Bayesian Nonparametric Causal Inference: Information Rates and Learning Algorithms

Ahmed M. Alaa, *Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—We investigate the problem of estimating the causal effect of a treatment on *individual* subjects from observational data; this is a central problem in various application domains, including healthcare, social sciences, and online advertising. Within the *Neyman-Rubin potential outcomes* model, we use the Kullback-Leibler (KL) divergence between the estimated and true distributions as a measure of accuracy of the estimate, and we define the information rate of the Bayesian causal inference procedure as the (asymptotic equivalence class of the) expected value of the KL divergence between the estimated and true distributions as a function of the number of samples. Using Fano's method, we establish a fundamental limit on the information rate that can be achieved by *any* Bayesian estimator, and show that this fundamental limit is independent of the *selection bias* in the observational data. We characterize the Bayesian priors on the potential (factual and counterfactual) outcomes that achieve the optimal information rate. As a consequence, we show that a particular class of priors that have been widely used in the causal inference literature cannot achieve the optimal information rate. On the other hand, a broader class of priors can achieve the optimal information rate. We go on to propose a prior adaptation procedure (which we call the *information-based empirical Bayes* procedure) that optimizes the Bayesian prior by maximizing an information-theoretic criterion on the recovered causal effects rather than maximizing the marginal likelihood of the observed (factual) data. Building on our analysis, we construct an information-optimal Bayesian causal inference algorithm. This algorithm embeds the potential outcomes in a *vector-valued reproducing kernel Hilbert space* (vvRKHS), and uses a multi-task Gaussian process prior over that space to infer the individualized causal effects. We show that for such a prior, the proposed information-based empirical Bayes method adapts the smoothness of the *multi-task Gaussian process* to the true smoothness of the causal effect function by balancing a tradeoff between the *factual bias* and the *counterfactual variance*. We conduct experiments on a well-known real-world dataset and show that our model significantly outperforms the state-of-the-art causal inference models.**

***Index Terms*—Bayesian nonparametrics, causal effect inference, Gaussian processes, multitask learning, selection bias.**

I. Introduction

THE problem of estimating the *individualized* causal effect of a particular intervention from *observational data* is central in many application domains and research fields, including public health and healthcare [1], computational advertising [2], and social sciences [3]. With the increasing availability of data in all these domains, machine learning algorithms can be used to obtain estimates of the effect of an intervention, an action, or a treatment on individuals given their features and traits. For instance, using observational electronic health record data[1], machine learning-based recommender system can learn the *individual-level* causal effects of treatments currently deployed in clinical practice and help clinicians refine their current treatment policies [4]. There is a growing interest in using machine learning methods to infer the individualized causal effects of medical treatments; this interest manifests in recent initiatives such as STRATOS [4], which focuses on guiding observational medical research, in addition to various recent works on causal effect inference by the machine learning community [5]–[10].

The problem of estimating individual-level causal effects is usually formulated within the classical *potential outcomes* framework, developed by Neyman and Rubin [11], [12]. In this framework, every subject (individual) in the observational dataset possesses two "potential outcomes": the subject's outcome under the application of the treatment, and the subject's outcome when no treatment is applied. The treatment effect is the difference between the two potential outcomes, but since we only observe the "factual" outcome for a specific treatment assignment, and never observe the corresponding "counterfactual" outcome, we never observe any samples of the true treatment effect in an observational dataset. *This is what makes the problem of causal inference fundamentally different from standard supervised learning (regression).* Moreover, the policy by which treatments are assigned to subjects induces a *selection bias* in the observational data, creating a discrepancy in the feature distributions for the *treated* and *control* patient groups, which makes the problem even harder. Many of the classical works on causal inference have focused on the simpler problem of estimating *average* treatment effects, where unbiased estimators based on propensity score weighting were developed to alleviate the impact of selection bias on the causal estimands (see [13] and the references therein).

While more recent works have developed machine learning algorithms for estimating individualized treatment effects from observational data in the past few years [2], [5], [8], [14]–[19], the inference machinery built in most of these works seem to be rather ad-hoc. The causal inference problem entails a richer set of modeling choices and decisions compared to that of the standard supervised learning (regression) problem, which includes deciding what model to use, how to model the treatment assignment variables in the observational data, and how to handle selection bias, etc. In order to properly address

A. Alaa is with the Department of Electrical Engineering, University of California Los Angeles (UCLA), Los Angeles, CA, 90095, USA (e-mail: ahmedmalaa@ucla.edu).

M. van der Schaar is with the Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK (e-mail: mihaela.vanderschaar@eng.ox.ac.uk).

Section VII of this work was presented in part at the thirty-first annual conference on Neural Information Processing Systems (NIPS), 2017.

[1] https://www.healthit.gov/sites/default/files/briefs/

all these modeling choices, one needs to understand the fundamental limits of performance in causal effect estimation problems, and how different modeling choices impact the achievable performance.

In this paper, we establish the fundamental limits on the amount of information that a learning algorithm can gather about the causal effect of an intervention given an observational data sample. We also provide guidelines for building proper causal inference models that "do not leave any information on the table" because of poor modeling choices. A summary of our results is provided in the following Section.

II. Summary of the Results

We address the individualized causal effect estimation problem on the basis of the Neyman-Rubin potential outcomes model [11], [12]. We focus on Bayesian nonparametric learning algorithms, as they are immune to model mis-specification, and can learn highly heterogeneous response functions that one would expect to encounter in datasets with medical or social outcomes [3], [20]. In Section IV, we introduce the notion of *information rate* as a measure for the quality of Bayesian nonparametric learning of the individualized causal effects. The information rate is defined in terms of a measure of the Kullback-Leibler divergence between the true and posterior distributions for the causal effect. In Theorem 1, we establish the equivalence between Bayesian information rates and frequentist estimation rate. In the rest of the paper, we characterize: (1) the optimal information rates that can be achieved by any Bayesian nonparametric learning algorithm, and (2) the nature of the priors that would give rise to "informationally optimal" Bayesian nonparametric causal inference procedure.

In Section V, we establish the fundamental limit on the information rate that can be achieved by any Bayesian causal inference procedure using an information-theoretic lower bound based on Fano's method. The optimal information rate is a property of the function classes to which the potential outcomes belong, and is independent of the inference algorithm. We show that the optimal information rate for causal inference is governed by the "rougher" of the two potential outcomes functions. We also show that the optimal information rates for causal inference are *insensitive* to selection bias (Theorem 2).

In Section VI, we characterize the Bayesian priors that achieve the optimal rate. We show that the most common modeling choice adopted in the literature, which is to augment the treatment assignment variable to the feature space, leads to priors that are suboptimal in terms of the achievable rate (Theorem 3). We show that informationally optimal priors are ones that place a probability distribution over a vector-valued function space, where the function space has its smoothness matching the rougher of the two potential outcomes functions. Since the true smoothness parameter of the potential outcomes functions is generally unknown a priori, we propose a prior adaptation procedure, called the *information-based empirical Bayes* procedure, which optimizes the Bayesian prior by maximizing an information-theoretic criterion on the recovered causal effects rather than maximizing the marginal likelihood of the observed (factual) data.

We conclude the paper by building an information-optimal Bayesian causal inference algorithm that is based on our analysis. The inference procedure embeds the potential outcomes in a *vector-valued reproducing kernel Hilbert space* (vvRKHS), and uses a *multi-task Gaussian process* prior (with a Matérn kernel) over that space to infer the individualized causal effects. We show that for such a prior, the proposed information-based empirical Bayes method exhibits an insightful *factual bias* and *counterfactual variance* decomposition. Experiments conducted on a standard dataset that is used for benchmarking causal inference models show that our model significantly outperforms the state-of-the-art.

III. Related Work

We conduct our analysis within the *potential outcomes* framework developed by Neyman and Rubin [11], [12]. The earliest works on estimating causal effects have focused on the problem of obtaining unbiased estimates for the *average* treatment effects using observational samples. The most common well-known estimator for the average causal effect of a treatment is the propensity score weighting estimator, which simply removes the bias introduced by selection bias by giving weights to different samples that are inversely proportional to their propensity scores [13]. More recently, the machine learning community has also developed estimators for the average treatment effects that borrows ideas from representation learning, i.e. see for instance the work in [9]. In this paper, we focus on the *individual*, rather than the *average* causal effect estimation problem.

To the best of our knowledge, non of the previous works have attempted to characterize the limits of learning causal effects in either the frequentist or Bayesian setups. Instead, most previous works on causal effect inference have focused on model development, and various algorithms have been recently developed for estimating individualized treatment effects from observational data, mostly based on either tree-based methods [7], [8], [16], or deep learning methods [14], [15]. Most of the models that were previously developed for estimating causal effects relied on regression models that treat the treatment assignment variables (i.e. whether or not the intervention was applied to the subject) as an extended dimension in the feature space. Examples of such models include Bayesian additive regression trees (BART) [8], causal forests [7], balanced counterfactual regression [18], causal multivariate additive regression splines (MARS) [19], propensity-dropout networks [15], or random forests [21]. In all these methods, augmenting the treatment assignment variable to the feature space introduces a mismatch between the training and testing distribution (i.e. covariate shift induced by the selection bias [18]). The different methods followed different approaches for handling the selection bias: causal forests use estimates of the propensity score for deriving a tree splitting rule that attempts to balance the treated and control populations, propensity-dropout networks use larger dropout regularization for training points with very high or very low propensity scores, whereas balanced counterfactual regression uses deep neural networks to learn a balanced representation

(i.e. a feature transformation) that tries to alleviate the effect of the selection bias. Bayesian methods, like BART, do not address selection bias since the Bayesian posterior naturally incorporates uncertainty in regions of poor overlap in the feature space. As we show later in Sections VI and VIII, our analysis and experimental results indicated that, by augmenting the treatment assignment variable to the feature space, all these methods achieve a suboptimal information rate.

Our analysis is related to a long strand of literature that studied frequentist (minimax) estimation rates, or posterior contraction rates in standard regression problems [22]–[26]. In Theorem 2, we show that the optimal information rate for causal inference has the same form as the optimal minimax estimation rate obtained by Stone in [25] for standard nonparametric regression problems, when the true regression function is set to be the rougher of the two potential outcomes functions. Our analysis for the achievable information rates for Gaussian process priors uses the results by van Zanten and van der Vaart in [27].

IV. Bayesian Nonparametric Causal Inference from Observational Data

In this section, we provide a general description for the Neyman-Rubin causal model considered in this paper (Subsection IV-A), and present the Bayesian nonparametric inference framework under study (Subsection IV-B).

A. The Neyman-Rubin Causal Model

Consider a population of subjects with each subject i possessing a d-dimensional *feature* $X_i \in \mathcal{X}$. An intervention is applied to some subjects in the population: subject i's response to the intervention is a random variable denoted by $Y_i^{(1)}$, whereas the subject's natural response when no intervention is applied is denoted by $Y_i^{(0)}$. The two random variables, $Y_i^{(1)}, Y_i^{(0)} \in \mathbb{R}$, are known as the *potential outcomes*. The causal effect of the intervention (treatment) on subject i is characterized through the difference between the two (random) potential outcomes $(Y_i^{(1)} - Y_i^{(0)}) \,|\, X_i = x$, and is generally assumed to be dependent on the subject's features $X_i = x$. Hence, we define the *individualized treatment effect* (ITE) for a subject i with a feature $X_i = x$ as

$$T(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \,\middle|\, X_i = x \right]. \tag{1}$$

Our goal is to estimate the function $T(x)$ from an *observational* dataset $\mathcal{D}_n$, which comprises n independent samples of the random tuple $\{X_i, \omega_i, Y_i^{(\omega_i)}\}$, where $\omega_i \in \{0,1\}$ is an intervention assignment indicator that indicates whether or not subject i has received the intervention (treatment) under consideration. The outcomes $Y_i^{(\omega_i)}$ and $Y_i^{(1-\omega_i)}$ are known in the literature as the *factual* and the *counterfactual* outcomes, respectively [18], [28]. Intervention assignments generally depend on the subjects' features, i.e. $\omega_i \not\perp\!\!\!\perp X_i$. This dependence is quantified via the conditional distribution $\mathbb{P}(\omega_i = 1 | X_i = x)$, also known as the *propensity score* of subject i [13], [11]. In the rest of this paper, we denote the propensity score of a feature point x as $\gamma(x)$.

The observational dataset $\mathcal{D}_n = \{X_i, \omega_i, Y_i^{(\omega_i)}\}_{i=1}^n$ is drawn from a joint density $d\mathbb{P}(X_i, \omega_i, Y_i^{(0)}, Y_i^{(1)})$, with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports the following standard conditions [11], [12]:

- **Condition 1 (unconfoundedness):** Treatment assignment decisions are independent of the outcomes given the subject's features, i.e. $(Y_i^{(0)}, Y_i^{(1)}) \perp\!\!\!\perp \omega_i \,|\, X_i$.
- **Condition 2 (overlap):** Every subject has a non-zero chance of receiving the treatment, and treatment assignment decisions are non-deterministic, i.e. $0 < \gamma(x) < 1$.

B. Bayesian Nonparametric Causal Inference

Throughout this paper, we consider the following *signal-in-white-noise* random design regression model for the potential outcomes:

$$Y_i^{(\omega)} = f_\omega(X_i) + \epsilon_{i,\omega},\ \omega \in \{0,1\}, \tag{2}$$

where $\epsilon_{i,\omega} \sim \mathcal{N}(0, \sigma_\omega^2)$ is a Gaussian noise variable. It follows from (2) that $\mathbb{E}[Y_i^{(\omega)} \,|\, X_i = x] = f_\omega(x)$, and hence the ITE is given by $T(x) = f_1(x) - f_0(x)$. The functions $f_1(x)$ and $f_0(x)$ correspond to the *response surfaces* over the subjects' feature space with and without the intervention; the difference between these two surfaces correspond to the individualized effect of the intervention. We assume that $\mathcal{X}$ is a compact metric space (e.g. bounded, closed sets in $\mathbb{R}^d$), and that the true regression function $f_\omega : \mathcal{X} \to \mathbb{R}$, $\omega \in \{0,1\}$, lives in a space of "smooth" or "regular" functions $\mathcal{F}^{\alpha_\omega}$, where α_ω is a smoothness (or regularity) parameter. This roughly means that f_ω is α_ω-differentiable; precise definitions for α_ω-regular function classes will be provided in subsequent Sections.

