psi(\hat{T})$ is simply a functional that belongs to the *doubly-robust* class of functionals analyzed by Robins in (Robins et al., 2008). Thus, one can construct the "most" efficient estimator of $\psi(\hat{T})$ using the most *efficient influence function* of $\psi(\hat{T})$ as follows (Robins et al., 2008; Robins, 2004):

$$\hat{\psi}^*(\hat{T}) = \textstyle\sum_{i=1}^n \left(\frac{Y_i^{(W_i)} - (W_i - p(X_i)) \cdot \hat{T}(X_i)}{p(X_i) \cdot (1 - p(X_i))} \right)^2.$$

The derivation of the estimator above can be found in Theorem 9 in (Robins, 2004) and Section 5 in (Robins et al., 2008). When the propensity function $p(.)$ is known, this estimator approximate the PEHE at its optimal minimax rate. We estimate $p(.)$ via standard kernel density estimation methods. It can be easily shown using the results in (Dudoit & van der Laan, 2005) that when using the estimator above to tune the GP hyperparameters via cross-validation, then the learned length-scale parameters will satisfy the matching condition for minimax optimality.



(a) Impact of selection bias. (b) Impact of using over-smoothed priors. (c) Performance of different models.

Figure 2. Scatter-plots and linear fits for the PEHE of NSGP on a log-log scale in different simulation setups.

5. Experiments

In this Section, we check the validity of our analyses using a synthetic simulation setup (Subsection 5.1), and then evaluate the performance of our proposed model using data from a real-world clinical trial with simulated potential outcomes (Subsection 5.2). We will use the acronym **NSGP** to refer to the non-stationary GP model proposed in Section 4.

5.1. Learning Brownian Response Surfaces

5.1.1. Synthetic Model

Let $\mathcal{X} = [0, 1]$, and define a κ-fold integrated Brownian motion B_κ, $\kappa \in \mathbb{N}_+$, on $\mathcal{X}$ as follows:

$$B_\kappa(x) = \int_0^x \int_0^{x_\kappa} \cdots \int_0^{x_2} B_0(x_1)\, dx_1\, dx_2 \cdots dx_{x_\kappa},$$

where $B_0(.)$ is a standard Brownian motion (Wiener process). Sample paths of B_0 are almost surely Hölder regular with exponent $\frac{1}{2}$ (Karatzas & Shreve, 2012). Since $B_0(x)$ is almost surely non-differentiable everywhere in $\mathcal{X}$, then sample paths of $B_\kappa(x)$ are Hölder with exponent $\kappa + \frac{1}{2}$, i.e. $B_\kappa \in H^{\kappa+\frac{1}{2}}$ with probability 1. Therefore, when the true response surfaces are κ-fold integrated Brownian paths, the optimality and achievability results in Theorems 1 and 2 should hold. To this end, we simulate the true response surfaces $f_0 \in H^{\alpha_0}$ and $f_1 \in H^{\alpha_1}$ as $\boldsymbol{f_0} \sim \boldsymbol{B_{\alpha_0 - \frac{1}{2}}}$, and $\boldsymbol{f_1} \sim \boldsymbol{B_{\alpha_1 - \frac{1}{2}}}$, where we set $\boldsymbol{\alpha_0 = 2.5}$ and $\boldsymbol{\alpha_1 = 5.5}$.

The propensity score is modeled as a parametrized logistic function $\boldsymbol{p(x \mid \eta) = (1 + e^{-\eta\,(x - \frac{1}{2})})^{-1}}$, where $\eta \in \mathbb{R}$ is a parameter that determines the severity of selection bias. For a pair of fixed Brownian paths f_0 and f_1, synthetic observational samples $(\boldsymbol{X_i}, \boldsymbol{W_i}, \boldsymbol{Y_i^{(W_i)}})_i$ are generated as follows: $\boldsymbol{X_i} \sim \textbf{Uniform}[\boldsymbol{0, 1}]$, $\boldsymbol{W_i} \sim \textbf{Bernoulli}(\boldsymbol{p(x \mid \eta)})$, and $\boldsymbol{Y_i^{(W_i)}} \sim \boldsymbol{f_{W_i} + \mathcal{N}(0, \sigma^2)}$, where $\boldsymbol{\sigma^2 = 0.1}$.

5.1.2. Experiments and Results

Using the setup in Section 5.1.1, we conducted the following Monte Carlo simulations to verify our theoretical findings and highlight the merits of our NSGP model.

• ***Verifying Theorems 1 and 2*:** In order to check the validity of the results of Theorems 1 and 2, we use a NSGP Matérn prior $\mathcal{GP}(0, \mathbf{K}_\beta)$, with length-scale parameters β_0 and β_1 that are matched exactly with the regularities of the Brownian paths f_0 and f_1 (i.e. $\beta_0 = 2.5$ and $\beta_1 = 5.5$). According to Theorem 1, the optimal rate for estimating the CATE $T = f_1 - f_0$ is $n^{\frac{-5}{6}}$, and from Theorem 2, the NSGP with $\beta_0 = 2.5$ and $\beta_1 = 5.5$ should achieve that rate.

