# Open Problem: Structure-Agnostic Minimax Risk for Partial Linear Model

**Yihong Gu** YIHONG_GU@HMS.HARVARD.EDU
*Department of Biomedical Informatics, Harvard University*

**Editors:** Nika Haghtalab and Ankur Moitra

## Abstract

Double machine learning is a theoretically grounded and practically efficient procedure for a variety of causal estimands and functional estimation problems when adopting black-box machine learning models for estimating nuisance parameters. It is known that double machine learning may have sub-optimal performance in the structure-aware settings, e.g., the nuisances are Hölder smooth functions, and recent articles (Balakrishnan et al., 2023) are delivering the message that double machine learning is optimal in structure-agnostic settings. This note claims that whether double machine learning is optimal for black-box machine learning models remains open, even for the simplest linear coefficient estimation in the partial linear model. We argue that the key gap that differentiates structure-agnostic and structure-aware settings, and also the previous lower bound results do not address, is the role of variance – the awareness of well-conditioned structures offers the possibility to mitigate the effects of variance, while that is not clear for structure-agnostic settings. The answer to this question has significant implications both in theory and practice.

**Keywords:** Double machine learning; Minimax optimality; Partial linear model; Structure-agnostic estimation

## 1. Introduction

Suppose we observe $(X, T, Y) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$ from the model

$$\begin{aligned} Y &= \beta_0 \cdot T + \mu_0(X) + \varepsilon_Y \qquad \text{with} \qquad \mathbb{E}[\varepsilon_Y | X, T] = 0, \\ T &= \pi_0(X) + \varepsilon_T \qquad\quad\ \ \text{with} \qquad \mathbb{E}[\varepsilon_T | X] = 0. \end{aligned} \tag{1}$$

Here we call $Y$ the *outcome*, $T$ the *treatment*, $X$ the *covariate*. $\mu_0$ and $\pi_0$ are usually referred to as *outcome function* and *propensity score*, respective. The goal is to estimate the homogeneous treatment effect $\theta_0 = \beta_0 \in \mathbb{R}$ using $2n$ i.i.d. observations $\mathcal{D} = \{(X_i, T_i, Y_i)\}_{i=1}^{2n}$ from (1). We consider the setting that best fits the modern machine learning pipeline – the *structure-agnostic* setting. To be specific, we are unaware of any structure of $\mu_0$ and $\pi_0$ like additivity or (sparse) linearity, but know $\mu_0$ and $\pi_0$ can be estimated well by some black-box machine learning models $\mathcal{G}_\mu$ and $\mathcal{G}_\pi$, represented as the function classes, with certain approximation error (or misspecification error) and stochastic error (or variance). We are interested in the best structure-agnostic algorithm on top of black-box machine learning models under the worst case.

We first define the notation of approximation error and stochastic error. With loss of generality, we assume $X$ is uniformly distributed on $[0, 1]^d$ and all the functions are uniformly bounded by 3. Let $\|h\|_2 = \sqrt{\mathbb{E}[h^2(X)]}$, for a machine learning model $\mathcal{G}_h$ to estimate the function $h$, we can define its approximation error and stochastic error as

$$\delta_h^{\mathrm{appr}} = \inf_{g \in \mathcal{G}_h} \|h - g\|_2 \qquad \text{and} \qquad \delta_h^{\mathrm{stoc}} = \inf\left\{\bar{\delta} > 0 : \mathsf{R}(\delta; \partial\mathcal{G}_h) \le \delta \cdot \bar{\delta} \ \ \forall \delta \ge \bar{\delta}\right\}$$

where $\mathsf{R}(\delta;\mathcal{F}) = \mathbb{E}_{\varepsilon_{1:n}\sim[\mathrm{Unif}\{-1,+1\}]^n}[\sup_{f\in\mathcal{F},\|f\|_2\le\delta}\frac{1}{n}\sum_{i=1}^n f(X_i)\varepsilon_i]$ is the local Rademacher complexity. It follows from standard empirical process theory that one can estimate $h$ with error rate $\delta_h^{\mathrm{appr}}+\delta_h^{\mathrm{stoc}}$ using least squares. For example, one has $\|\hat\pi-\pi_0\|_2 \lesssim \delta_\pi^{\mathrm{appr}}+\delta_\pi^{\mathrm{stoc}}$ for $\hat\pi = \mathrm{argmin}_{\pi\in\mathcal{G}_\pi}\frac{1}{n}\sum_{i=1}^n\{Y_i-\pi(X_i)\}^2$.