A Bayesian procedure for estimating the ITE function entails specifying a prior distribution Π over the response surfaces $f_1(x)$ and $f_0(x)$, which in turn induces a prior over $T(x)$. The nonparametric nature of inference follows from the fact that Π is a prior over functions, and hence the estimation problem involves an infinite-dimensional parameter space. For a given prior Π, the Bayesian inference procedure views the observational dataset $\mathcal{D}_n$ as being sampled according to the following generative model:

$$\begin{aligned} f_0, f_1 &\sim \Pi,\ X_i \sim d\mathbb{P}(X_i = x) \\ \omega_i \,|\, X_i = x &\sim \text{Bernoulli}(\gamma(x)) \\ Y_i^{(\omega_i)} \,|\, f_0, f_1, \omega_i &\sim \mathcal{N}(f_{\omega_i}(x), \sigma_{\omega_i}^2),\ i = 1, \ldots, n. \end{aligned} \tag{3}$$

Since we are interested in estimating an underlying true ITE function $T(x)$, we will analyze the Bayesian causal inference procedure within the so-called frequentist setup, which assumes that the subjects' outcomes $\{Y_i^{(\omega_i)}\}_{i=1}^n$ are generated according to the model in (3) for a given true (and fixed) regression functions $f_0(x)$ and $f_1(x)$. That is, in the next Subsection, we will assess the quality of a Bayesian inference procedure by quantifying the amount of information the posterior distribution $d\Pi_n(T \,|\, \mathcal{D}_n) = d\Pi_n(f_1 - f_0 \,|\, \mathcal{D}_n)$ has about the true ITE function T. This type of analysis is sometimes referred to as the "Frequentist-Bayes" analysis [29].

C. Information Rates

How much information about the true causal effect function $T(.)$ is conveyed in the posterior $d\Pi_n(T \,|\, \mathcal{D}_n)$? A natural measure of the "informational quality" of a posterior $d\Pi_n(T \,|\, \mathcal{D}_n)$ is the information-theoretic criterion due to Barron [30], which quantifies the quality of a posterior via the Kullback-Leibler (KL) divergence between the posterior and true distributions. In that sense, the quality (or informativeness) of the posterior $d\Pi_n(T \,|\, \mathcal{D}_n)$ at a feature point x is given by the KL divergence between the posterior distribution at x, $d\Pi_n(T(x) \,|\, \mathcal{D}_n)$, and the true distribution of $(Y^{(1)} - Y^{(0)}) \,|\, X = x$. The overall quality of a posterior is thus quantified by marginalizing the pointwise KL divergence over the feature space $\mathcal{X}$. For a prior Π, true responses f_0 and f_1, propensity function γ, and observational datasets of size n, the *expected KL risk* is:

$$\mathbb{D}_n(\Pi; f_0, f_1, \gamma) = \mathbb{E}_x \left[\mathbb{E}_{\mathcal{D}_n} \left[D_{\mathrm{KL}} \left(P(x) \parallel Q_{\mathcal{D}_n}(x) \right) \right] \right], \quad (4)$$

where $D_{\mathrm{KL}}(.\|.)$ is the KL divergence[2], $P(x)$ is the true distribution of $T(x)$, i.e. $P(x) = d\mathbb{P}(Y^{(1)} - Y^{(0)} \,|\, X = x)$, and $Q_{\mathcal{D}_n}(x)$ is the posterior distribution of $T(x)$, given by:

$$\begin{aligned} Q_{\mathcal{D}_n}(x) &= d\Pi_n(Y^{(1)} - Y^{(0)} \,|\, X = x, \mathcal{D}_n) \\ &\overset{(\star)}{=} d\Pi_n(T(x) + \mathcal{N}(0, \sigma_0^2 + \sigma_1^2) \,|\, \mathcal{D}_n), \\ &\overset{(*)}{=} \int \mathcal{N}(T(x), \sigma_0^2 + \sigma_1^2)\, d\Pi_n(T(x) \,|\, \mathcal{D}_n), \end{aligned} \quad (5)$$

where steps $(\star)$ and $(*)$ in (5) follow from the sampling model in (3). The expected KL risk $\mathbb{D}_n$ in (4) marginalizes the pointwise KL divergence $D_{\mathrm{KL}}(P(x) \,\|\, Q_{\mathcal{D}_n}(x))$ over the distribution of the observational dataset $\mathcal{D}_n$ (generated according to (3)), and the feature distribution $d\mathbb{P}(X = x)$. Variants of the expected KL risk in (4) have been widely used in the analysis of nonparametric regression models, usually in the form of (cumulative) Cesàro averages of the pointwise KL divergence at certain points in the feature space [27], [30], [32], [33]. Assuming posterior consistency, the (asymptotic) dependence of $\mathbb{D}_n(\Pi; f_0, f_1, \gamma)$ on n reflects the rate by which the posterior $d\Pi_n(T \,|\, \mathcal{D}_n)$ "sandwiches" the true ITE function $T(x)$ everywhere in $\mathcal{X}$. An efficient causal inference procedure would exhibit a rapidly decaying $\mathbb{D}_n$: this motivates the definition of an "*information rate*".

Definition 1. (Information Rate) We say that the *information rate* of a Bayesian causal inference procedure is $\mathcal{I}_n$, for a sequence $\mathcal{I}_n \to 0$, if $\mathbb{D}_n(\Pi; f_0, f_1, \gamma)$ is $\Theta(\mathcal{I}_n)$. ■

Note that $\mathcal{I}_n$ is the equivalence class of all sequences that have the same asymptotic rate of convergence. In the rest of this paper, we use the notation $\mathcal{I}_n(\Pi; f_0, f_1, \gamma)$ to denote the information rate achieved by a prior Π in a causal inference problem instance described by the tuple (f_0, f_1, γ). The notion of an "information rate" for a Bayesian causal effect inference procedure is closely connected to the frequentist estimation rate (with respect to the L_2 loss) with $T(.)$ as the estimand [30], [34], [35]. The following Theorem establishes such a connection.

Theorem 1. Let $\mathcal{I}_n(\Pi; f_0, f_1, \gamma)$ be the information rate of a given Bayesian causal inference procedure, then we have that

$$\mathbb{E}_{\mathcal{D}_n} \left[\left\| \mathbb{E}_\Pi \left[T \,|\, \mathcal{D}_n \right] - T \right\|_2^2 \right] \leq 2(\sigma_0^2 + \sigma_1^2) \cdot \mathcal{I}_n(\Pi; f_0, f_1, \gamma),$$

where $\|.\|_2^2$ is the $L_2(\mathbb{P})$-norm with respect to the feature distribution, i.e. $\|f\|_2^2 = \int f^2(x) d\mathbb{P}(X = x)$.

Proof. Recall from (4) that $\mathbb{D}_n$ is given by

$$\mathbb{D}_n(\Pi; f_0, f_1, \gamma) = \mathbb{E}_x \left[\mathbb{E}_{\mathcal{D}_n} \left[D_{\mathrm{KL}} \left(P(x) \parallel Q_{\mathcal{D}_n}(x) \right) \right] \right].$$

Based on (5), $D_{\mathrm{KL}} \left(P(x) \parallel Q_{\mathcal{D}_n}(x) \right)$ can be written as

$$D_{\mathrm{KL}} \left(P(x) \,\middle\|\, \int \mathcal{N}(T(x), \sigma_0^2 + \sigma_1^2)\, d\Pi_n(T(x) \,|\, \mathcal{D}_n) \right),$$

which by the concavity of the KL divergence in its second argument, and using Jensen's inequality, is bounded below by

$$D_{\mathrm{KL}} \left(P(x) \,\middle\|\, \mathcal{N} \left(\mathbb{E}_\Pi[T(x) \,|\, \mathcal{D}_n], \sigma_0^2 + \sigma_1^2 \right) \right).$$

From the regression model in (2), we know that $P(x) = d\mathbb{P}(Y^{(1)} - Y^{(0)} \,|\, X = x) \sim \mathcal{N}(T(x), \sigma_0^2 + \sigma_1^2)$, and hence KL divergence above can be written as

$$D_{\mathrm{KL}} \left(\mathcal{N}(T(x), \sigma_0^2 + \sigma_1^2) \,\middle\|\, \mathcal{N} \left(\mathbb{E}_\Pi[T(x) \,|\, \mathcal{D}_n], \sigma_0^2 + \sigma_1^2 \right) \right),$$

which is given by $\frac{1}{2(\sigma_0^2 + \sigma_1^2)} \left| \mathbb{E}_\Pi[T(x) \,|\, \mathcal{D}_n] - T(x) \right|^2$ since $D_{\mathrm{KL}}(\mathcal{N}(\mu_0, \sigma^2) \| \mathcal{N}(\mu_1, \sigma^2)) = \frac{1}{\sigma^2} |\mu_1 - \mu_0|^2$ [31]. Hence, the expected KL risk is bounded below as follows

$$\begin{aligned} \mathbb{D}_n(\Pi; f_0, f_1, \gamma) &\geq \mathbb{E}_{\mathcal{D}_n} \left[\mathbb{E}_x \left[\frac{\left| \mathbb{E}_\Pi[T(x) \,|\, \mathcal{D}_n] - T(x) \right|^2}{2(\sigma_0^2 + \sigma_1^2)} \right] \right], \\ &= \frac{1}{2(\sigma_0^2 + \sigma_1^2)} \mathbb{E}_{\mathcal{D}_n} \left[\left\| \mathbb{E}_\Pi[T \,|\, \mathcal{D}_n] - T \right\|_2^2 \right], \end{aligned}$$

for all $n \in \mathbb{N}_+$. ■

Theorem 1 says that the information rate of causal inference upper bounds the rate of convergence of the $L_2(\mathbb{P})$ risk of the sequence of estimates $\hat{T}_n$ induced by the posterior mean $\int T d\Pi_n(T \,|\, \mathcal{D}_n)$. The $L_2(\mathbb{P})$ risk $\| \mathbb{E}_\Pi[T \,|\, \mathcal{D}_n] - T \|_2^2$ was dubbed the *precision in estimating heterogeneous effects* (PEHE) by Hill in [8], and is the most commonly used metric for evaluating causal inference model [8], [18], [21], [28], [36]. Theorem 1 tells us that the PEHE is $O(\mathcal{I}_n)$, and hence inference procedures with good information rates should also exhibit fast convergence for the PEHE. Thus, the asymptotic behavior of $\mathcal{I}_n(\Pi; f_0, f_1, \gamma)$ is revealing of both the informational quality of the Bayesian posterior, as well as the convergence rates of frequentist loss functions.

V. Optimal Information Rates for Bayesian Causal Inference

In this Section, we establish a fundamental limit on the information rate that can be achieved by *any* sequence of posteriors $d\Pi_n(T \,|\, \mathcal{D}_n)$ for a given causal inference problem. Let

[2]The KL divergence between probability measures P and Q is given by $D_{\mathrm{KL}}(P\|Q) = \int \log(dP/dQ) dP$ [31]. The existence of the Radon-Nikodym derivative $\frac{dP}{dQ_n}$ in (4) is guaranteed since P and Q are mutually absolutely continuous.

the *achievable information rate* for a given prior Π and function classes $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$, denoted by $I_n(\Pi; \mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}, \gamma)$, be the rate obtained by taking the supremum of the information rate over functions in $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$. This is a quantity that depends only on the prior but not on the specific realizations of f_0 and f_1. The *optimal information rate* is defined to be the maximum achievable information rate by any possible prior, and is denote by $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}, \gamma)$. While the information rate $\mathcal{I}_n(\Pi; f_0, f_1, \gamma)$ characterizes a particular instance of a causal inference problem with (f_0, f_1, γ) and a given Bayesian prior Π, the optimal information rate $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}, \gamma)$ is an abstract (prior-independent) measure of the "information capacity" or the "hardness" of a class of causal inference problems (corresponding to response surfaces in $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$). Intuitively, one expects that the limit on the achievable information rate will be higher for smooth (regular) response surfaces and for propensity functions that are close to 0.5 everywhere in $\mathcal{X}$. Whether or not the Bayesian inference procedure achieves the optimal information rate will depend on the prior Π. In the next Section, we will investigate different design choices for the prior Π, and characterize the "capacity-achieving" priors that achieve the optimal information rate.

Before presenting the main result of this Section (Theorem 2), we first give a formal definition for the optimal information rate. We say that $g(n)$ is a *lower information rate* if there is a positive constant c such that

$$\lim_{n\to\infty} \inf_{\Pi} \sup_{f_0, f_1} \mathbb{P}(\mathcal{I}_n(\Pi; f_0, f_1, \gamma) \geq c \cdot g(n)) = 1, \qquad (6)$$

where the infimum is taken over all possible priors, and the supremum is taken over the function spaces $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$. We say that $g(n)$ is an *achievable information rate* if there exists a prior Π and a positive constant c such that

$$\lim_{n\to\infty} \sup_{f_0, f_1} \mathbb{P}(\mathcal{I}_n(\Pi; f_0, f_1, \gamma) \geq c \cdot g(n)) = 0. \qquad (7)$$

The sequence $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}, \gamma) \to 0$ is an optimal information rate if it is both a lower and an achievable information rate [25]. In Theorem 2, we will use the notion of *metric entropy* $H(\delta; \mathcal{F}^{\alpha})$ to characterize the "size" of general (nonparametric or parametric) function classes. The metric entropy $H(\delta; \mathcal{F}^{\alpha})$ of a function space $\mathcal{F}^{\alpha}$ is given by the logarithm of the *covering number* $N(\delta, \mathcal{F}^{\alpha}, \rho)$ of that space with respect to a metric ρ, i.e. $H(\delta; \mathcal{F}^{\alpha}) = \log(N(\delta, \mathcal{F}^{\alpha}, \rho))$. A formal definition for covering numbers is provided below.

Definition 2. (Covering number) A δ-cover of a given function space $\mathcal{F}^{\alpha}$ with respect to a metric ρ is a set of functions $\{f^1, \ldots, f^N\}$ such that for any function $f \in \mathcal{F}^{\alpha}$, there exists some $v \in \{1, \ldots, N\}$ such that $\rho(f, f^v) \leq \delta$. The δ-covering number of $\mathcal{F}^{\alpha}$ is [29]

$$N(\delta, \mathcal{F}^{\alpha}, \rho) := \inf\{N \in \mathbb{N} : \exists \text{ a } \delta\text{-cover of } \mathcal{F}^{\alpha}\}. \quad \blacksquare$$

That is, the covering number of a function class $\mathcal{F}^{\alpha}$ is the number of balls (in a given metric ρ) of a fixed radius $\delta > 0$ required to cover it. Throughout this paper, the metric entropy will always be evaluated with respect to the $L_2(\mathbb{P})$ norm. In the light of the definition above, the metric entropy can be thought of as a measure of the complexity of a function class; smoother function classes would generally display a smaller metric entropy. All function classes considered in this paper comprise bounded functions over bounded domains, and hence all such classes have finite metric entropy. Figure 1 shows a pictorial depiction for two exemplary function classes $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$ for the treated and control responses, respectively. In this depiction, α_0 is smaller than α_1, hence the δ-cover of $\mathcal{F}^{\alpha_0}$ contains more balls than the δ-cover of $\mathcal{F}^{\alpha_1}$, and it follows that $\mathcal{F}^{\alpha_0}$ has a larger metric entropy. This manifests in the control response surface f_0 being less smooth than the treated response surface f_1. This is usually the case for real-world data on responses to medical treatments, where the untreated population typically display more heterogeneity than the treated population [20].