Figure 2a provides a scatter-plot for the PEHE achieved by the NSGP with respect to the number of samples on a log-log scale for different settings of η. We fit a linear regression model that describes the PEHE behavior in the log-log scale. We found the slope of the linear fit to be 0.8437, which is very close[2] to the slope of $\frac{5}{6} \approx 0.833$ predicted by Theorem 1. Moreover, by changing the magnitude of η from 0 to $\frac{1}{2}$, the PEHE curve did not exhibit any significant change in its slope, and was only moved upwards by a constant offset. On the contrary, Figure 2b shows the PEHE behavior when the NSGP prior is over-smoothed ($\beta_0 > \alpha_0$) for $\eta = 0$: as predicted by Theorem 2, learning becomes sluggish (slopes become less steep) as β_0 increase since the matching condition in (10) does not hold any more.

• ***NSGPs do not leave any money on the table:*** In this experiment, we show that the different components of the NSGP model allow it to perform well in small and large sample regimes. We set a strong selection bias of $\eta = \frac{1}{2}$ and compare the log(PEHE) characteristic of NSGP with a model that uses the same non-stationary kernel as NSGP, and another model that uses a standard stationary kernel, but both models are tuned using marginal likelihood maximization. As we can see in Figure 2c, the model with the non-stationary kernel achieves the same learning rate as NSGP, but exhibits a large offset as it does not account for selection bias, whereas the stationary model fails to learn the smoothness of the rougher Brownian motion since it assigns the same length-scale to both surfaces, and hence it over-smooths the prior, achieving a suboptimal rate.

[2]The minor discrepancy is a result of the residual error in the linear regression fit.

[Q1]	[Q2]	Model	In-sample $\sqrt{\text{PEHE}}$	Out-of-sample $\sqrt{\text{PEHE}}$	[Q1]	[Q2]	Model	In-sample $\sqrt{\text{PEHE}}$	Out-of-sample $\sqrt{\text{PEHE}}$
✓	✓	**NSGP**	**0.51 ± 0.013**	**0.64 ± 0.030**	✓		T-XGBoost	1.46 ± 0.081	1.98 ± 0.152
		SGP	0.95 ± 0.021	1.21 ± 0.052			S-XGBoost	2.97 ± 0.211	3.04 ± 0.216
✓	✓	CMGP	0.61 ± 0.011	0.76 ± 0.012	✓		T-AdaBoost	2.40 ± 0.177	2.79 ± 0.212
✓		TARNet	0.88 ± 0.021	0.95 ± 0.025			S-AdaBoost	4.53 ± 0.317	4.56 ± 0.312
	✓	BNN	2.21 ± 0.115	2.15 ± 0.125	✓		T-OLS	1.85 ± 0.107	1.94 ± 0.122
✓	✓	CFR Wass.	0.71 ± 0.018	0.76 ± 0.032			S-OLS	5.06 ± 0.357	5.05 ± 0.352
✓	✓	CFR MMD	0.73 ± 0.021	0.78 ± 0.022	✓		T-DNN	3.36 ± 0.137	3.46 ± 0.142
✓		T-Random Forest	1.41 ± 0.071	2.21 ± 0.162			S-DNN	3.56 ± 0.217	3.64 ± 0.212
		S-Random Forest	2.72 ± 0.241	2.91 ± 0.252		✓	MARS	1.66 ± 0.106	1.74 ± 0.112
	✓	Causal Forest	2.41 ± 0.141	2.82 ± 0.181			k-NN	2.69 ± 0.177	4.06 ± 0.212
		BART	2.00 ± 0.141	2.22 ± 0.151		✓	PSM	4.92 ± 0.312	4.92 ± 0.312
	✓	BCF	1.31 ± 0.061	1.71 ± 0.102		✓	TMLE	5.27 ± 0.357	5.27 ± 0.352

Table 1. Simulation results for the IHDP dataset. The values reported correspond to the average PEHE ($\pm$ 95% confidence intervals).

5.2. The Infant Health and Development Program

We evaluated the performance of the NSGP model presented in Section 4.1 using the standard semi-synthetic experimental setup designed by Hill in (Hill, 2011). We report a state-of-the-art result in this setup, and draw connections between our experimental results and our analyses.

5.2.1. Data and Benchmarks

The Infant Health and Development Program (IHDP) is an interventional program intended to enhance the health of premature infants (Hill, 2011). (Hill, 2011) extracted features and treatment assignments from a real-world clinical trial, and introduced selection bias to the data artificially by removing a subset of the patients. The potential outcomes are simulated according to the standard non-linear "Response Surface B" setting in (Hill, 2011). The dataset comprised 747 subjects, with 25 features for each subject. Our experimental setup is identical to (Hill, 2011; Johansson et al., 2016; Shalit et al., 2017; Alaa & van der Schaar, 2017): we run 1000 experiments in which we compute the in-sample and out-of-sample $\sqrt{\text{PEHE}}$ (with 80/20 training/testing splits), and report average results in Table 1.

We compared the performance of NSGP with a total of 23 CATE estimation benchmarks. We considered: tree-based algorithms (BART (Hill, 2011), Causal forests (Wager & Athey, 2017), Bayesian causal forests (Hahn et al., 2017)), methods based on deep learning (CFR Wass., CFR MMD, BNN, TARnet (Shalit et al., 2017)), multivariate additive regression splines (MARS) (Powers et al., 2017), Gaussian processes (CMGP) (Alaa & van der Schaar, 2017), nearest neighbor matching (k-NN), propensity score matching (PSM), and targeted maximum likelihood (TMLE) (Porter et al., 2011). We also composed a number of T-learners and S-learners as in (Künzel et al., 2017), using a variety of baseline machine learning algorithms (DNN stands for deep networks and OLS stands for linear regression).