A purely structure-agnostic approach is to split the dataset $\mathcal{D}$ with size $2n$ into two subsets $\mathcal{D}_1\cup\mathcal{D}_2$ with size $n$, use the $\mathcal{D}_1$ to estimate $\pi_0$ and $\mu_0$ via least squares above and joint least squares $\hat\beta,\hat\mu = \mathrm{argmin}_{\beta\in\mathbb{R},g(x)\in\mathcal{G}_\mu}\frac{1}{n}\sum_{i=1}^n\{Y_i-\beta\cdot T_i-g(X_i)\}^2$, respectively, and then use the estimated $(\hat\mu,\hat\pi)$ and $\mathcal{D}_2$ to construct the AIPW-style estimator as

$$\hat\theta = \frac{\sum_{(X,Y,T)\in\mathcal{D}_2}(T-\hat\pi(X))(Y-\hat\mu(X))}{\sum_{(X,Y,T)\in\mathcal{D}_2}T^2-T\cdot\hat\pi(X)}. \tag{2}$$

The estimator (2) is called a double machine learning (DML) estimator. One can show that

$$|\hat\theta-\theta_0| \lesssim (\delta_\mu^{\mathrm{appr}}+\delta_\mu^{\mathrm{stoc}})\cdot(\delta_\pi^{\mathrm{appr}}+\delta_\pi^{\mathrm{stoc}})+\frac{1}{\sqrt{n}}. \tag{3}$$

We wish to answer the following question, whose answer is still open as illustrated below.

**Question 1** *Is the rate (3) the best we can obtain under the worst-case structure-agnostic scenario?*

### 1.1. Previous results

Such a model has been widely studied in various structure-aware settings, and a faster rate can be obtained if the black-box machine learning model can be represented as a linear class or a sparse combination of well-conditioned linear bases.

**Linear basis.** The most simple machine learning model is the linear model by a set of fixed basis, that is $\mathcal{G}_h^N = \{f(x)=\sum_{\ell=1}^N\alpha_\ell\phi_\ell(x):(\alpha_1,\ldots,\alpha_N)\in\mathbb{R}^N\}$ with a set of fixed basis $\{\phi_\ell(x)\}_{\ell=1}^N$. The stochastic error for this machine learning model is $(N/n)^{1/2}$. In this case, we simply let $\hat\theta=\hat\beta$ be the parameter estimate in the above joint least squares with minimization function class being $\mathcal{G}_\pi+\mathcal{G}_\mu$ instead of $\mathcal{G}_\mu$. Following the proofs in Donald and Newey (1994) or Section 7.3.3 of Fan et al. (2020) gives the error rate explicitly stated in approximation and stochastic errors:

$$|\hat\theta-\theta_0| \lesssim \delta_\mu^{\mathrm{appr}}\cdot\delta_\pi^{\mathrm{appr}}+\frac{1}{\sqrt{n}}\left[\delta_\mu^{\mathrm{stoc}}+\delta_\pi^{\mathrm{stoc}}+1\right] \qquad \text{if} \quad \delta_\mu^{\mathrm{stoc}}+\delta_\pi^{\mathrm{stoc}}=o(1). \tag{4}$$

The idea of using a set of fixed bases has been adopted to derive minimax optimal estimation when the nuisance parameters are Hölder functions.

**Well-conditioned sparse basis.** Another machine learning model is a sparse combination of linear basis, that is, $\mathcal{G}^{s,N} = \{f(x)=\sum_{\ell=1}^N\alpha_\ell\phi_\ell(x):\|\alpha\|_0\le s\}$ with stochastic error $\{s\log(N)/n\}^{1/2}$. We assume $\mathcal{G}_\pi$ and $\mathcal{G}_\mu$ use the same set of basis and this set of basis is well-conditioned, i.e., $\lambda_{\max}(\mathbb{E}[\phi(X)\phi(X)^\top])+\lambda_{\min}^{-1}(\mathbb{E}[\phi(X)\phi(X)^\top])=O(1)$ for $\phi(X)=[\phi_1(X),\ldots,\phi_N(X)]$. In this case, the debiased Lasso estimator (Zhang and Zhang, 2014) attains the error rate

$$|\hat\beta_{\mathsf{db}}-\theta_0| \lesssim \delta_\mu^{\mathrm{appr}}\cdot\delta_\pi^{\mathrm{appr}}+\left[\delta_\mu^{\mathrm{stoc}}\right]^2+\frac{1}{\sqrt{n}}. \tag{5}$$