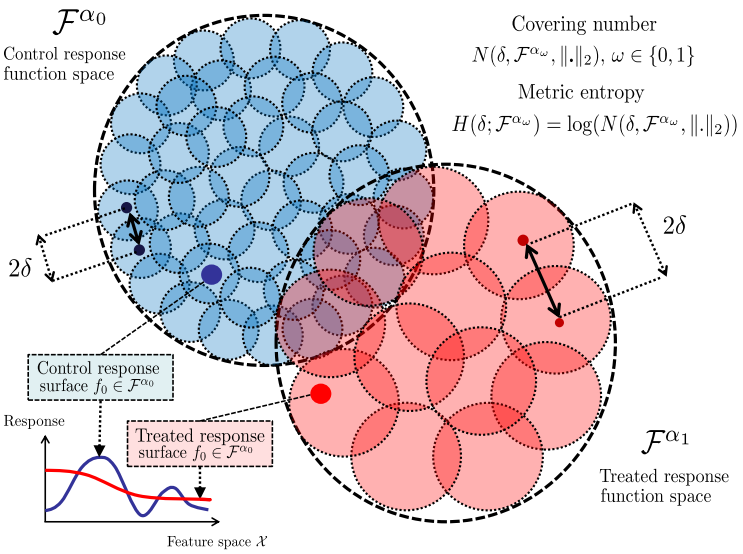


Fig. 1: Pictorial depiction of covering sets for $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$.

We now present the main result of this Section. In the following Theorem, we provide a general characterization for the optimal information rates of Bayesian causal inference when the treated and control surfaces are known to belong to function classes $\mathcal{F}^{\alpha_1}$ and $\mathcal{F}^{\alpha_0}$. The proofs for all the Theorems in this paper are provided in an online appendix[3].

Theorem 2. (Optimal Information Rates) Suppose that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, and that Conditions 1-2 hold. Then the optimal information rate is $\Theta(\delta_0^2 \wedge \delta_1^2)$, where δ_ω is the solution for $H(\delta_\omega; \mathcal{F}^{\alpha_\omega}) \asymp n\,\delta_\omega^2$, $\omega \in \{0, 1\}$.

Proof. See Appendix A. $\blacksquare$

Theorem 2 characterizes $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}, \gamma)$ in terms of the metric entropies $H(\delta; \mathcal{F}^{\alpha_0})$ and $H(\delta; \mathcal{F}^{\alpha_1})$ for general function classes $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$. We used the local *Fano method* to derive an information-theoretic lower bound on the information rate that can be achieved by any estimator [30]. The characterization in Theorem 2 implies that selection bias *has*

[3] http://medianetlab.ee.ucla.edu/papers/JSTSPappendix.pdf

TABLE I: OPTIMAL INFORMATION RATES FOR BAYESIAN CAUSAL INFERENCE IN STANDARD FUNCTION SPACES.

SPACE $\mathcal{F}^\alpha$	METRIC ENTROPY $H(\delta;\mathcal{F}^\alpha)$	RESPONSE SURFACES $f_0,\, f_1$	OPTIMAL INFORMATION RATE $I_n^*(\mathcal{F}^{\alpha_0},\mathcal{F}^{\alpha_1})$
Analytic $C^\omega(\mathcal{X})$	$H(\delta;C^\omega)=\log\left(\frac{1}{\delta}\right)$	$f_0,f_1\in C^\omega(\mathcal{X})$	$\Theta(n^{-1})$
Smooth $C^\infty(\mathcal{X})$	$H(\delta;C^\infty)=\log\left(\frac{1}{\delta}\right)$	$f_0,f_1\in C^\infty(\mathcal{X})$	$\Theta(n^{-1})$
α-Smooth $C^\alpha(\mathcal{X})$	$H(\delta;C^\alpha)=\delta^{-\frac{d}{\alpha}}$	$f_0\in C^{\alpha_0}(\mathcal{X}),\, f_1\in C^{\alpha_1}(\mathcal{X})$	$\Theta\left(n^{-2(\alpha_0\wedge\alpha_1)/(2(\alpha_0\wedge\alpha_1)+d)}\right)$
Hölder $H^\alpha(\mathcal{X})$	$H(\delta;H^\alpha)=\delta^{-\frac{d}{\alpha}}$	$f_0\in H^{\alpha_0}(\mathcal{X}),\, f_1\in H^{\alpha_1}(\mathcal{X})$	$\Theta\left(n^{-2(\alpha_0\wedge\alpha_1)/(2(\alpha_0\wedge\alpha_1)+d)}\right)$
Sobolev $S^\alpha(\mathcal{X})$	$H(\delta;S^\alpha)=\delta^{-\frac{d}{\alpha}}$	$f_0\in S^{\alpha_0}(\mathcal{X}),\, f_1\in S^{\alpha_1}(\mathcal{X})$	$\Theta\left(n^{-2(\alpha_0\wedge\alpha_1)/(2(\alpha_0\wedge\alpha_1)+d)}\right)$
Besov $B^\alpha_{p,q}(\mathcal{X})$	$H(\delta;B^\alpha)=\delta^{-\frac{d}{\alpha}}$	$f_0\in B^{\alpha_0}_{p,q}(\mathcal{X}),\, f_1\in B^{\alpha_1}_{p,q}(\mathcal{X})$	$\Theta\left(n^{-2(\alpha_0\wedge\alpha_1)/(2(\alpha_0\wedge\alpha_1)+d)}\right)$
Lipschitz $L^\alpha(\mathcal{X})$	$H(\delta;L^\alpha)=\delta^{-\frac{d}{\alpha}}$	$f_0\in L^{\alpha_0}(\mathcal{X}),\, f_1\in L^{\alpha_1}(\mathcal{X})$	$\Theta\left(n^{-2(\alpha_0\wedge\alpha_1)/(2(\alpha_0\wedge\alpha_1)+d)}\right)$
Parametric models	$H(\delta;\Theta)=K\cdot\log\left(\frac{1}{\delta}\right),\lvert\Theta\rvert=K$	$f_\omega(\theta_\omega),\theta_\omega\in\Theta_\omega,\lvert\Theta_\omega\rvert=K_\omega,\omega\in\{0,1\}$	$\Theta\left((K_0\wedge K_1)^2\cdot n^{-1}\right)$

no effect on the achievable information rate. (Thus, in the rest of the paper we drop the dependency on γ when referring to I_n^*.) That is, as long as the overlap condition holds, selection bias does not hinder the information rate that can be achieved by a Bayesian causal inference procedure, and we can hope to find a good prior Π that achieves the optimal rate of posterior contraction around the true ITE function $T(x)$ irrespective of the amount of bias in the data. Theorem 2 also says that the achievable information rate is bottle-necked by the more "complex" of the two response surfaces f_0 and f_1. Hence, we cannot hope to learn the causal effect at a fast rate if either of the treated or the control response surfaces are rough, even when the other surface is smooth.

The general characterization of the optimal information rate for causal inference provided in Theorem 2 can be cast into specific forms by specifying the regularity classes $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$. Table I demonstrates the optimal information rates for standard function classes, including the space of analytic, smooth, Hölder, Sobolev, Besov, and Lipschitz functions. A rough description for the optimal information rates of all nonparametric function spaces (α-smooth, Hölder, Sobolev, Besov, and Lipschitz) can be given as follows. If f_0 is α_0-regular (e.g. α_0-differentiable) and f_1 is α_1-regular, then the optimal information rate for causal inference is

$$I_n^*(\mathcal{F}^{\alpha_0},\mathcal{F}^{\alpha_1})\asymp n^{\frac{-2(\alpha_0\wedge\alpha_1)}{2(\alpha_0\wedge\alpha_1)+d}}, \tag{8}$$

where $\asymp$ denotes *asymptotic equivalence*, i.e. in Bachmann-Landau notation, $g(x)\asymp f(x)$ if $g(x)=\Theta(f(x))$. That is, the regularity parameter of the rougher response surface, i.e. $\alpha_0\wedge\alpha_1$, dominates the rate by which any inference procedure can acquire information about the causal effect. This is because, if one of the two response surfaces is much more complex (rough) than the other (as it is the case in the depiction in Figure 1), then the ITE function $T(x)$ would naturally lie in a function space that is at least as complex as the one that contains the rough surface. Moreover, the best achievable information rate depends only on the smoothness of the response surfaces and the dimensionality of the feature space, and is independent of the selection bias. As it is the case in standard nonparametric regression, the optimal information rate for causal inference gets exponentially slower as we add more dimensions to the feature space [23], [26], [35], [37]. Note that in Theorem 2, we assumed that for the surfaces f_0 and f_1, all of the d dimensions of $\mathcal{X}$ are *relevant*, and that the responses have the same smoothness level on all dimensions. Now assume that surfaces f_0 and f_1 have relevant feature dimensions in the sets $\mathcal{P}_0$ and $\mathcal{P}_1$, respectively, where $|\mathcal{P}_\omega|=p_\omega\leq d,\,\omega\in\{0,1\}$ [37]. In this case, the optimal information rate becomes

$$I_n^*(\mathcal{F}^{\alpha_0}_{\mathcal{P}_0},\mathcal{F}^{\alpha_1}_{\mathcal{P}_1})\asymp n^{\frac{-2\alpha_0}{2\alpha_0+p_0}}\wedge n^{\frac{-2\alpha_1}{2\alpha_1+p_1}}, \tag{9}$$

where $\mathcal{F}^{\alpha_\omega}_{\mathcal{P}_\omega}$ denotes the space of functions in $\mathcal{F}^{\alpha_\omega}$ for which the relevant dimensions are in $\mathcal{P}_\omega$. In (9), the rate is dominated by the more complex response surface, where "complexity" here is manifesting as a combination of the number of relevant dimensions and the smoothness of the response over the those dimensions. One implication of (9) is that the information rate can be bottle-necked by the smoother of the response surfaces f_0 and f_1, if such a response has more relevant dimensions in the feature space[4]. More precisely, if $\alpha_0<\alpha_1$, then the information rate can still be bottle-necked by the smoother surface f_1 as long as $p_1>\frac{\alpha_1}{\alpha_0}\,p_0$.

Since the optimal (Bayesian) information rate is an upper bound on the optimal (frequentist) minimax estimation rate (Theorem 1), we can directly compare the limits of estimation in the causal inference setting (established in Theorem 2) with that of the standard nonparametric regression setting. It is well known that the optimal minimax rate for estimating an α-regular function is $\Theta(n^{-2\alpha/(2\alpha+d)})$; a classical result due to Stone [25], [26]. The result of Theorem 2 (and the tabulated results in Table I) asserts that the causal effect estimation problem is as hard as the problem of estimating the "rougher" of the two surfaces f_0 and f_1 in a standard regression setup. The fact that selection bias does not impair the optimal information rate for causal inference is consistent with previous results on minimax-optimal kernel density estimation under selection bias or length bias [38]–[41]. In these settings, selection bias did not affect the optimal minimax rate for density estimation, but the kernel bandwidth optimization strategies that achieve the optimal rate needed to account for selection bias [40], [42]. In Section VI, we show that the same holds for causal inference: in order to achieve the optimal information rate, the strategy for selecting the prior Π needs to account for selection bias. This means that even though the optimal information rates in the causal inference and standard regression settings

[4]A more general characterization of the information rate would consider the case when the responses have different smoothness levels on each of the d-dimensions. Unfortunately, obtaining such a characterization is technically daunting.

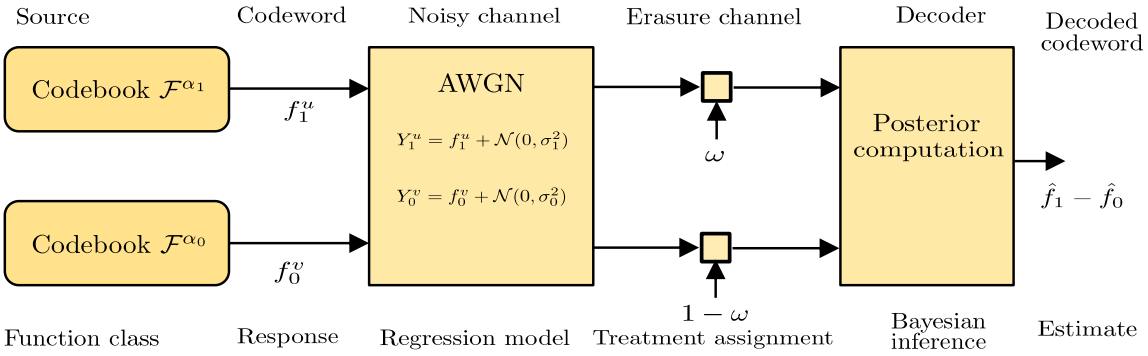


Fig. 2: Duality of Bayesian causal inference and channel coding in a cascade of AWGN and erasure channels.

are *similar*, the optimal estimation strategies in both setups are *different*.

In the rest of this Section, we present a sketch of the proof of Theorem 2. The detailed proof is provided in Appendix A. (Readers can skip the following Subsection and proceed to Section VI without loss of context.) In Section VI, we study different strategies for selecting the prior Π, and identify the priors that achieve the optimal information rate in Theorem 2.

Sketch of the Proof of Theorem 2

We will first show that $\delta_0^2 \wedge \delta_1^2$, where δ_ω is the solution for the transcendental equation $H(\delta_\omega; \mathcal{F}^{\alpha_\omega}) = n\,\delta_\omega^2$, $\omega \in \{0, 1\}$, is a lower information rate (condition (6)) by proving that it is an optimal lower bound for $\mathbb{D}_n$. We will then show that such a rate is achievable (7), and hence $I_n^* \asymp \delta_0^2 \wedge \delta_1^2$.

VI. Rate-adaptive Bayesian Causal Inference

In Section V, we have established the optimal rates by which any Bayesian inference procedure can gather information about the causal effect of a treatment from observational data. In this Section, we investigate different strategies for selecting the prior Π, and study their corresponding achievable information rates. (An optimal prior Π^* is one that achieves the optimal information rate I_n^*.) A strategy for selecting Π comprises the following three modeling choices:

1) How to incorporate the treatment assignment variable ω in the prior Π?
2) What function (regularity) class should the prior Π place a probability distribution over?
3) What should be the smoothness (regularity) parameter of the selected function class?