5.2.2. Results and Conclusions

As we can see in Table 1, the proposed NSGP model significantly outperforms **all** competing benchmarks. The combined benefit of the two components of an NSGP (non-stationary kernel and doubly-robust hyperparameters) is highlighted by comparing its performance to a vanilla SGP (stationary GP) with marginal likelihood maximization. The gain with respect to such a model is a 2-fold improvement in the PEHE.

Because the IHDP dataset has a "moderate" sample size, both selection bias and learning rate seem to impact the performance. Thus, our method took advantage of having addressed modeling questions **[Q1]** and **[Q2]** appropriately by being both "rate-optimal" and "bias-aware".

The check marks in columns **[Q1]** and **[Q2]** designate methods that address modeling questions **[Q1]** and **[Q2]** "appropriately" in the light of the analysis presented in Section 3. Methods with **[Q1]** checked use a regression structure with "outcome-specific" hyperparameters, and methods with **[Q2]** checked adjust for selection bias. A general observation is that the structure of the regression model seem to matter much more than the strategy for handling selection bias. This is evident from the fact that the TARnet model (does not handle bias but models outcomes separately) significantly outperforms BNN (handles bias but uses a single-surface model (Shalit et al., 2017)), and that all T-learners (models 2 separate response surfaces) outperformed their S-shaped counterparts (models a single surface). For parametric models, such as OLS, the issue of selecting the right regression structure is even more crucial.

To sum up, the results in Table 1 imply that selecting the right regression structure is crucial for rate-optimality in sufficiently large dataset, whereas handling selection bias provides an extra bonus. In Table 1, methods that address both **[Q1]** and **[Q2]** (NSGP, CMGP, and CFR. Wass and MMD) displayed a superior performance.

References

Alaa, Ahmed M and van der Schaar, Mihaela. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Alaa, Ahmed M and van der Schaar, Mihaela. Bayesian nonparametric causal inference: Information rates and learning algorithms. *Journal on Selected Topics in Signal Processing*, 2018.

Alvarez, Mauricio A, Rosasco, Lorenzo, Lawrence, Neil D, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3): 195–266, 2012.

Atan, O, Jordan, J, and van der Schaar, M. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. *AAAI*, 2018.

Bhattacharyya, A. On some analogues of the amount of information and their use in statistical estimation. *Sankhyā: The Indian Journal of Statistics*, pp. 1–14, 1946.

Bickel, Peter J, Klaassen, Chris A, Bickel, Peter J, Ritov, Y, Klaassen, J, Wellner, Jon A, and Ritov, YA'Acov. *Efficient and adaptive estimation for semiparametric models*, volume 2. Springer New York, 1998.

Borrajo, Maria Isabel, González-Manteiga, Wenceslao, and Martínez-Miranda, María Dolores. Bandwidth selection for kernel density estimation with length-biased data. *Journal of Nonparametric Statistics*, 29(3):636–668, 2017.

Bottou, Léon, Peters, Jonas, Quiñonero-Candela, Joaquin, Charles, Denis X, Chickering, D Max, Portugaly, Elon, Ray, Dipankar, Simard, Patrice, and Snelson, Ed. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Castillo, I. Lower bounds for posterior rates with gaussian process priors. *Electron. J. Stat.*, 2:1281–1299, 2008.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2012.

Dudoit, Sandrine and van der Laan, Mark J. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2 (2):131–154, 2005.

Foster, Jared C, Taylor, Jeremy MG, and Ruberg, Stephen J. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Hahn, P Richard, Murray, Jared S, and Carvalho, Carlos M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. 2017.

Heckman, James J. Sample selection bias as a specification error (with an application to the estimation of labor supply functions), 1977.

Hill, Jennifer L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Huang, Jiayuan, Gretton, Arthur, Borgwardt, Karsten M, Schölkopf, Bernhard, and Smola, Alex J. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pp. 601–608, 2007.

Johansson, Fredrik, Shalit, Uri, and Sontag, David. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Karatzas, Ioannis and Shreve, Steven. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

Kennedy, Edward H. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 2018.

Künzel, Sören, Sekhon, Jasjeet, Bickel, Peter, and Yu, Bin. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

Li, Sheng and Fu, Yun. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pp. 930–940, 2017.

Lu, Min, Sadiq, Saad, Feaster, Daniel J, and Ishwaran, Hemant. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 2017.

Porter, Kristin E, Gruber, Susan, Van Der Laan, Mark J, and Sekhon, Jasjeet S. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, 7(1):1–34, 2011.

Powers, Scott, Qian, Junyang, Jung, Kenneth, Schuler, Alejandro, Shah, Nigam H, Hastie, Trevor, and Tibshirani, Robert. Some methods for heterogeneous treatment effect estimation in high-dimensions. *arXiv preprint arXiv:1707.00102*, 2017.

Raskutti, Garvesh, Yu, Bin, and Wainwright, Martin J. Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems*, pp. 1563–1570, 2009.

Rasmussen, Carl Edward and Williams, Christopher KI. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

Robins, James, Li, Lingling, Tchetgen, Eric, van der Vaart, Aad, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics, 2008.

Robins, James M. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pp. 189–326. Springer, 2004.

Rosenbaum, Paul R and Rubin, Donald B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

Rubin, Donald B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Shalit, Uri, Johansson, Fredrik, and Sontag, David. Estimating individual treatment effect: generalization bounds and algorithms. pp. 3076–3085, 2017.

Shimodaira, Hidetoshi. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90 (2):227–244, 2000.

Sniekers, Suzanne, van der Vaart, Aad, et al. Adaptive bayesian credible sets in regression with a gaussian process prior. *Electronic Journal of Statistics*, 9(2):2475–2527, 2015.

Stone, Charles J. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Sugiyama, Masashi and Storkey, Amos J. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems*, pp. 1337–1344, 2007.

Sugiyama, Masashi, Krauledat, Matthias, and MÃžller, Klaus-Robert. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

Vaart, Aad van der and Zanten, Harry van. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

Van der Vaart, Aad W. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

van der Vaart, Aad W and van Zanten, J Harry. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, pp. 1435–1463, 2008.

van der Vaart, Aad W, van Zanten, J Harry, et al. Reproducing kernel hilbert spaces of gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics, 2008.

Wager, Stefan and Athey, Susan. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Xie, Yu, Brand, Jennie E, and Jann, Ben. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

Yang, Yuhong and Barron, Andrew. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pp. 1564–1599, 1999.

Yang, Yun, Tokdar, Surya T, et al. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.

Zhang, Qinghua. Using wavelet network in nonparametric estimation. *IEEE Transactions on Neural networks*, 8 (2):227–236, 1997.

6. Supplementary Material

6.1. Proof of Theorem 1

Let δ_ω be the solution to $H(\delta_\omega; \mathcal{F}^{\alpha_\omega}) \asymp n\,\delta_\omega^2$, where $H(\delta_\omega; \mathcal{F}^{\alpha_\omega})$ is the metric entropy of the function space $\mathcal{F}^{\alpha_\omega}$. We will prove that the optimal rate is $\Theta(\delta_0^2 \vee \delta_1^2)$ by first showing that $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1})$ is lower bounded by, i.e. $\psi(\hat{T}) = \Omega(\delta_0^2 \vee \delta_1^2)$, and then show that $\psi(\hat{T}) = O(\delta_0^2 \vee \delta_1^2)$. We start by observing that the causal inference problem can be described through the following Markov chain

$$(f_0, f_1) \to \mathcal{D}_n \to (\hat{f}_0, \hat{f}_1) \to \hat{T}.$$

The amount of information shared between the true function $T(.)$ and the estimate $\hat{T}(.)$ can be quantified by the *mutual information* $I(T; \hat{T})$. Given the Markov chain above, we can upper bound $I(T; \hat{T})$ as follows

$$I(T; \hat{T}) \overset{(*)}{\leq} I(T; \mathcal{D}_n) \overset{(\star)}{\leq} \sup_{\Pi} I(T; \mathcal{D}_n), \tag{11}$$

where $(*)$ follows from the *data processing* inequality (Cover & Thomas, 2012), and the supremum in $(\star)$ is taken over all possible priors. $I(T; \hat{T})$ is bounded below by the *rate-distortion* function

$$I(T; \hat{T}) \geq \inf_{T, \hat{T}: \mathbb{E}\|T - \hat{T}\|_2^2 \leq R_{\Pi}^*} I(T; \hat{T}), \tag{12}$$

for any $\hat{T}$ satisfying $\mathbb{E}\|T - \hat{T}\|_2^2 \leq R_{\Pi}^*$, where the infimum is taken over all joint distributions of $(T, \hat{T})$. Combining (11) and (12), we can upper and lower bound the mutual information $I(T; \hat{T})$ as follows

$$\inf_{\mathbb{E}\|T - \hat{T}\|_2^2 \leq R_{\Pi}^*} I(T; \hat{T}) \leq I(T; \hat{T}) \leq \sup_{\Pi} I(T; \mathcal{D}_n). \tag{13}$$

The lower bound in the chain of inequalities above is intractable, and hence we further lower bound $I(T; \hat{T})$ using *Fano's* method (**?**Yang & Barron, 1999). That is, we take discrete subsets $\tilde{\mathcal{F}}^{\alpha_0}$ and $\tilde{\mathcal{F}}^{\alpha_1}$ of the function spaces $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$, and convert the estimation problem to a testing problem. The spaces

$$\tilde{\mathcal{F}}^{\alpha_\omega} = \{\tilde{f}_\omega^1, \ldots, \tilde{f}_\omega^{\tilde{M}_\omega}\},\ \tilde{\mathcal{F}}^{\alpha_\omega} \subset \mathcal{F}^{\alpha_\omega},\ \omega \in \{0, 1\},$$

are constructed such that $\|\tilde{f}_\omega^i - \tilde{f}_\omega^j\| \geq \delta, \forall i \neq j$. Let Q be a quantizer that maps elements of $\mathcal{F}^{\alpha_\omega}$ to $\tilde{\mathcal{F}}^{\alpha_\omega}$, $\omega \in \{0, 1\}$. Thus, the causal inference problem can be described through the following Markov chain:

$$(f_0, f_1) \to \mathcal{D}_n \to (\hat{f}_0, \hat{f}_1) \to Q(\hat{f}_0, \hat{f}_1). \tag{14}$$