The first modeling decision involves *two* possible choices. The *first* choice is to give no special role to the treatment assignment indicator ω, and build a model that treats it in a manner similar to all other features by augmenting it to the feature space $\mathcal{X}$. This leads to models of the form

$$f(x, \omega) : \mathcal{X} \times \{0, 1\} \to \mathbb{R}.$$

We refer to priors over models of the form above as **Type-I priors**. The *second* modeling choice is to let ω index two different models for the two response surfaces. This leads to models of the form $\mathbf{f}(x) = [f_0(x), f_1(x)]^T$, where $f_0 \in \mathcal{F}^{\beta_0}$ and $f_1 \in \mathcal{F}^{\beta_1}$ for some $\beta_0, \beta_1 > 0$. We refer to priors over models of the form $\mathbf{f}(.)$ as **Type-II priors**.

Type I and Type II priors induce different estimators for the ITE function: we will study the information rates achieved by those estimators in order to assess the qualities of the two modeling choices. The posterior mean ITE estimator (which is the optimal estimator with respect to the $L_2(\mathbb{P})$ risk) for a Type-I prior is given by

$$\hat{T}_n(x) = \mathbb{E}_\Pi[f(x, 1) \,|\, \mathcal{D}_n] - \mathbb{E}_\Pi[f(x, 0) \,|\, \mathcal{D}_n],$$

whereas for a Type-II prior, the posterior mean ITE estimator is given by $\hat{T}_n(x) = \mathbb{E}_\Pi[\mathbf{f}^T(x)\,\mathbf{v} \,|\, \mathcal{D}_n]$, where $\mathbf{v} = [-1, 1]^T$. Figure 3 is a pictorial depiction for the posterior mean ITE estimates obtained via Type-I and Type-II priors.

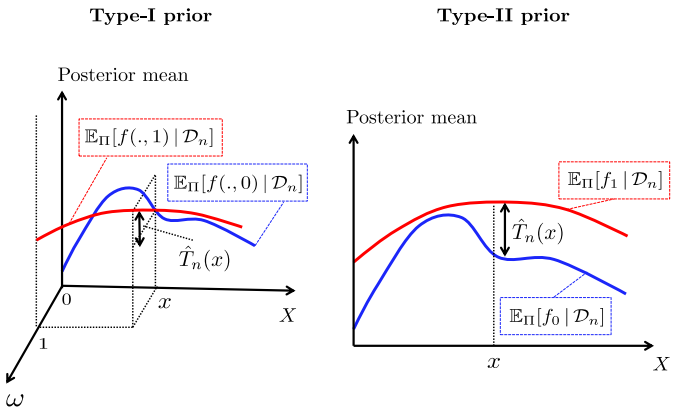


Fig. 3: Depiction for estimates obtained by Type-I and Type-II priors.

The main difference between Type-I and Type-II priors is that the former restricts the smoothness of $f(x, \omega)$ on any feature dimension to be the same for $\omega = 0$ and $\omega = 1$. This also entails that the relevant dimensions for the two response surfaces ($\omega = 0$ and $\omega = 1$) need to be the same under a Type-I prior. (This is a direct consequence of the fact that Type-I priors give no special role to the variable ω.) As a result, a priori knowledge (or even data-driven a priori knowledge) on the different characteristics between responses f_0 and f_1 (e.g. different smoothness levels or relevant dimensions) cannot be incorporated in a Type-I prior. Type-II priors can incorporate such information as they provide separate models for f_0 and f_1. However, while Type-I priors give a posterior of f_0 and f_1

using *all* the observational data (by virtue of joint modeling of the two surfaces), Type-II priors use only the data for the treated population to compute posteriors of f_1 (and uses the control population for f_0), which can be problematic if the two populations posses highly unbalanced relative sizes (e.g. treated populations are usually much smaller than control populations [1]).

In order to better illustrate the difference between Type-I and Type-II priors, we look at their simpler parametric counterparts. A Type-I linear regression model defines $f(x,\omega)$ as a linear function $Y = \beta^T \mathbf{x} + \tilde{\gamma} \cdot \omega + \epsilon$, where $\beta \in \mathbb{R}^d$, $\tilde{\gamma} \in \mathbb{R}$, and ϵ is a Gaussian noise variable. (Here the Type-I prior is a prior on the model coefficients β and $\tilde{\gamma}$) As we can see, this model restricts the two responses f_0 and f_1 to have the exact same interactions with the features through the coefficients in β. If we know a priori that f_0 and f_1 have different "slopes" or different relevant dimensions, we cannot incorporate this knowledge into the model. What would such a model learn? Assuming consistency, the estimated ITE function would be $\hat{T}_n(x) \to \tilde{\gamma}$ everywhere in $\mathcal{X}$. Thus, the restricted nature of a Type-I parametric model led to a constant (non-individualized) estimate of $T(.)$. Contrarily, a Type-II model of the form $Y^{(\omega)} = \beta_\omega^T \mathbf{x} + \epsilon$, $\omega \in \{0,1\}$, would allow for learning a linear estimate $\hat{T}_n(x)$ of the true function $T(x)$, with potentially different relevant dimensions for both surfaces (different non-zero entries in β_0 and β_1). However, Type-II model will only use data with $\omega = w$ to fit the model for $Y^{(w)}$, $w \in \{0,1\}$.

Unlike their parametric counterparts, the nonparametric Type-I and II priors can (in general) learn the ITE function consistently, but how do their information rates compare? Subsection VI-A studies the *achievable* information rates for "oracle" Type-I and Type-II priors that are informed with the true smoothness parameters (α_0 and α_1) and relevant dimensions of the function classes $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$. In Subsection VI-B, we study the (more realistic) setting when $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$ are unknown, and investigate different strategies for adapting the prior Π to the smoothness of the treated and control response surface in a data-driven fashion. The *achievability* and *adaptivity* analyses in Subsections VI-A and VI-B will guide the design of a practical Bayesian inference algorithm, which we present in Section VII.

A. Oracle Priors

In this Subsection, we assume that the true smoothness and relevant dimensions for f_0 and f_1 are known a priori. In the following Theorem, we show that Type-II priors are generally a better modeling choice than Type-I priors.

Theorem 3. (Sub-optimality of Type-I priors) Let Π°_β be the space of all Type-I priors that give probability one to draws from H^β, and let $\Pi^{\circ\circ}_{\beta_{0,1}}$ be the space of all Type-II priors that give probability one to draws from $(H^{\beta_0}, H^{\beta_1})$. If $f_0 \in H^{\alpha_0}_{\mathcal{P}_0}$ and $f_1 \in H^{\alpha_1}_{\mathcal{P}_1}$, then

$$\inf_{\beta} \inf_{\Pi \in \Pi^\circ_\beta} I_n(\Pi; H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}) \lesssim I^*_n(H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}),$$

$$\inf_{\beta_0,\beta_1} \inf_{\Pi \in \Pi^{\circ\circ}_{\beta_{0,1}}} I_n(\Pi; H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}) \asymp I^*_n(H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}).$$

Proof. See Appendix B. ■

The proof of Theorem 3 utilizes an information-theoretic lower bound (based on Fano's inequality) on the minimax information rate achieved by Type-I and Type-II priors. Theorem 3 says that the best minimax information rate that *any* Type-I prior can achieve is always suboptimal, even if we know the true smoothness of the response surfaces f_0 and f_1. Contrarily, Theorem 3 also says that an oracle Type-II prior can achieve the optimal information rate. When the the surfaces f_0 and f_1 have the same relevant dimensions and the same smoothness, the gap between the best information achieved by a Type-I prior and the optimal information is not large, and it diminishes for high-dimensional feature spaces or very smooth response surfaces (see Appendix B). The gap becomes larger when the surfaces f_0 and f_1 exhibit different relevant dimensions and smoothness levels. More precisely, the best achievable information rate for a Type-I prior is given by

$$\inf_{\beta} \inf_{\Pi \in \Pi^\circ_\beta} I_n(\Pi; H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}) = \Theta\left(n^{\frac{-2(\alpha_0 \wedge \alpha_1)}{2(\alpha_0 \wedge \alpha_1) + |\mathcal{P}_0 \cup \mathcal{P}_1| + 1}} \right),$$

whereas for Type-II priors, the best achievable rate is

$$\inf_{\beta_0,\beta_1} \inf_{\Pi \in \Pi^{\circ\circ}_{\beta_{0,1}}} I_n(\Pi; H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}) = \Theta\left(n^{\frac{-2\alpha_0}{2\alpha_0 + |\mathcal{P}_0|}} \wedge n^{\frac{-2\alpha_1}{2\alpha_1 + |\mathcal{P}_1|}} \right).$$

We note that most state-of-the-art causal inference algorithms, such as causal forests [7], Bayesian additive regression trees [8], and counterfactual regression [18], [28], use Type-I regression structures for their estimates. (While causal forests and counterfactual regression are frequentist methods, Theorems 1 and 3 imply the sub-optimality of their minimax estimation rate.) The sub-optimality of Type-I priors, highlighted in Theorem 3, suggests that improved estimates can be achieved over state-of-the-art algorithms via a Type-II regression structure.

We now focus on the second and third modeling questions: on what function space should the prior Π be defined, and what smoothness level to select? Obviously, with oracle knowledge of $\mathcal{F}^{\alpha_0}$ and $\mathcal{F}^{\alpha_1}$, the oracle prior Π should give probability one to draws from the function space $\mathcal{F}$, since the true response surfaces are known to lie in $\mathcal{F}$; it is non-obvious though what smoothness level should be selected for the prior. In the following Theorem, we characterize the achievable information rates as a function of the smoothness of the prior.

Theorem 4. (The Matching Condition) Suppose that f_0 and f_1 are in Hölder spaces H^{α_0} and H^{α_1}, respectively. Let $\Pi^{\circ\circ}_\beta$ be a Type-II prior on $(H^{\beta_0}, H^{\beta_1})$. Then we have that:

$$I_n(\Pi^{\circ\circ}_{\beta_{0,1}}; H^{\alpha_0}, H^{\alpha_1}) \asymp n^{\frac{-2(\beta_0 \wedge \alpha_0)}{2\beta_0 + d}} + n^{\frac{-2(\beta_1 \wedge \alpha_1)}{2\beta_1 + d}}.$$

Proof. See Appendix C. ■

Recall that the optimal information rate for causal inference in Hölder spaces is $I^*_n(H^{\alpha_0}, H^{\alpha_1}) = n^{\frac{-2(\alpha_0 \wedge \alpha_1)}{2(\alpha_0 \wedge \alpha_1) + d}}$ (Table I). Theorem 4 quantifies the information rates achieved by a Type-II prior with smoothness levels β_0 and β_1 for treated

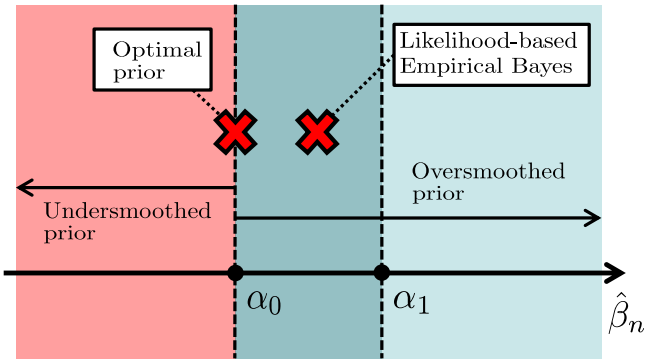
and control responses. Theorem 4 says that a Type-II prior can achieve the optimal information rate if and only if it *captures the smoothness of the rougher of the two response surfaces*. This gives rise to a *matching condition* that a Type-II prior requires in order to guarantee optimality: for a prior to achieve the optimal information rate, it must have a smoothness level that matches the smoothness of the rougher of the two response surfaces. For the prior $\Pi^{\circ\circ}_{\beta}$, the matching condition is $\beta_\omega = \alpha_0 \wedge \alpha_1$, $\omega = \arg\max_{w\in\{0,1\}} \alpha_w$. Note that Theorems 3 and 4 assume that the true response surfaces are Hölder continuous. Deriving the matching conditions for general function spaces is a technically hard task; we believe that the α-Hölder functions are good representatives of the generic behavior of α-regular functions. This is supported by the results in Table I, which show that all α-regular functions have the same optimal information rate.

To conclude this Subsection, we summarize the conclusions distilled from analyses of the achievable information rates for oracle priors. In order to achieve the optimal information rate, the prior should be of Type-II, and it needs to satisfy the matching condition. Since in practice we (generally) do not know the true smoothness of the response surfaces, we cannot build a prior that satisfies the matching condition as we neither know which of the response surfaces is rougher, nor do we know the exact level of smoothness. Practical causal inference thus requires *adapting* the prior to the smoothness of the true function in a data-driven fashion; we discuss this in the next Subsection.

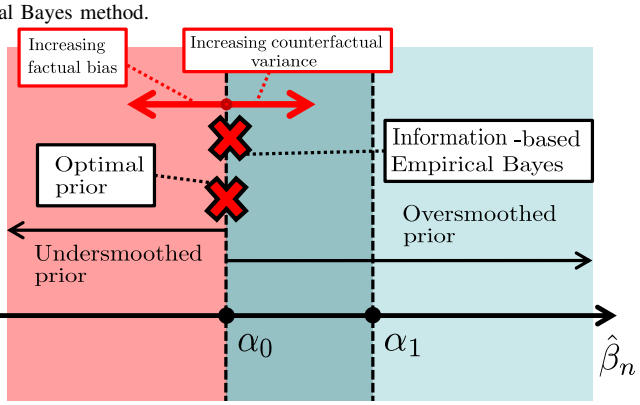
B. Rate-adaptive Data-driven Priors

Assume that the true response surfaces f_0 and f_1 belong to the Hölder spaces H^{α_0} and H^{α_1}, respectively. In this case, we know from the analysis in Subsection VI-B that the optimal prior is a Type-II prior over $H^{\alpha_0\wedge\alpha_1}$. In practice, we do not know $\alpha_0 \wedge \alpha_1$ a priori, and hence we need to obtain an estimate $\hat{\beta}_n$ of the optimal smoothness $\alpha_0 \wedge \alpha_1$ from the data in order to satisfy the matching condition in Theorem 4. Note that, unlike in standard nonparametric regression, estimating the optimal regularity parameter for causal inference ($\alpha_0 \wedge \alpha_1$) entails a mixed problem of testing and estimation, i.e. we need to test whether $\alpha_0 < \alpha_1$, and then estimate α_0. Hence, one would expect that the prior adaptation methods used in standard regression problems would not suffice in the causal inference setup. Prior adaptation can be implemented via *hierarchical Bayes* or *empirical Bayes* methods. Hierarchical Bayes methods specify a prior over $\alpha_0 \wedge \alpha_1$ (also known as the *hyper-prior* [24]), and then obtain a posterior over priors in a fully Bayesian fashion. Empirical Bayes simply obtains a point estimate $\hat{\beta}_n$ of $\alpha_0 \wedge \alpha_1$, and then conducts inference via the prior specified by $\hat{\beta}_n$. We focus on empirical Bayes methods since the hierarchical methods are often impractically expensive in terms of memory and computational requirements. A prior $\Pi_{\hat{\beta}_n}$ induced by $\hat{\beta}_n$ (obtained via empirical Bayes) is called *rate-adaptive* if it achieves the optimal information rate, i.e. $I_n(\Pi_{\hat{\beta}_n}) = I^*_n$.