Let $\tilde{T} = \tilde{f}_1^u - \tilde{f}_0^v$, where $\tilde{f}_0^v$ and $\tilde{f}_1^u$ are the functions in $\tilde{\mathcal{F}}^{\alpha_0}$ and $\tilde{\mathcal{F}}^{\alpha_1}$ that are closest to f_0 and f_1. The discrete element $\tilde{T}$ belongs to a set $\{\tilde{T}^1, \ldots, \tilde{T}^{\tilde{M}_T}\}$, which corresponds to a discretized version of the function space to which T belongs. Using the data processing inequality, we have that

$$I(\tilde{T}; \hat{T}) \geq I(\tilde{T}; Q(\hat{T})). \tag{15}$$

An "error event" is an event where $Q(\hat{T})$ does not correspond to the true discretized function $\tilde{T}$, i.e. the event $\{\tilde{T} \neq Q(\hat{T})\}$. The error event occurs when

$$\|\hat{T} - Q(\hat{T})\| \leq \|\hat{T} - \tilde{T}\|,\ \{\tilde{T} \neq Q(\hat{T})\}. \tag{16}$$

Thus, the error event implies that $\delta \leq \|Q(\hat{T}) - \tilde{T}\|$. Using the triangular inequality, (16) can be further bounded as follows:

$$\begin{aligned}
\delta &\leq \|Q(\hat{T}) - \tilde{T}\| = \|Q(\hat{T}) - \hat{T} + \hat{T} - \tilde{T}\| \\
&\leq \|Q(\hat{T}) - \hat{T}\| + \|\hat{T} - \tilde{T}\| \\
&\leq 2\,\|\hat{T} - \tilde{T}\| \implies \|\hat{T} - \tilde{T}\| \geq \frac{\delta}{2}.
\end{aligned} \tag{17}$$

Let P_e be the probability of the error event $\{\tilde{T} \neq Q(\hat{T})\}$. From (17), P_e can be bounded above as follows

$$\begin{aligned}
P_e &= \mathbb{P}(\{\tilde{T} \neq Q(\hat{T})\}) \\
&= \mathbb{P}(\|Q(\hat{T}) - \tilde{T}\| \geq \delta) = \mathbb{P}(\|\hat{T} - \tilde{T}\| \geq \delta/2) \\
&= \mathbb{P}(\|\hat{T} - \tilde{T}\|_2^2 \geq \delta^2/4) \\
&\overset{(\bullet)}{\leq} \frac{4}{\delta^2}\,\mathbb{E}[\|\hat{T} - \tilde{T}\|_2^2] \leq \frac{4}{\delta^2}\,R_{\Pi}^*,
\end{aligned} \tag{18}$$

where $(\bullet)$ is an application of Markov's inequality. By combining (15) with the result in (18), the lower bound in (13) can be further bounded below as follows

$$\begin{aligned}
\inf_{\mathbb{E}\|T - \hat{T}\|_2^2 \leq R_{\Pi}^*} I(T; \hat{T}) &\geq \inf_{\mathbb{E}\|T - \hat{T}\|_2^2 \leq R_{\Pi}^*} I(\tilde{T}; \hat{T}) \\
&= \inf_{P_e \leq \frac{4}{\delta^2} R_{\Pi}^*} I(\tilde{T}; \hat{T}) \\
&\geq \inf_{P_e \leq \frac{4}{\delta^2} R_{\Pi}^*} I(\tilde{T}; Q(\hat{T})).
\end{aligned}$$

The mutual information $I(\tilde{T}; Q(\hat{T}))$ can be bounded above as follows

$$\begin{aligned}
I(\tilde{T}; Q(\hat{T})) &= I(\tilde{f}_1 - \tilde{f}_0; Q(\hat{f}_1 - \hat{f}_0)) \\
&\overset{(\odot)}{\leq} I(\tilde{f}_0, \tilde{f}_1; Q(\hat{f}_1 - \hat{f}_0)) \\
&\leq I(\tilde{f}_0, \tilde{f}_1; Q(\hat{f}_0), Q(\hat{f}_1)) \\
&= I(\tilde{f}_0; Q(\hat{f}_0)) + I(\tilde{f}_1; Q(\hat{f}_1)) \\
&\leq 2 \max\{I(\tilde{f}_0; Q(\hat{f}_0)), I(\tilde{f}_1; Q(\hat{f}_1))\},
\end{aligned} \tag{19}$$

where $(\odot)$ follows from the data processing inequality. Note that the mutual information $I(\tilde{T}; Q(\hat{T}))$ can be written in terms of the KL divergence as (Cover & Thomas, 2012)

$$\begin{aligned}
I(\tilde{T}; Q(\hat{T})) &= D(\mathbb{P}(\tilde{T}; Q(\hat{T})) \,||\, \mathbb{P}(\tilde{T}) \cdot \mathbb{P}(Q(\hat{T}))) \\
&\geq D(\text{Bern}(P_e) \,||\, \text{Bern}(1 - 1/n)) \\
&= P_e \log\left(\frac{P_e}{1 - 1/\tilde{M}_T}\right) + (1 - P_e) \log\left(\frac{1 - P_e}{1/\tilde{M}_T}\right) \\
&= -h(P_e) + \log(\tilde{M}_T) - P_e\,\log(\tilde{M}_T - 1) \\
&\geq -\log(2) + \log(\tilde{M}_T) - P_e\,\log(\tilde{M}_T),
\end{aligned} \tag{20}$$

where $h(.)$ is the binary entropy. From (20), we have that

$$P_e \geq 1 - \frac{I(\tilde{T}; Q(\hat{T})) + \log(2)}{\log(\tilde{M}_T)}, \tag{21}$$

which is an incarnation of Fano's inequality. By combining (19) with (21), we have the following inequality

$$P_e \geq 1 - \frac{I(\tilde{f}_0; Q(\hat{f}_0)) \vee I(\tilde{f}_1; Q(\hat{f}_1)) + \log(\sqrt{2})}{\frac{1}{2}\,\log(\tilde{M}_T)}. \tag{22}$$

From (18), the minimax risk R^*_{Π} is bounded below by

$$R^*_{\Pi} \geq \frac{\delta^2}{4}\left(1 - \frac{I(\tilde{f}_0; Q(\hat{f}_0)) \vee I(\tilde{f}_1; Q(\hat{f}_1)) + \log(\sqrt{2})}{\frac{1}{2}\log(\tilde{M}_T)}\right).$$

The discretization $\tilde{\mathcal{F}}^{\alpha_\omega} = \{\tilde{f}^1_\omega, \ldots, \tilde{f}^{\tilde{M}_\omega}_\omega\}$ corresponds to a δ-packing of the function space $\mathcal{F}^{\alpha_\omega}$, and hence $\tilde{M}_\omega$ is given by the covering number $N(\delta, \mathcal{F}^{\alpha_\omega})$, for $\omega \in \{0, 1\}$. It follows that $\tilde{M}_T \geq N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1})$, and hence we have that

$$R^*_{\Pi} \geq \frac{\delta^2}{4}\left(1 - \frac{I(\tilde{f}_0; Q(\hat{f}_0)) \vee I(\tilde{f}_1; Q(\hat{f}_1)) + \log(\sqrt{2})}{\frac{1}{2}\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1}))}\right).$$

The mutual information $I(\tilde{f}_\omega; Q(\hat{f}_\omega))$ can be bounded via the KL divergence as

$$\begin{aligned} I(\tilde{f}_\omega; Q(\hat{f}_\omega)) &\leq \frac{1}{N^2(\delta, \mathcal{F}^{\alpha_\omega})} \sum_{i,j} D(\mathbb{P}(\tilde{f}^i_\omega) \,||\, \mathbb{P}(\tilde{f}^j_\omega)) \\ &\leq 2\, n\, \delta^2. \end{aligned}$$

Thus, the minimax risk can be bounded below as follows

$$R^*_{\Pi} \geq \frac{\delta^2}{4}\left(1 - \frac{4\, n\, \delta^2 + \log(2)}{\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1}))}\right),$$

and hence we have that

$$R^*_{\Pi} \gtrsim \delta^2 - \frac{\delta^4\, n + \delta^2}{\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1}))}. \tag{23}$$

Since R^*_{Π} is strictly positive, then we have that

$$R^*_{\Pi} \gtrsim \delta^2,$$

where δ is the solution to the transcendental equation

$$\delta^2 \asymp \frac{\delta^4\, n}{\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1}))},$$

or equivalently

$$\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1})) \asymp \delta^2\, n. \tag{24}$$

The metric entropy of a function space $\mathcal{F}^{\alpha_\omega}$ is given by $H(\delta, \mathcal{F}^{\alpha_\omega}) = \log(N(\delta, \mathcal{F}^{\alpha_\omega})$, and hence (24) is written as

$$H(\delta, \mathcal{F}^{\alpha_0}) \vee H(\delta, \mathcal{F}^{\alpha_1}) \asymp \delta^2\, n. \tag{25}$$

Since the metric entropy $H(\delta, \mathcal{F}^{\alpha_\omega})$ is a decreasing function of the smoothness parameter α_ω, then it follows that the solution δ^* of the transcendental equation in (25) is given by $\delta^* = \delta_0 \vee \delta_1$, where δ_ω is the solution to the equation

$$H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0, 1\}. \tag{26}$$

The equation in (26) has a solution for all n when the function space $\mathcal{F}^{\alpha_\omega}$ has a polynomial or a logarithmic metric entropy (Van der Vaart, 1998), which is the case for all function spaces of interest (see Table I for evaluations of $\delta_0 \vee \delta_1$ for various function spaces). It follows from (23) and (26) that

$$R^*_{\Pi} = \Omega(\delta^2_0 \vee \delta^2_1),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0, 1\},$$

and hence, from (**??**), we have that

$$\psi(\hat{T}) = \Omega(\delta^2_0 \vee \delta^2_1),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0, 1\}. \tag{27}$$

We now focus on upper bounding R^*_{Π}. From (**?**), we know that the minimax risk is upper bounded by the channel capacity in (11), which is further bounded above by the covering numbers as follows

$$R^*_{\Pi} \lesssim \frac{1}{n}\left(\log(N(\delta, \mathcal{F}^{\alpha_0})) \vee \log(N(\delta, \mathcal{F}^{\alpha_1})) + n\,\delta^2\right).$$