In the rest of this Subsection, we show that marginal likelihood maximization, which is the dominant strategy for empirical Bayes adaptation in standard nonparametric regression [24], [43], can fail to adapt to the optimal smoothness $\alpha_0 \wedge \alpha_1$ in the general case when $\alpha_0 \neq \alpha_1$. (This is crucial since in most practical problems of interest, the treated and control response surfaces have different levels of heterogeneity [20].) We then propose a novel *information-based empirical Bayes* strategy, and prove that it always asymptotically satisfies the matching condition in Theorem 4. Finally, we conclude the Subsection by identifying candidate function spaces over which we can define the prior Π such that we are able to both adapt to functions in Hölder spaces, and also conduct practical Bayesian inference in an algorithmically efficient manner.



(a) An exemplary data-driven prior obtained via the likelihood-based empirical Bayes method.



(b) An illustration for the factual bias and counterfactual variance trade-off.

Fig. 4: Pictorial depiction for the operation of likelihood-based and information-based empirical Bayes adaptation methods.

1) Information-based Empirical Bayes and the failure of Maximum Likelihood-based Prior Adaptation: To see why the marginal likelihood-based empirical Bayes method may fail in adapting priors for causal inference, consider the following example. Suppose that $f_0 \in H^{\alpha_0}$ and $f_1 \in H^{\alpha_1}$, where $\alpha_0 < \alpha_1$. Let $\Pi^{\circ}_{\hat{\beta}_n}$ be a Type-I data-driven prior, where $\hat{\beta}_n$ is an empirical Bayes estimate of the optimal smoothness $\alpha_0 \wedge \alpha_1$. For the likelihood-based empirical Bayes, $\hat{\beta}_n$ is obtained by maximizing the marginal likelihood $d\mathbb{P}(\mathcal{D}_n \,|\, \beta)$ with respect to β. Note that since f_0 and f_1 possess different smoothness parameters, then the "true" model for generating $\mathcal{D}_n$ has is characterized by a likelihood function $d\mathbb{P}(\mathcal{D}_n \,|\, \alpha_0, \alpha_1)$.

Assume that the true model $d\mathbb{P}(\mathcal{D}_n \,|\, \alpha_0, \alpha_1)$ is *identifiable*, i.e. the mapping $(\alpha_0, \alpha_1) \mapsto \mathbb{P}$ is one-to-one. Type-I priors re-parametrize the observation model so that the likelihood function $d\mathbb{P}(\mathcal{D}_n \,|\, \beta)$ is parametrized with a single smoothness parameter β. Hence, as long as $\alpha_0 \neq \alpha_1$, the new parametrization renders an unidentifiable model, since the mapping $\beta \mapsto \mathbb{P}$ is not one-to-one (i.e. different combinations of α_0 and α_1 can map to the same β). This means that likelihood-based empirical Bayes can never satisfy the matching condition in Theorem 4, even in the limit of infinite samples ($n \uparrow \infty$). In most practical Bayesian models (e.g. Gaussian processes), the empirical Bayes estimate $\hat{\beta}_n$ will be in the interval (α_0, α_1) with high probability as depicted in Figure 4a. This means that with high probability, the likelihood-based empirical Bayes method will prompt an *oversmoothed* prior, from which all draws are smoother than the true ITE function, leading to a poor information rate. This problem is not specific to Type-I priors, but would arise in Type-II priors as well if the prior fixes the smoothness parameters for both components of its vector-valued output.

The failure of likelihood-based empirical Bayes in the causal inference setup is not surprising as maximum likelihood adaptation is only optimal in the sense of minimizing the Kullback-Leibler loss for the individual potential outcomes. *Optimal prior adaptation in our setup should be tailored to the causal inference task*. Hence, we propose an information-based empirical Bayes scheme in which, instead of maximizing the marginal likelihood, we pick the smoothness level $\hat{\beta}_n$ that minimizes the posterior Bayesian KL divergence defined in (4), i.e.

$$\hat{\beta}_n = \arg\min_{\beta} \mathbb{E}_{f_0, f_1 \sim d\Pi_\beta(. \,|\, \mathcal{D}_n)}[\, \mathbb{D}_n(\Pi_\beta; f_0, f_1) \,|\, \mathcal{D}_n \,]. \quad (10)$$

The information-based empirical Bayes estimator is simply a Bayesian estimator of β with the loss function defined as the information-theoretic criterion for the quality of the posterior distribution of the ITE function $T(x)$ (see (4)). Unlike the likelihood-based method, the objective in (10) is an direct measure for the quality of causal inference conducted with a prior Π_β. In the following Theorem, we show that the information-based empirical Bayes asymptotically satisfies the matching condition in Theorem 4.

Theorem 5. (Asymptotic Matching) Suppose that f_0 and f_1 are in Hölder spaces H^{α_0} and H^{α_1}, respectively. Let Π_β be a Type-I (or Type-II) prior defined over H^β. If $\hat{\beta}_n$ is obtained as in (10), then we have that $\hat{\beta}_n \xrightarrow{p} (\alpha_0 \wedge \alpha_1)$.
Proof. See Appendix D. ■

Theorem 5 says that the information-based empirical Bayes estimator is consistent. That is, the estimate $\hat{\beta}_n$ will eventually converge to the optimal smoothness $\alpha_0 \wedge \alpha_1$ as $n \to \infty$. Note that this is a weaker result than adaptivity: consistency of $\hat{\beta}_n$ does not imply that the corresponding prior will necessarily achieve the optimal information rate. However, the consistency result in Theorem 5 is both strongly suggestive of adaptivity, and also indicative of the superiority of the information-based empirical Bayes method to the likelihood-based approach.

Note that, while information-based empirical Bayes guarantees the recovery of the optimal smoothness $\alpha_0 \wedge \alpha_1$ (in the asymptotic sense), it still oversmoothes the prior for the smoother response surface, even when using a Type-II prior, as long as the prior restricts itself to a single smoothness parameter β for both f_0 and f_1. This can problematic if we wish the posterior credible interval on $T(x)$ to be "honest", i.e. possess frequentist coverage [7], [36], [44]. A more flexible Type-II prior that assigns different smoothness parameters β_0 and β_1 to response surfaces f_0 and f_1 can potentially guarantee honest frequentist coverage in a manner similar to that provided by causal forests [7]. As a consequence of Theorem 1, it turns out that the information-based empirical Bayes estimator in (10) is structurally similar to the risk-based empirical Bayes estimator proposed in [24]. Hence, we conjecture that our proposed empirical Bayes procedure can guarantee frequentist coverage for the estimated causal effects under some conditions (the polished tail condition in [24]).

2) Concrete Priors for Bayesian Causal Inference: So far we have been assuming that, given that f_0 and f_1 belong to Hölder spaces, the prior Π places a probability distribution over a Hölder space with a (data-driven) regularity parameter. In practice, we cannot build an inference algorithm that operates on a general Hölder space: we need to specify simpler function spaces over which machine learning/Bayesian inference methods become feasible. Two function spaces that are most commonly used in Bayesian nonparametric inference are the reproducing kernel Hilbert space (RKHS) [45]–[47], and the space of piece-wise constant functions (trees)[8], [16], [16]. The machine learning objects operating on those spaces are *Gaussian processes* (GPs), and *Bayesian additive regression trees* (BART), respectively. BART was especially proven successful in causal inference problems, and was one of the winning algorithms in the 2016 Atlantic Causal Inference Conference Competition[5].

In order to decide which function spaces we would build our inference procedure on, we evaluated the information rates achievable by oracle priors in the form of GPs and BART. The following is a summary of our analysis: details are provided in Appendix E. Let $\Pi_\beta(\text{GP})$ be a (Type-II) Gaussian process prior over a RKHS, defined by a Matérn covariance kernel with parameter β, then the achievable information rate is $I_n(\Pi_\beta(\text{GP}); H^{\alpha_0}, H^{\alpha_1}) \asymp n^{-2(\beta \wedge \alpha_0 \wedge \alpha_1)/(2\beta + d)}$. Hence, the optimal information rate is achieved if the matching condition is satisfied (Theorem 4). The squared exponential kernel displayed a suboptimal information rate of $(\log(n))^{-(\alpha_0 \wedge \alpha_1)/2 + d/4}$, whereas spline kernels achieved a rate of $(n/\log(n))^{-2(\beta \wedge \alpha_0 \wedge \alpha_1)/(2\beta + d)}$, which when satisfying the matching condition, becomes a near-optimal rate (optimal up to a logarithmic factor). The analyses in Appendix E was based on the work of van der Vaart and van Zanten in [27].

Since BART places a prior on a space of non-differentiable (piece-wise constant) functions, one would expect that the information rates achieved by BART would be inferior to those achieved by a GP. We show in Appendix E that a

[5]http://jenniferhill7.wixsite.com/acic-2016

carefully designed BART can only achieve a near-optimal rate of $(n/\log(n))^{-2(\beta\wedge\alpha_0\wedge\alpha_1)/(2\beta+d)}$ [16], [22], [23]. Our conclusion was that GPs are better choice for causal modeling, not only because it can achieve better rates than BART, but also because its relatively tractable nature would allow for an easy implementation for the information-based empirical Bayes scheme in (10).

VII. Practical Rate-adaptive Causal Inference with Multitask Gaussian Process Priors

The previous Section provided a detailed recipe for the informationally optimal Bayesian causal inference procedure. In particular, inference should be conducted through a Type-II Gaussian process prior on an RKHS space (Theorem 3 and Subsection VI-B). Moreover, the RKHS space should be defined through a Matérn covariance kernel with parameters β_0 and β_1 for response surfaces f_0 and f_1 (Subsection VI-B), and the parameters $\beta = (\beta_0, \beta_1)$ should be optimized via the information-based empirical Bayes procedure in (10). In this Section, we construct a practical learning algorithm that follows this recipe.

Type-II GP priors place a probability distribution on functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^2$ in a *vector-valued Reproducing Kernel Hilbert Space* (vvRKHS). A vvRKHS $\mathcal{H}_{\mathbf{K}}$ is equipped with an inner product $\langle ., . \rangle_{\mathcal{H}_{\mathbf{K}}}$, and a *reproducing kernel* $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$, where $\mathbf{K}$ is a (symmetric) positive semi-definite matrix-valued function [5], [46]–[48]. Note that by operating in a vvRKHS we get the algorithmic advantage of being able to conduct posterior inference in an infinite-dimensional function space by estimating a finite number of coefficients evaluated at the input feature points (this is a consequence of the well-known *representer Theorem* [49]). GP regression in vvRKHS is often associated with *multi-task learning* [47], and the corresponding GP is often known as a *multi-task* GP [48]. Multi-task learning is a common setup in machine learning where one model shares parameters between different tasks to improve statistical efficiency. The results of Theorem 3 can be thought of as suggesting multi-task learning as a framework for causal inference, where learning each of the potential outcomes (f_0 and f_1) is thought of as a separate learning task, and a single model is used to execute the two tasks simultaneously.

We chose the Matérn covariance kernel as the underlying regularity of the vvRKHS since it can achieve the optimal information rate (see Appendix E). In order to avoid under-smoothing any of the two surfaces, we also chose to assign separate smoothness parameters β_0 and β_1 to f_0 and f_1, respectively. Standard *intrinsic coregionalization models* for vector-valued kernels impose the same covariance parameters for all outputs [48], which implies that the prior will have the same smoothness on both f_0 and f_1. Thus, we constructed a *linear model of coregionalization* (LMC) [47], which mixes two intrinsic coregionalization models as follows

$$\mathbf{K}_\theta(x, x') = \mathbf{A}\, k_0(x, x') + \mathbf{B}\, k_1(x, x'),$$

where $k_\omega(x, x') = \text{Matérn}(\beta_\omega)$, $\omega \in \{0, 1\}$, whereas $\mathbf{A}$ and $\mathbf{B}$ are given by

$$\mathbf{A} = \begin{bmatrix} a_{00}^2 & a_{01} \\ a_{10} & \epsilon \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} \epsilon & b_{01} \\ b_{10} & b_{11} \end{bmatrix}, \tag{11}$$

where $\epsilon \to 0$ is a small positive number that is determined a priori and kept fixed during the prior adaptation procedure. The LCM kernel structure in (11) ensures that the response surfaces f_0 and f_1 have smoothness levels β_0 abd β_1 respectively. The constant ϵ ensures that $\mathbf{K}_\theta(x, x')$ is positive semi-definite for any selection of the other parameters. The parameters a_{00} and b_{11} represent the variances of f_0 and f_1, whereas all other variables $(a_{01}, a_{10}, b_{01}, b_{10})$ are correlation variables that share information among the two learning tasks (learning f_0 and f_1). The set of *all* kernel parameters is denoted as β. Given a set of "hyper-parameters" β, the ITE function estimate $\hat{T}_n$ is obtained in terms of the posterior mean[6] as follows: $\hat{T}_n = \mathbb{E}_{\Pi_\beta}[\,\mathbf{f}^T\mathbf{v} \,|\, \mathcal{D}_n]$, where $\mathbf{v} = [-1, 1]^T$.

Now that we completely specified the multi-task GP prior for a given hyper-parameter set β, the only remaining ingredient in the recipe is to implement the information-based empirical Bayes adaptation criterion in (10). The following Theorem gives an insightful decomposition of the information-based empirical Bayes objective for the multi-task GP model. (In the following Theorem, $\mathbf{Y}^{(\mathbf{W})} = [Y_i^{(\omega_i)}]_i$ and $\mathbf{Y}^{(\mathbf{1-W})} = [Y_i^{(1-\omega_i)}]_i$ are vectors comprising all factual and counterfactual outcomes associated with an observational dataset $\mathcal{D}_n$.)

Theorem 6. (Factual bias and counterfactual variance decomposition) The minimizer β^* of the information-based empirical Bayes adaptation criterion in (10) is given by

$$\arg\min_\beta \underbrace{\left\| \mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_{\Pi_\beta}[\,\mathbf{f} \,|\, \mathcal{D}_n\,] \right\|_2^2}_{\text{Factual bias}} + \underbrace{\left\| \text{Var}_{\Pi_\beta}[\,\mathbf{Y}^{(\mathbf{1-W})} \,|\, \mathcal{D}_n\,] \right\|_1}_{\text{Counterfactual variance}},$$

where Var_{Π_β} is the posterior variance and $\|.\|_p$ is the p-norm.
Proof. See Appendix F. ■

Theorem 6 states that, for a multi-task GP prior, the information-based empirical Bayes criterion in (10) decomposes to *factual bias* and *counterfactual variance* terms[7]. The factual bias term quantifies the empirical error in the observed factual outcome that results from selecting a particular smoothness level β. In that sense, the factual bias is a measure of the goodness-of-fit for the posterior mean resulting from a prior smoothness β. On the other hand, the counterfactual variance term quantifies the posterior uncertainty that would be induced in the unobserved counterfactual outcomes when selecting a smoothness level β. A small value for β would lead to a rough posterior mean function, which corresponds to a good empirical fit for the data. On the contrary, a small value for β would induce large uncertainty in the unobserved outcomes, which corresponds

[6]Closed-form expressions for the posterior mean of a multi-task GP can be found in [46], [48].

[7]The objective function in Theorem 6 can be easily optimized via a leave-one-out cross-validation procedure. Refer to [5] for a detailed explanation.

to large uncertainty in the counterfactual outcomes. The couterfactual variance thus acts as a regularizer for the factual bias that helps solving the joint testing-estimation problem of identifying the minimum of α_0 and α_1, and estimating the value of $\alpha_0 \wedge \alpha_1$. That is, the regularizer attempts to protect the prior from falsely recognizing either α_0 or α_1 as being very low just because it over-fit the factual outcomes, and hence underestimating the true optimal smoothness $\alpha_0 \wedge \alpha_1$, thereby undersmoothing the prior and giving rise to a suboptimal information rate. The two terms work in opposite directions as shown in Figure 4b: factual bias pushes for undersmoothed priors and counterfactual variance pushes for oversmoothed priors. Theorem 6 says that the resulting prior will lie on the optimal boundary in the large data limit.

Finally, we note that the factual bias and counterfactual variance trade-off automatically handles selection bias. That is, when there is a poor overlap between the treated and control populations, the posterior counterfactual variances would tend to be higher, and the information-based empirical Bayes method would tend to oversmooth the prior rather than fitting the factual data. *Selection bias does not affect the optimal information rate, but it does affect the optimal strategy for achieving that rate as long as we decide to share parameters and data points between our models for the potential outcomes.*

VIII. Experiments

We sought to evaluate the finite-sample performance of the Bayesian causal inference procedure proposed in Section VII, and compare it with state-of-the-art causal inference models. Causal inference models are hard to evaluate [6], and obviously, it is impossible to validate a causal model using real-world data due to the absence of counterfactual outcomes. A common approach for evaluating causal models, which we follow in this paper, is to validate the model's predictions/estimates in a semi-synthetic dataset for which artificial counterfactual outcomes are randomly generated via a predefined probabilistic model. To ensure a fair and objective comparison, we did not design the semi-synthetic dataset used in the experiments by ourselves, but rather used the (standard) semi-synthetic experimental setup designed by Hill in [8]. In this setup, the features and treatment assignments are real but outcomes are simulated. The experimental setup was based on the IHDP dataset, a public dataset for data from a randomized clinical trial. We describe the dataset in more detail in the following Subsection.

A. The IHDP dataset

The Infant Health and Development Program (IHDP) is an interventional program that is intended to enhance the cognitive and health status of low birth weight, premature infants through pediatric follow-ups and parent support groups [8]. The semi-simulated dataset in [8], [18], [28] is based on features for premature infants enrolled in a real randomized experiment that evaluated the impact of the IHDP on the subjects' IQ scores at the age of three. Because the data was originally collected from a randomized trial, selection bias was introduced in the treatment assignment variable by removing a subset of the treated population. All outcomes (response surfaces) are simulated. The response surface data generation process was not designed to favor our method: we used the standard non-linear "Response Surface B" setting in [8]. The dataset comprises 747 subjects (608 control and 139 treated), and there are 25 features associated with each subject.

B. Benchmarks

We compared our algorithm with various causal models and standard machine learning benchmarks which we list in what follows: ♣ Tree-based methods (BART [8], [16], [22], causal forests (CF) [7], [36], ♠ Balancing counterfactual regression (balancing neural networks (BNN) [18], and counterfactual regression with Wasserstein distance metric (CFRW) [28]), ★ Propensity-based and matching methods (k nearest-neighbor (kNN), propensity score matching (PSM)), a ♢ nonparametric spline regression model (causal MARS [19]), and ⊙ Doubly-robust methods (Targeted maximum likelihood (TML) [50]). We also compared the performance of our model with standard machine learning benchmarks, including linear regression (LR), random forests (RF), AdaBoost, XGBoost, and neural networks (NN). We evaluated two different variants of all the machine learning benchmarks: a □ Type-I regression structure, in which we use the treatment assignment variable as an input feature to the machine leaning algorithm, and a ⊗ Type-II regression structure, in which we fit two separate models for treated and control populations. We compare all these benchmarks with our proposed model: a Type-II multi-task GP prior (MTGP) with a Matérn kernel optimized through information-based empirical Bayes. We also compare the proposed model with a Type-I multi-task GP model (with a Matérn kernel) optimized through likelihood-based empirical Bayes in order to verify the conclusions drawn from our analyses.

All machine learning benchmarks had their hyperparameters optimized via grid search using a held-out validation set. Hyper-parameter optimization was using the mean square error in the observed factual outcomes as the optimization objective. For BART, we used the default prior as in [8], and did not tune the model's hyper-parameters. For BNN and CFRW, we used the neural network configurations reported in [18] and [28]. Causal MARS was implemented as described in [19]. PSM was implemented as described in [8], and its performance was obtained by assuming that every patient's estimated ITE is equal to the average treatment effect estimated by PSM. All benchmarks were implemented in `Python`, with the exception of BART, causal forests and TMLE, all of which were implemented in `R`. We used the `R` libraries `bartMachine`, `grf`, and `tmle` for the implementation of BART, causal forests and TMLE, respectively. Our method was implemented in `Python` using `GPy`, an open source library for Gaussian processes [51].

C. Evaluation

We evaluate the performance of all benchmarks by reporting the square-root of the PEHE. The empirical PEHE is estimated as $\text{PEHE} = \frac{1}{n}\sum_{i=1}^{n}((f_1(X_i) - f_0(X_i)) - \mathbb{E}[Y_i^{(1)} - Y_i^{(0)}|X_i =$

TABLE II: SIMULATION RESULTS FOR THE IHDP DATASET. NUMERICAL VALUES CORRESPOND TO THE AVERAGE PEHE $\pm$ 95% CONFIDENCE INTERVALS.

		In-sample $\sqrt{\text{PEHE}}$	Out-of-sample $\sqrt{\text{PEHE}}$			In-sample $\sqrt{\text{PEHE}}$	Out-of-sample $\sqrt{\text{PEHE}}$
♡	**MTGP (Type-II)**	**0.59 ± 0.01**	**0.76 ± 0.01**	♢	Causal MARS	1.66 ± 0.10	1.74 ± 0.10
	GP (Type-I)	1.85 ± 0.12	2.10 ± 0.16	□	NN-1	3.56 ± 0.20	3.64 ± 0.20
♣	BART	2.0 ± 0.13	2.2 ± 0.15		AdaBoost-1	4.53 ± 0.31	4.56 ± 0.31
	CF	2.4 ± 0.14	2.8 ± 0.18		XGBoost-1	2.97 ± 0.21	3.04 ± 0.21
	RF-1	2.7 ± 0.24	2.9 ± 0.25		LR-1	5.06 ± 0.35	5.05 ± 0.35
	RF-2	1.4 ± 0.07	2.2 ± 0.16	⊗	NN-2	3.36 ± 0.13	3.46 ± 0.14
♠	BNN	2.1 ± 0.11	2.2 ± 0.13		AdaBoost-2	2.40 ± 0.17	2.79 ± 0.20
	CFRW	1.0 ± 0.07	1.2 ± 0.08		XGBoost-2	1.46 ± 0.08	1.98 ± 0.15
★	kNN	2.69 ± 0.17	4.0 ± 0.21		LR-2	1.85 ± 0.10	1.94 ± 0.12
	PSM	4.9 ± 0.31	4.9 ± 0.31	⊙	TMLE	5.27 ± 0.35	5.27 ± 0.35

$x])^2$, where $f_1(X_i) - f_0(X_i)$ is the estimated treatment effect. We evaluate the PEHE of all algorithms via a Monte Carlo simulation with 1000 realizations of the IHDP semi-synthetic model, where in each experiment/realization we run all the benchmarks with a 60/20/20 train-validation-test splits. (For models that do not need hyper-parameter tuning, such as BART and our GP models, the entire training set is used to compute the posterior distributions.) We report both the in-sample and out-of-sample PEHE estimates: the former corresponds to the accuracy of the estimated ITE in a retrospective cohort study, whereas the latter corresponds to the performance of a clinical decision support system that provides out-of-sample patients with ITE estimates [28]. The in-sample PEHE results are non-trivial since we never observe counterfactuals even in the training phase. Recall that, from Theorem 1, we know that the achieved information rate by a Bayesian inference procedure is equivalent to the PEHE estimation rate. Thus, the PEHE performance is a direct proxy of the achieved information rate, and since it is an essentially frequentist quantity, we can use it to compare the performance of our model with the frequentist benchmarks.

D. Results

As can be seen in Table II, the proposed Bayesian inference algorithm (Type-II MTGP) outperforms all other benchmarks in terms of the (in-sample and out-of-sample) PEHE. This result suggests that the proposed model was capable of adapting its prior to the data, and may have achieved the optimal (or a near-optimal) information rate. The PEHE results in Table II are the averages of 1000 experiments with 1000 different random realizations of the semi-synthetic outcome model. This means that our algorithm is consistently outperforming all other benchmarks as it is displaying a very tight confidence interval.

The benefit of the information-based empirical Bayes method manifests in the comparison with the Type-I MTGP prior optimized via likelihood-based empirical Bayes. The performance gain of the Type-II MTGP prior with respect to the Type-I MTGP prior results from the fact that the two response surfaces in the synthetic outcomes model have different levels of heterogeneity (the control response is non-linear whereas the treated response is linear. See the description of Response surface B in [8]). Our algorithm is also performing better than all other nonparametric tree-based algorithms. This is expected since, as we have discussed earlier in Subsection VI-B, an oracle BART prior can only achieve the optimal information rate up to a logarithmic factor. With the default prior, it is expected that BART would display a slow information rate as compared to our adapted, information-optimal Matérn kernel prior. Similar insights apply to the frequentist random forest algorithms, which approximates the true regression functions through non-differentiable, piecewise functions (trees), and hence is inevitably suboptimal in terms of the achievable minimax estimation rate.

Our model also outperforms all the standard machine learning benchmarks, whether the ones trained with a Type-I regression structure, or those trained with a Type-II structure. We believe that this is because our model outperforms the standard machine learning benchmarks since the information-based empirical Bayes method provides a natural protection against selection bias (via the counterfactual variance regularization). Selection bias introduces a mismatch between the training and testing datasets for all the machine learning benchmarks (i.e. a covariate shift [18]), and hence all machine learning methods exhibit high generalization errors.

IX. CONCLUSIONS

In this paper, we studied the problem of estimating the causal effect of an intervention on *individual* subjects using observational data in the Bayesian nonparametric framework. We characterized the optimal Kullback-Leibler information rate that can be achieved by any learning procedure, and showed that it depends on the dimensionality of the feature space, and the smoothness of the "rougher" of the two potential outcomes. We characterized the priors that are capable of achieving the optimal information rates, and proposed a novel empirical Bayes procedure that is adapts the Bayesian prior to the causal effect function through an information-theoretic criterion. Finally, we used the conclusions drawn from our analysis and designed a practical Bayesian causal inference algorithm with a multi-task Gaussian process, and showed that it significantly outperforms the state-of-the-art causal inference models through experiments conducted on a standard semi-synthetic dataset.

APPENDIX A
PROOF OF THEOREM 2

Let δ_ω be the solution for $H(\delta_\omega;\, \mathcal{F}^{\alpha_\omega}) \asymp n\,\delta_\omega^2$, $\omega \in \{0,1\}$. We will prove that the optimal rate is $\Theta(\delta_0^2 \wedge \delta_1^2)$ by first showing that $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}) = \Omega(\delta_0^2 \wedge \delta_1^2)$, and then showing that $I_n^*(\mathcal{F}^{\alpha_0}, \mathcal{F}^{\alpha_1}) = O(\delta_0^2 \wedge \delta_1^2)$. 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

$$I(T;\hat{T}) \overset{(*)}{\leq} I(T;\mathcal{D}_n) \overset{(\star)}{\leq} \sup_{\Pi} I(T;\mathcal{D}_n), \tag{A.12}$$

where $(*)$ follows from the *data processing* inequality [31], and the supremum in $(\star)$ is taken over all possible priors. $I(T;\hat{T})$ is bounded below by the *rate-distortion* function

$$I(T;\hat{T}) \geq \inf_{T,\hat{T}:\, \mathbb{E}\|T-\hat{T}\|_2^2 \leq R_\Pi^*} I(T;\hat{T}), \tag{A.13}$$

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 $(T,\hat{T})$, and with $\|.\|_2^2$ being the $L_2(\mathbb{P})$ norm. Combining (A.12) and (A.13), we can upper and lower bound the mutual information $I(T;\hat{T})$ as follows

$$\inf_{\mathbb{E}\|T-\hat{T}\|_2^2 \leq R_\Pi^*} I(T;\hat{T}) \leq I(T;\hat{T}) \leq \sup_{\Pi} I(T;\mathcal{D}_n). \tag{A.14}$$

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 [30], [52]. 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:

$$(f_0, f_1) \to \mathcal{D}_n \to (\hat{f}_0, \hat{f}_1) \to Q(\hat{f}_0, \hat{f}_1). \tag{A.15}$$

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 the true response surfaces 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

$$I(\tilde{T};\hat{T}) \geq I(\tilde{T};Q(\hat{T})). \tag{A.16}$$

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

$$\|\hat{T} - Q(\hat{T})\| \leq \|\hat{T} - \tilde{T}\|. \tag{A.17}$$

Thus, the error event implies that $\delta \leq \|Q(\hat{T}) - \tilde{T}\|$. Using the triangular inequality, (A.17) can be further bounded as follows:

$$\begin{aligned}\delta &\leq \|Q(\hat{T}) - \tilde{T}\| = \|Q(\hat{T}) - \hat{T} + \hat{T} - \tilde{T}\| \\ &\leq \|Q(\hat{T}) - \hat{T}\| + \|\hat{T} - \tilde{T}\| \\ &\leq 2\,\|\hat{T} - \tilde{T}\| \implies \|\hat{T} - \tilde{T}\| \geq \frac{\delta}{2}.\end{aligned} \tag{A.18}$$