For δ satisfying (26), we have that

$$\log(N(\delta, \mathcal{F}^{\alpha_0})) \vee \log(N(\delta, \mathcal{F}^{\alpha_1})) = \delta^2\, n,$$

and hence $R^*_{\Pi} \lesssim \delta^2_0 \vee \delta^2_1$. It follows that

$$\psi(\hat{T}) = O(\delta^2_0 \vee \delta^2_1),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0, 1\}. \tag{28}$$

By combining (27) and (28), we have that $I^*_n = \Omega(\delta^2_0 \wedge \delta^2_1)$ and $I^*_n = O(\delta^2_0 \vee \delta^2_1)$, and hence it follows that

$$\psi(\hat{T}) = \Theta(\delta^2_0 \vee \delta^2_1),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0, 1\}. \tag{29}$$

For Hölder balls, $\delta_w = n^{\frac{-\alpha_w}{2\alpha_w + d_w}}$. The variable selection term follows straightforwardly from model $M2$ in (Yang et al., 2015).

6.2. Proof of Theorem 2

We start by providing the following Lemma, which we will use to prove the statement of the Theorem.

Lemma 1. The PEHE $\psi(\hat{T}) = \mathbb{E}\left[\,\|\,\hat{T} - T\,\|_2^2\,\right]$ is asymptotically bounded above as follows:

$$\psi(\hat{T}) \lesssim \mathbb{E}\left[\,\|\,\mathbb{E}[\,\hat{f}_0 - f_0\,\|_2^2\,\right] + \mathbb{E}\left[\,\|\,\mathbb{E}[\,\hat{f}_1 - f_1\,\|_2^2\,\right].$$

Proof. The PEHE $\psi(\hat{T})$ for a given observational dataset is given by:

$$\psi(\hat{T}) = \|\,(\hat{f}_1(x) - \hat{f}_0(x)) - (f_1(x) - f_0(x))\,\|_2^2, \tag{30}$$

where $\hat{f}_w(x) = \mathbb{E}[\,f_w(x)\,|\,\mathcal{D}_n\,]$, for $w \in \{0, 1\}$. The norm in (30) can be expressed as follows:

$$\begin{aligned}
\psi(\hat{T}) &= \|\,(\hat{f}_1(x) - \hat{f}_0(x)) - (f_1(x) - f_0(x))\,\|_2^2, \\
&= \int_{\mathcal{X}} ((\hat{f}_1(x) - \hat{f}_0(x)) - (f_1(x) - f_0(x)))^2\, d\mathbb{P}(x) \\
&= \int_{\mathcal{X}} ((\hat{f}_1(x) - f_1(x)) + (f_0(x) - \hat{f}_0(x)))^2\, d\mathbb{P}(x) \\
&\leq 2 \int_{\mathcal{X}} ((\hat{f}_1(x) - f_1(x))^2 + (\hat{f}_0(x) - f_0(x))^2)\, d\mathbb{P}(x) \\
&= 2 \int_{\mathcal{X}} (\hat{f}_1(x) - f_1(x))^2\, d\mathbb{P}(x, w = 1) \\
&+ 2 \int_{\mathcal{X}} (\hat{f}_0(x) - f_0(x))^2\, d\mathbb{P}(x, w = 0).
\end{aligned} \tag{31}$$

Thus, we have that

$$\begin{aligned}
\psi(\hat{T}) &= 2 \int_{\mathcal{X}} (\hat{f}_1(x) - f_1(x))^2\, p(x) \cdot d\mathbb{P}(x) \\
&+ 2 \int_{\mathcal{X}} (\hat{f}_0(x) - f_0(x))^2\, (1 - p(x)) \cdot d\mathbb{P}(x) \\
&= 2\,\|\sqrt{p(x)} \cdot (\hat{f}_1(x) - f_1(x))\|_{L_2(\mathbb{P})}^2 \\
&+ 2\,\|\sqrt{1 - p(x)} \cdot (\hat{f}_0(x) - f_0(x))\|_{L_2(\mathbb{P})}^2.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we obtain the following:

$$\|\sqrt{p(x)}(\hat{f}_1(x) - f_1(x))\|_2^2 \leq \|p(x)\|_2 \cdot \|(\hat{f}_1(x) - f_1(x))^2\|_2,$$

and similarly for $\|\sqrt{1 - p(x)} \cdot (\hat{f}_0(x) - f_0(x))\|_2^2$.

The proof of the Lemma is concluded by observing that $\|p(x)\|_2$ is $O(1)$ and

$$\|(\hat{f}_1(x) - f_1(x))^2\|_2 \asymp \|(\hat{f}_1(x) - f_1(x))\|_2^2. \tag{32}$$

The same can be arrived at via Minkowski inequality. □

The minimax rate achieved by the prior $\Pi(\beta_0, \beta_1)$ is upper bounded by the posterior contraction rates (van der Vaart & van Zanten, 2008) (rate of convergence of the $L^2(\mathbb{P})$ loss) over the surfaces f_0 and f_1. For a prior $\mathcal{GP}(\text{Matérn}(\beta_w))$ and a true function $f_\omega \in H^{\alpha_w}$, the contraction rate ε^2 is given by solving the following transcendental equation (Vaart & Zanten, 2011):

$$\phi_{f_\omega}(\varepsilon) \asymp n \cdot \varepsilon^2, \tag{33}$$

where $\phi_{f_\omega}(\varepsilon)$ is the *concentration function* defined as (van der Vaart & van Zanten, 2008):

$$\phi_{f_\omega}(\varepsilon) = -\log(\mathbb{P}_{\Pi(\beta_\omega)}(\|f - f_\omega\|_\infty < \varepsilon)). \tag{34}$$

The concentration function measures the amount of prior mass that Π places around the true function f_w. In Lemma 4 in (Vaart & Zanten, 2011), the concentration function $\phi_{f_w}(\varepsilon)$ for a sufficiently smooth prior $\mathcal{GP}(\text{Matérn}(\beta_w))$, with $\beta_w > d/2$, and a sufficiently smooth true function $f_w \in H^{\alpha_w}$, with $\beta_w > d/2$, was obtained as follows:

$$\phi_{f_w}(\varepsilon) \lesssim \varepsilon^{-\frac{d}{\beta_w}} + \varepsilon^{-\frac{2\beta_w - 2\alpha_w + d}{\alpha_w}}. \tag{35}$$

Thus, combining (33) and (35), the posterior contraction rate for $\Pi(\beta_w)$ around f_w is the solution to:

$$n \cdot \varepsilon^2 \lesssim \varepsilon^{-\frac{d}{\beta_w}} + \varepsilon^{-\frac{2\beta_w - 2\alpha_w + d}{\alpha_w}}, \tag{36}$$

The solution to (36) is given by

$$\begin{aligned}
\varepsilon &\lesssim n^{-\frac{\beta_w}{2\beta_w + d}} + n^{-\frac{\alpha_w}{2\beta_w + d}}, \\
&\asymp n^{-\frac{(\beta_w \wedge \alpha_w)}{2\beta_w + d}}.
\end{aligned} \tag{37}$$

That is, we can characterize the $L^2(\mathbb{P})$ loss surfaces on f_0 and f_1 as follows:

$$\mathbb{E}_{\mathcal{D}_n}\left[\,\|\,\mathbb{E}_\Pi[\,f_w\,|\,\mathcal{D}\,] - f_w\,\|_2^2\,\right] \lesssim n^{-\frac{(\beta_w \wedge \alpha_w)}{2\beta_w + d}}, \tag{38}$$

and so it follows that:

$$\psi(\hat{T}) \lesssim n^{-\frac{(\beta_0 \wedge \alpha_0)}{2\beta_0 + d}} \vee n^{-\frac{(\beta_1 \wedge \alpha_1)}{2\beta_1 + d}},$$

which concludes the proof of the Theorem.

6.3. Derivation of Equation (6)

From Lemma 1, we have that:

$$\|\sqrt{p(x)}(\hat{f}_1(x) - f_1(x))\|_2^2 \leq \|p(x)\|_2 \cdot \|(\hat{f}_1(x) - f_1(x))^2\|_2,$$

and similarly for $\|\sqrt{1-p(x)} \cdot (\hat{f}_0(x) - f_0(x))\|_2^2$. Thus, we can upper bound the PEHE as follows:

$$\begin{aligned}\psi(\hat{T}) \lesssim\ & \|p(x)\|_{L^2(\mathbb{P}_1)} \cdot \|(\hat{f}_1(x) - f_1(x))^2\|_{L^2(\mathbb{P}_1)}, \\ & + \|(1-p(x))\|_{L^2(\mathbb{P}_0)} \cdot \|(\hat{f}_0(x) - f_0(x))^2\|_{L^2(\mathbb{P}_0)}.\end{aligned}$$

Note that:

$$\begin{aligned}\|p(x)\|_{L^2(\mathbb{P}_1)} &= \int_{\mathcal{X}} p(x)\, d\mathbb{P}(X = x \,|\, W = 1) \\ &= \int_{\mathcal{X}} \mathbb{P}(W = 1 \,|\, X = x)\, d\mathbb{P}(X = x \,|\, W = 1) \\ &= \int_{\mathcal{X}} \frac{d\mathbb{P}(W = 1, X = x)}{d\mathbb{P}(X = x)}\, d\mathbb{P}(X = x \,|\, W = 1) \\ &= \int_{\mathcal{X}} \frac{d\mathbb{P}(X = x \,|\, W = 1), \mathbb{P}(W = 1)}{d\mathbb{P}(X = x)}\, d\mathbb{P}(X = x \,|\, W = 1) \\ &= \mathbb{P}(W = 1) \cdot \int_{\mathcal{X}} \frac{d\mathbb{P}^2(X = x \,|\, W = 1)}{d\mathbb{P}(X = x)} \\ &= \mathbb{P}(W = 1) \cdot \exp\left(\log\left(\int_{\mathcal{X}} \frac{d\mathbb{P}^2(X = x \,|\, W = 1)}{d\mathbb{P}(X = x)}\right)\right) \\ &= \mathbb{P}(W = 1) \cdot \exp\left(D_2(d\mathbb{P}(X = x \,|\, W = 1) \| d\mathbb{P}(X = x))\right) \\ &= \mathbb{P}(W = 1) \cdot \exp\left(D_2(Q_1 \| Q)\right).\end{aligned}$$

The same can be shown for $\|1 - p(x)\|_{L^2(\mathbb{P}_0)}$, and the bound in (6) follows.