Let P_e be the probability of the error event $\{\tilde{T} \neq Q(\hat{T})\}$. From (A.18), the probability of the error event can be bounded above as follows

$$\begin{aligned}P_e &:= \mathbb{P}(\{\tilde{T} \neq Q(\hat{T})\}) \\ &= \mathbb{P}(\|Q(\hat{T}) - \tilde{T}\| \geq \delta) = \mathbb{P}(\|\hat{T} - \tilde{T}\| \geq \delta/2) \\ &= \mathbb{P}(\|\hat{T} - \tilde{T}\|_2^2 \geq \delta^2/4) \\ &\overset{(\bullet)}{\leq} \frac{4}{\delta^2}\,\mathbb{E}[\|\hat{T} - \tilde{T}\|_2^2] \leq \frac{4}{\delta^2}\,R_\Pi^*,\end{aligned} \tag{A.19}$$

where $(\bullet)$ is an application of Markov's inequality. By combining (A.16) with the result in (A.19), the lower bound in (A.14) 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

$$\begin{aligned}I(\tilde{T};Q(\hat{T})) &= I(\tilde{f}_1 - \tilde{f}_0; Q(\hat{f}_1 - \hat{f}_0)) \\ &\overset{(\odot)}{\leq} I(\tilde{f}_0, \tilde{f}_1; Q(\hat{f}_1 - \hat{f}_0)) \\ &\leq I(\tilde{f}_0, \tilde{f}_1; Q(\hat{f}_0), Q(\hat{f}_1)) \\ &= I(\tilde{f}_0; Q(\hat{f}_0)) + I(\tilde{f}_1; Q(\hat{f}_1)) \\ &\leq 2\max\{I(\tilde{f}_0; Q(\hat{f}_0)), I(\tilde{f}_1; Q(\hat{f}_1))\},\end{aligned} \tag{A.20}$$

where $(\odot)$ follows from the data processing inequality. The mutual information $I(\tilde{T};Q(\hat{T}))$ can be written in terms of the KL divergence as [31]

$$\begin{aligned}I(\tilde{T};Q(\hat{T})) &= D(\mathbb{P}(\tilde{T};Q(\hat{T}))\,||\,\mathbb{P}(\tilde{T})\cdot\mathbb{P}(Q(\hat{T}))) \\ &\geq D(\text{Bern}(P_e)\,||\,\text{Bern}(1-1/n)) \\ &= P_e \log\left(\frac{P_e}{1-1/\tilde{M}_T}\right) + (1-P_e)\log\left(\frac{1-P_e}{1/\tilde{M}_T}\right) \\ &= -h(P_e) + \log(\tilde{M}_T) - P_e\,\log(\tilde{M}_T - 1) \\ &\geq -\log(2) + \log(\tilde{M}_T) - P_e\,\log(\tilde{M}_T),\end{aligned} \tag{A.21}$$

where $h(.)$ is the binary entropy. From (A.21), we have that

$$P_e \geq 1 - \frac{I(\tilde{T};Q(\hat{T})) + \log(2)}{\log(\tilde{M}_T)}, \tag{A.22}$$

which is an incarnation of Fano's inequality. By combining (A.20) with (A.22), we have the following inequality

$$P_e \geq 1 - \frac{I(\tilde{f}_0; Q(\hat{f}_0)) \vee I(\tilde{f}_1; Q(\hat{f}_1)) + \log(2)}{\frac{1}{2}\,\log(\tilde{M}_T)}. \tag{A.23}$$

From (A.19), the minimax risk R^*_Π is bounded below as follows

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

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

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

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

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

From Theorem 1, we have that

$$R^*_\Pi \leq 2(\sigma_0^2 + \sigma_1^2) \cdot I^*_n,$$

and hence we have that

$$\begin{aligned} I^*_n &\geq \frac{\delta^2}{8(\sigma_0^2+\sigma_1^2)}\left(1 - \frac{4\, n\, \delta^2 + \log(2)}{\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1}))}\right). \\ &\gtrsim \delta^2 - \frac{\delta^4\, n + \delta^2}{\log(N(\delta, \mathcal{F}^{\alpha_0}) \vee N(\delta, \mathcal{F}^{\alpha_1}))}. \end{aligned} \tag{A.24}$$

Since I^*_n is strictly positive, then the information rate is bounded below by δ^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

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

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 (A.25) is written as

$$H(\delta, \mathcal{F}^{\alpha_0}) \vee H(\delta, \mathcal{F}^{\alpha_1}) \asymp \delta^2\, n. \tag{A.26}$$

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 (A.26) is given by $\delta^* = \delta_0 \wedge \delta_1$, where δ_ω is the solution to the equation

$$H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0,1\}. \tag{A.27}$$

The equation in $(A.27)$ has a solution for all n when the function space $\mathcal{F}^{\alpha_\omega}$ has a polynomial or a logarithmic metric entropy [29], which is the case for all function spaces of interest (see Table I for evaluations of $\delta_0 \wedge \delta_1$ for various function spaces). It follows from (A.24) and (A.27) that

$$I^*_n = \Omega(\delta_0^2 \wedge \delta_1^2),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0,1\}. \tag{A.28}$$

We now focus on upper bounding I^*_n. From [52], we know that the minimax information rate is upper bounded by the channel capacity in (A.12), which is further bounded above by the covering numbers as follows

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

For δ satisfying (A.27), we have that

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

and hence $\mathcal{I}_n \lesssim \delta_0^2 \wedge \delta_1^2$. It follows that

$$I^*_n = O(\delta_0^2 \wedge \delta_1^2),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0,1\}. \tag{A.29}$$

By combining (A.28) and (A.29), we have that $I^*_n = \Omega(\delta_0^2 \wedge \delta_1^2)$ and $I^*_n = O(\delta_0^2 \wedge \delta_1^2)$, and hence it follows that

$$I^*_n = \Theta(\delta_0^2 \wedge \delta_1^2),\ H(\delta_\omega, \mathcal{F}^{\alpha_\omega}) \asymp \delta^2_\omega\, n,\ \omega \in \{0,1\}. \tag{A.30}$$

APPENDIX B
PROOF OF THEOREM 3

When $f_0 \in H^{\alpha_0}_{\mathcal{P}_0}$ and $f_1 \in H^{\alpha_1}_{\mathcal{P}_1}$, the metric entropy of the function spaces $H^{\alpha_0}_{\mathcal{P}_0}$ and $H^{\alpha_1}_{\mathcal{P}_1}$ are given by

$$H(\delta, H^{\alpha_0}_{\mathcal{P}_0}) \asymp \delta^{\frac{-|\mathcal{P}_0|}{\alpha_0}},\ H(\delta, H^{\alpha_1}_{\mathcal{P}_1}) \asymp \delta^{\frac{-|\mathcal{P}_1|}{\alpha_1}}.$$

From Theorem 2, we know that the optimal information rate is given by $I^*_n(H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}) \asymp \delta_0^2 \wedge \delta_1^2$, where δ_0 and δ_1 are the solutions for

$$\delta^{\frac{-|\mathcal{P}_0|}{\alpha_0}} \asymp n\,\delta_0^2,\ \delta^{\frac{-|\mathcal{P}_1|}{\alpha_1}} \asymp n\,\delta_1^2,$$

and hence we have that

$$\delta_\omega \asymp n^{\frac{-2\alpha_\omega}{2\alpha_\omega + |\mathcal{P}_\omega|}},\ \omega \in \{0,1\},$$

and thus the optimal information rate is given by

$$I^*_n(H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}) \asymp n^{\frac{-2\alpha_0}{2\alpha_0+|\mathcal{P}_0|}} \wedge n^{\frac{-2\alpha_1}{2\alpha_1+|\mathcal{P}_1|}}.$$

From Theorem 1, we know that

$$\mathbb{D}_n(\Pi; f_0, f_1, \gamma) \asymp \frac{1}{2(\sigma_0^2+\sigma_1^2)}\,\mathbb{E}\left[\,\|\,\mathbb{E}_\Pi[\,T\,|\,\mathcal{D}_n\,] - T\,\|_2^2\,\right].$$

The term $\mathbb{E}\left[\,\|\,\mathbb{E}_\Pi[\,T\,|\,\mathcal{D}_n\,] - T\,\|_2^2\,\right]$ on the right hand side can be upper bounded as follows [28]:

$$\begin{aligned} \mathbb{E}\left[\,\|\,\mathbb{E}[\,T\,|\,\mathcal{D}\,] - T\,\|_2^2\,\right] \leq\ & 2\,\bar{\gamma}\,\mathbb{E}\left[\,\|\,\mathbb{E}[\,f_0\,|\,\mathcal{D}\,] - f_0\,\|_2^2\,\right] \\ & + 2\,(1-\bar{\gamma})\,\mathbb{E}\left[\,\|\,\mathbb{E}[\,f_1\,|\,\mathcal{D}\,] - f_1\,\|_2^2\,\right], \end{aligned}$$

where $\bar{\gamma} = \mathbb{E}_x[\gamma(x)]$. For a Type-II prior $\Pi^{\infty\infty}_{\beta_{0,1}}$ over the two Hölder spaces $H^{\alpha_0}_{\mathcal{P}_0}$ and $H^{\alpha_1}_{\mathcal{P}_1}$, with $\beta_0 = \alpha_0$ and $\beta_1 = \alpha_1$, we achieve the minimax estimation rates for nonparametric regression over f_0 and f_1 as follows

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}[\,f_0\,|\,\mathcal{D}\,] - f_0\,\|_2^2\,\right] &\asymp n^{\frac{-2\alpha_0}{2\alpha_0+|\mathcal{P}_0|}}, \\ \mathbb{E}\left[\mathbb{E}[\,f_1\,|\,\mathcal{D}\,] - f_1\,\|_2^2\,\right] &\asymp n^{\frac{-2\alpha_1}{2\alpha_1+|\mathcal{P}_1|}}, \end{aligned}$$

and it follows that

$$\mathbb{D}_n(\Pi^{\infty\infty}_{\beta_{0,1}}; f_0, f_1, \gamma) = O\left(n^{\frac{-2\alpha_0}{2\alpha_0+|\mathcal{P}_0|}} \wedge n^{\frac{-2\alpha_1}{2\alpha_1+|\mathcal{P}_1|}}\right),$$

which matches the optimal information rate. Similarly, for a Type-I prior Π°_{β} over a Hölder space $H^{\beta}_{\mathcal{P}_0 \cup \mathcal{P}_1}$, the achieved information rate is

$$\mathbb{D}_n(\Pi^{\circ}_{\beta}; f_0, f_1, \gamma) \asymp n^{\frac{-2(\beta \wedge \alpha_0 \wedge \alpha_1)}{2\beta + |\mathcal{P}_0| + |\mathcal{P}_1| + 1}},$$

where the number of feature dimensions $|\mathcal{P}_0| + |\mathcal{P}_1| + 1$ correspond to all the relevant dimensions for the regression function $f(x, \omega)$. Since the rate $n^{\frac{-2\beta}{2\beta + |\mathcal{P}_0| + |\mathcal{P}_1| + 1}}$ is strictly slower than the optimal rate of $n^{\frac{-2\alpha_0}{2\alpha_0 + |\mathcal{P}_0|}} \wedge n^{\frac{-2\alpha_1}{2\alpha_1 + |\mathcal{P}_1|}}$ for all $\beta > 0$, it follows that

$$\mathbb{D}_n(\Pi^{\circ}_{\beta}; f_0, f_1, \gamma) \lesssim I^*_n(H^{\alpha_0}_{\mathcal{P}_0}, H^{\alpha_1}_{\mathcal{P}_1}).$$

Appendix C
Proof of Theorem 4

From Appendix B (and Theorem 1), we know that

$$\mathbb{D}_n(\Pi; f_0, f_1, \gamma) \asymp \frac{1}{2(\sigma_0^2 + \sigma_1^2)} \mathbb{E}\left[\| \mathbb{E}_{\Pi}[T \,|\, \mathcal{D}_n] - T \|_2^2 \right],$$

and the $L_2(\mathbb{P})$ risk $\mathbb{E}\left[\| \mathbb{E}_{\Pi}[T \,|\, \mathcal{D}_n] - T \|_2^2 \right]$ is upper bounded as follows:

$$\begin{aligned}\mathbb{E}\left[\| \mathbb{E}[T \,|\, \mathcal{D}] - T \|_2^2 \right] &\leq 2\,\bar{\gamma}\, \mathbb{E}\left[\| \mathbb{E}[\, f_0 \,|\, \mathcal{D}] - f_0 \|_2^2 \right] \\ &\quad + 2\,(1 - \bar{\gamma})\, \mathbb{E}\left[\| \mathbb{E}[\, f_1 \,|\, \mathcal{D}] - f_1 \|_2^2 \right],\end{aligned}$$

From Lemma 4 in [27], we know that

$$\mathbb{E}\left[\| \mathbb{E}[\, f_\omega \,|\, \mathcal{D}] - f_\omega \|_2^2 \right] \lesssim \ (\psi_\omega^{-1}(n))^2,$$

where

$$\psi_\omega(\delta) = \frac{1}{n^2}\, \log(N(\delta, \mathcal{F}^{\beta_\omega})) + \frac{1}{n^2}\, \delta^{\frac{2\beta_\omega + d - 2\alpha_\omega}{\alpha_\omega}},$$

for $\omega \in \{0, 1\}$, which for a Hölder space H^{α_ω} with metric entropy of $\log(N(\delta, H^{\beta_\omega})) = \delta^{\frac{-d}{\beta_\omega}}$ is given by

$$\psi_\omega(\delta) = \frac{1}{n^2}\, \delta^{\frac{d}{\beta_\omega}} + \frac{1}{n^2}\, \delta^{\frac{2\beta_\omega + d - 2\alpha_\omega}{\alpha_\omega}}.$$

Thus, an upper bound on the minimax estimation rate for f_ω is given by

$$\mathbb{E}\left[\| \mathbb{E}[\, f_\omega \,|\, \mathcal{D}] - f_\omega \|_2^2 \right] \lesssim n^{\frac{-2\beta_\omega}{2\beta_\omega + d}} + n^{\frac{-2\alpha_\omega}{2\beta_\omega + d}} \quad = n^{\frac{-2(\alpha_\omega \wedge \beta_\omega)}{2\beta_\omega + d}}.$$

Hence, the achieved information rate is upper bounded by

$$\begin{aligned} I_n &\lesssim 2(1 - \bar{\gamma})\, n^{\frac{-2(\alpha_0 \wedge \beta_0)}{2\beta_0 + d}} + 2\bar{\gamma}\, n^{\frac{-2(\alpha_1 \wedge \beta_1)}{2\beta_1 + d}} \\ &\asymp n^{\frac{-2(\alpha_0 \wedge \beta_0)}{2\beta_0 + d}} + n^{\frac{-2(\alpha_1 \wedge \beta_1)}{2\beta_1 + d}}. \end{aligned}$$

The tightness of these upper bounds have been demonstrated by Castillo in [53].

Appendix D
Proof of Theorem 5

The empirical smoothness estimate $\hat{\beta}_n$ is obtained by minimizing the objective

$$Q_n = \mathbb{E}_{f_0, f_1 \sim d\Pi_\beta(\,.\,|\,\mathcal{D}_n)}[\, \mathbb{D}_n(\Pi_\beta; f_0, f_1) \,|\, \mathcal{D}_n\,]. \qquad \text{(D.31)}$$

From the capacity-redundancy theorem [54], we know that the Bayesian risk above is upper bounded by the minimax risk (i.e. the information rate) and decays at the same rate. Hence, the optimal solution to $\hat{\beta}^*_n = \alpha_0 \wedge \alpha_1$ also minimizes Q_n. If the estimate $\hat{\beta}_n$ that maximizes the empirical objective Q_n is obtained by searching over values in a compact set $[0, \bar{\beta}]$ for some $\bar{\beta} > 0$ that contains the optimal solution $\alpha_0 \wedge \alpha_1$, then we know, by the consistency of M-estimators, that $\hat{\beta}_n \xrightarrow{p} \alpha_0 \wedge \alpha_1$ if Q_n satisfies Wald's regularity conditions [29].

Appendix E
Some Concrete Priors

Appendix F
Proof of Theorem 6

Minimizing the objective in (D.31) is equivalent to minimzing the posterior Bayesian $L_2(\mathbb{P})$ risk $R(\theta, \hat{\mathbf{f}}; \mathcal{D})$ for a point estimate $\hat{\mathbf{f}}$, which is given by

$$R(\theta, \hat{\mathbf{f}}; \mathcal{D}) = \mathbb{E}_\theta\left[\hat{\mathcal{L}}(\hat{\mathbf{f}}; \mathbf{K}_\theta, \mathbf{Y}^{(\mathbf{W})}, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})}) \,\middle|\, \mathcal{D} \right],$$

where the expectation in is taken with respect to $\mathbf{Y}^{(\mathbf{1}-\mathbf{W})}|\mathcal{D}$. The Bayesian risk can be written as

$$R(\theta, \hat{\mathbf{f}}; \mathcal{D}) = \int \hat{\mathcal{L}}(\hat{\mathbf{f}}; \mathbf{K}_\theta, \mathbf{Y}^{(\mathbf{W})}, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})})\, d\mathbb{P}_\theta(\mathbf{Y}^{(\mathbf{1}-\mathbf{W})}|\mathcal{D}).$$

The loss function $\hat{\mathcal{L}}$ conditional on a realization of the counterfactual outcomes is given by

$$\begin{aligned}&\hat{\mathcal{L}}(\hat{\mathbf{f}}; \mathbf{K}_\theta, \mathbf{Y}^{(\mathbf{W})}, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})}) = \\ &\frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mathbf{f}}^T(X_i)\mathbf{e} - (1 - 2W_i) \left(Y_i^{(1-W_i)} - Y_i^{(W_i)} \right) \right)^2.\end{aligned}$$

The optimal hyper-parameter and interpolant $(\hat{f}^*, \theta^*)$ are obtained through the following optimization problem in (F.1). The optimization problem can solved separately for θ and $\hat{f}$; we know from Theorem 1 that for any given θ, the optimal interpolant $\hat{\mathbf{f}} = \mathbb{E}_\theta[\mathbf{f} \,|\, \mathcal{D}]$. Hence, the optimal hyper-parameter θ^* can be found by solving the optimization problem in (F.2). The objective function R can thus be written as in (F.3) and further reduced as in (F.4).

Note that since $Y_i^{(W_i)} = f_{W_i}(X_i) + \epsilon_{i, W_i}$, then we have that $\mathbb{E}_\theta[f_{W_i}(X_i) \,|\, \mathcal{D}] = \mathbb{E}_\theta[Y_i^{(W_i)} \,|\, \mathcal{D}]$ and $\mathbb{E}_\theta[f_{1-W_i}(X_i) \,|\, \mathcal{D}] = \mathbb{E}_\theta[Y_i^{(1-W_i)} \,|\, \mathcal{D}]$. Therefore, we can evaluate the terms R_1, R_2 and R_3 as follows

$$\begin{aligned} R_1 &= \frac{1}{n} \sum_{i=1}^{n} \int (Y_i^{(W_i)} - \mathbb{E}_\theta[f_{W_i}(X_i) \,|\, \mathcal{D}])^2\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \int (Y_i^{(W_i)} - \mathbb{E}_\theta[Y_i^{(W_i)} \,|\, \mathcal{D}])^2\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}) \\ &= \frac{1}{n} \|\mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\mathbf{f} \,|\, \mathcal{D}]\|_2^2, \end{aligned} \qquad \text{(F.36)}$$

and

$$\begin{aligned} R_2 &= \frac{1}{n} \sum_{i=1}^{n} \int (Y_i^{(1-W_i)} - \mathbb{E}_\theta[f_{1-W_i} \,|\, \mathcal{D}])^2\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \int (Y_i^{(1-W_i)} - \mathbb{E}_\theta[Y_i^{(1-W_i)} \,|\, \mathcal{D}])^2\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}[\, Y_i^{(1-W_i)} \,|\, \mathcal{D}\,], \\ &= \frac{1}{n} \|\mathrm{Var}[\, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})} \,|\, \mathcal{D}\,]\|_1, \end{aligned} \qquad \text{(F.37)}$$

$$(\hat{f}^*, \theta^*) = \arg\min_{\hat{f},\theta} \int \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\mathbf{f}}^T(X_i)\mathbf{e} - (1-2W_i)\left(Y_i^{(1-W_i)} - Y_i^{(W_i)}\right)\right)^2 d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}). \tag{F.32}$$

$$\theta^* = \arg\min_{\theta} \int \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_\theta[\mathbf{f}^T(X_i)\,|\,\mathcal{D}]\mathbf{e} - (1-2W_i)\left(Y_i^{(1-W_i)} - Y_i^{(W_i)}\right)\right)^2 d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}). \tag{F.33}$$

$$R = \frac{1}{n}\sum_{i=1}^{n} \int \left((1-2W_i)\left((Y_i^{(W_i)} - \mathbb{E}_\theta[f_{W_i}(X_i)\,|\,\mathcal{D}]) - (Y_i^{(1-W_i)} - \mathbb{E}_\theta[f_{1-W_i}\,|\,\mathcal{D}])\right)\right)^2 d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}). \tag{F.34}$$

$$R = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\int (Y_i^{(W_i)} - \mathbb{E}_\theta[f_{W_i}(X_i)\,|\,\mathcal{D}])^2\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D})}_{R_1} + \underbrace{\int (Y_i^{(1-W_i)} - \mathbb{E}_\theta[f_{1-W_i}\,|\,\mathcal{D}])^2\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D})}_{R_2}$$
$$\underbrace{- 2\int (Y_i^{(W_i)} - \mathbb{E}_\theta[f_{W_i}(X_i)\,|\,\mathcal{D}])\,(Y_i^{(1-W_i)} - \mathbb{E}_\theta[f_{1-W_i}\,|\,\mathcal{D}])\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D})}_{R_3}. \tag{F.35}$$

and

$$R_3 = \frac{1}{n}\sum_{i=1}^{n} \int (Y_i^{(W_i)} - \mathbb{E}_\theta[f_{W_i}\,|\,\mathcal{D}]) \\ (Y_i^{(1-W_i)} - \mathbb{E}_\theta[f_{1-W_i}\,|\,\mathcal{D}])\, d\mathbb{P}_\theta(Y_i^{(1-W)}|\mathcal{D}) = 0$$

Therefore, θ^* is found by minimizing $\|\mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\mathbf{f}\,|\,\mathcal{D}]\|_2^2 + \|\mathrm{Var}[\,\mathbf{Y}^{(\mathbf{1-W})}\,|\,\mathcal{D}\,]\|_1$.

REFERENCES

[1] J. C. Foster, J. M. Taylor, and S. J. Ruberg, "Subgroup identification from randomized clinical trial data," *Statistics in medicine*, vol. 30, no. 24, pp. 2867–2880, 2011.

[2] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3207–3260, 2013.

[3] Y. Xie, J. E. Brand, and B. Jann, "Estimating heterogeneous treatment effects with observational data," *Sociological methodology*, vol. 42, no. 1, pp. 314–347, 2012.

[4] W. Sauerbrei, M. Abrahamowicz, D. G. Altman, S. Cessie, and J. Carpenter, "Strengthening analytical thinking for observational studies: The stratos initiative," *Statistics in medicine*, vol. 33, no. 30, pp. 5413–5432, 2014.

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

[6] D. Tran, F. J. Ruiz, S. Athey, and D. M. Blei, "Model criticism for bayesian causal inference," *arXiv preprint arXiv:1610.09037*, 2016.

[7] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, no. just-accepted, 2017.

[8] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[9] S. Li and Y. Fu, "Matching on balanced nonlinear representations for treatment effects estimation," in *Advances in Neural Information Processing Systems*, 2017, pp. 930–940.

[10] N. Kallus, "Recursive partitioning for personalization using observational data," in *International Conference on Machine Learning*, 2017, pp. 1789–1798.

[11] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[12] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[13] A. Abadie and G. W. Imbens, "Matching on the estimated propensity score," *Econometrica*, vol. 84, no. 2, pp. 781–807, 2016.

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

[15] A. Alaa, M. Weisz, and M. van der Schaar, "Deep counterfactual networks with propensity-dropout," *ICML Workshop on Principled Approaches to Deep learning*, 2017.

[16] P. R. Hahn, J. S. Murray, and C. M. Carvalho, "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects," 2017.

[17] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen *et al.*, "Double machine learning for treatment and causal parameters," *American Economic Review (AER)*, vol. 107, no. 5, pp. 261–265, 2017.

[18] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International Conference on Machine Learning*, 2016, pp. 3020–3029.

[19] S. Powers, J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani, "Some methods for heterogeneous treatment effect estimation in high-dimensions," *arXiv preprint arXiv:1707.00102*, 2017.

[20] Y. Xie, "Population heterogeneity and causal inference," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6262–6268, 2013.

[21] M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran, "Estimating individual treatment effect in observational data using random forest methods," *Journal of Computational and Graphical Statistics*, 2017.

[22] V. Rockova and S. van der Pas, "Posterior concentration for bayesian regression trees and their ensembles," *arXiv preprint arXiv:1708.08734*, 2017.

[23] A. R. Linero and Y. Yang, "Bayesian regression tree ensembles that adapt to smoothness and sparsity," *arXiv preprint arXiv:1707.09461*, 2017.

[24] S. Sniekers, A. van der Vaart *et al.*, "Adaptive bayesian credible sets in regression with a gaussian process prior," *Electronic Journal of Statistics*, vol. 9, no. 2, pp. 2475–2527, 2015.

[25] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *The annals of statistics*, pp. 1040–1053, 1982.

[26] ——, "Optimal rates of convergence for nonparametric estimators," *The annals of Statistics*, pp. 1348–1360, 1980.

[27] A. v. d. Vaart and H. v. Zanten, "Information rates of nonparametric gaussian process methods," *Journal of Machine Learning Research*, vol. 12, no. Jun, pp. 2095–2119, 2011.

[28] U. Shalit, F. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," pp. 3076–3085, 2017.

[29] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 1998, vol. 3.

[30] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, pp. 1564–1599, 1999.

[31] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[32] M. W. Seeger, S. M. Kakade, and D. P. Foster, "Information consistency of nonparametric gaussian process methods," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2376–2382, 2008.

[33] J. Bernardo, J. Burger, and A. D. E. M. Smith, "Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems," 1998.
[34] Y. Chen, X. Yi, and C. Caramanis, "Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates," *IEEE Transactions on Information Theory*, 2017.
[35] Y. Yang, S. T. Tokdar *et al.*, "Minimax-optimal nonparametric regression in high dimensions," *The Annals of Statistics*, vol. 43, no. 2, pp. 652–674, 2015.
[36] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.
[37] G. Raskutti, B. Yu, and M. J. Wainwright, "Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness," in *Advances in Neural Information Processing Systems*, 2009, pp. 1563–1570.
[38] S. Efromovich, "Density estimation for biased data," *Annals of Statistics*, pp. 1137–1161, 2004.
[39] C. O. Wu and A. Q. Mao, "Minimax kernels for density estimation with biased data," *Annals of the Institute of Statistical Mathematics*, vol. 48, no. 3, pp. 451–467, 1996.
[40] M. I. Borrajo, W. González-Manteiga, and M. D. Martínez-Miranda, "Bandwidth selection for kernel density estimation with length-biased data," *Journal of Nonparametric Statistics*, vol. 29, no. 3, pp. 636–668, 2017.
[41] E. Brunel, F. Comte, and A. Guilloux, "Nonparametric density estimation in presence of bias and censoring," *Test*, vol. 18, no. 1, pp. 166–194, 2009.
[42] C. O. Wu, "A cross-validation bandwidth choice for kernel density estimates with selection biased data," *Journal of multivariate analysis*, vol. 61, no. 1, pp. 38–60, 1997.
[43] M. Hoffmann, J. Rousseau, J. Schmidt-Hieber *et al.*, "On adaptive posterior concentration rates," *The Annals of Statistics*, vol. 43, no. 5, pp. 2259–2295, 2015.
[44] B. T. Knapik, A. W. van der Vaart, J. H. van Zanten *et al.*, "Bayesian inverse problems with gaussian priors," *The Annals of Statistics*, vol. 39, no. 5, pp. 2626–2657, 2011.
[45] A. W. van der Vaart, J. H. van Zanten *et al.*, "Reproducing kernel hilbert spaces of gaussian priors," in *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*. Institute of Mathematical Statistics, 2008, pp. 200–222.
[46] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006, vol. 1.
[47] M. A. Alvarez, L. Rosasco, N. D. Lawrence *et al.*, "Kernels for vector-valued functions: A review," *Foundations and Trends® in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012.
[48] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Advances in neural information processing systems*, 2008, pp. 153–160.
[49] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," in *Computational learning theory*. Springer, 2001, pp. 416–426.
[50] K. E. Porter, S. Gruber, M. J. Van Der Laan, and J. S. Sekhon, "The relative performance of targeted maximum likelihood estimators," *The International Journal of Biostatistics*, vol. 7, no. 1, pp. 1–34, 2011.
[51] GPy, "GPy: A gaussian process framework in python," http://github.com/SheffieldML/GPy, since 2012.
[52] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, 1990.
[53] I. Castillo, "Lower bounds for posterior rates with gaussian process priors," *Electron. J. Stat.*, vol. 2, pp. 1281–1299, 2008.
[54] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 714–722, 1995.