And $\hat{\beta}_{\mathsf{db}}$ admits the same form of (2) with $\hat{\mu}$ and $\hat{\pi}$ being $L_1$ penalized least squares solutions instead of the vanilla (joint) least squares before. At the same time, when $s_\mu > s_\pi$, another estimator can obtain the rate $\delta_\mu^{\mathrm{appr}} \cdot \delta_\pi^{\mathrm{appr}} + \left[\delta_\pi^{\mathrm{stoc}}\right]^2$ (Bellec and Zhang, 2022). In summary, the full picture is that if $\mathcal{G}_\mu$ and $\mathcal{G}_\pi$ can be represented as a sparse combination of well-conditioned linear bases, then with the knowledge of all the error quantities, one can obtain

$$|\hat{\beta}_{\mathsf{d}} - \beta_0| \lesssim \delta_\mu^{\mathrm{appr}} \cdot \delta_\pi^{\mathrm{appr}} + \left[\delta_\mu^{\mathrm{stoc}} \wedge \delta_\pi^{\mathrm{stoc}}\right]^2 + \frac{1}{\sqrt{n}}. \tag{6}$$

The lower bound of stochastic error is tight by the lower bound results (Bellec and Zhang, 2022).

**Summary.** Here, we can see that for different categories of machine learning models we adopted, the cost of the product of the approximation error or the intrinsic bias, i.e., $\delta_\mu^{\mathrm{appr}} \cdot \delta_\pi^{\mathrm{appr}}$ are inevitable in the upper bounds. The problem formulation and main result in Balakrishnan et al. (2023); Jin and Syrgkanis (2024) obtain the lower bound in a similar spirit to $\delta_\mu^{\mathrm{appr}} \cdot \delta_\pi^{\mathrm{appr}}$. However, this is the one appearing invariantly in different cases. In contrast, the role of stochastic error, or variance, is different in the three rates (4), (6), and (3). It varies from the most mild (4) that only requires $o(1)$, to contributing to the final error in square (6), and finally to contributing to the final error the same as the approximation error in (3).

### 1.2. Gap and significance

For the problem of estimating the linear coefficient in a partial linear model, the double machine learning procedure can get the error rate (3) in the structure-agnostic setting, while a faster rate (6), with a matching lower bound, can be obtained in a structure-aware setting. It is still unclear whether the best rate is $(\delta_\mu^{\mathrm{appr}} + \delta_\mu^{\mathrm{stoc}}) \cdot (\delta_\pi^{\mathrm{appr}} + \delta_\pi^{\mathrm{stoc}})$ (pessimistic world), $\delta_\mu^{\mathrm{appr}} \cdot \delta_\pi^{\mathrm{appr}} + \left[\delta_\mu^{\mathrm{stoc}} \wedge \delta_\pi^{\mathrm{stoc}}\right]^2$ (optimistic world), or something in between these two under the worst-case structure-agnostic scenario. In this regard, the insights Balakrishnan et al. (2023) offer are very few, given that the product of the approximation error is not the term that differentiates the structure-agnostic and structure-aware setting. The potential answer will have various implications as illustrated below.

(a) If the answer is [pessimistic], then it confirms the optimality of the double machine learning procedure in the structure-agnostic setting. A follow-up question is: What is the reason behind such a gap between (3) and (6)? Is the reason mild in that one can fill such a gap under a mild condition that doesn't hurt the generalization of the problem, or is the reason fundamentally related to the benefits of a generic machine learning model compared with a sparse pursuit of a fixed basis?

(b) If the answer is [optimistic], then either a finer analysis of the double machine learning procedure or a better algorithm should be proposed. The latter is more likely conditioned on this answer, and this implies that double machine learning procedures are not the best algorithm on top of black-box machine learning models.

(c) If the answer lies in between the two, then it enjoys consequences in both (a) and (b).

(d) Given that the partial linear model is a canonical example of the problem that double machine learning applies, it is likely that similar answers hold for many other causal estimand and functional estimation problems.

(e) For the practitioners, the message that the potential answer will indicate whether one can benefit from reducing bias (approximation error) at the cost of large variance (stochastic error) in estimating the nuisance functions, where the latter can be reduced automatically or using debias techniques for generic black-box machine learning models.

## 2. A formal statement

In this section, we provide a formal statement for the structure-agnostic minimax risk, which is a framework that disentangles the effects of approximation error and stochastic error as compared to Balakrishnan et al. (2023). Let $X \sim \nu_d$ be uniformly distributed on $[0,1]^d$, $\{X_i\}_{i=1}^n \sim \nu_d$ be i.i.d. random variables, and $\{\varepsilon_i\}_{i=1}^n$ be independent Rademacher random variables that is also independent of $\{X_i\}_{i=1}^n$, define $\mathsf{R}_n(\delta; \partial\mathcal{G})$ as the local Rademacher complexity for the function class $\partial\mathcal{G} = \mathcal{G} - \mathcal{G}$, i.e., $\mathsf{R}_n(\delta; \partial\mathcal{G}) := \mathbb{E}[\sup_{g,\tilde{g}\in\mathcal{G}, \|g-\tilde{g}\|_2 \le \delta} \frac{1}{n}\sum_{i=1}^n (g(X_i) - \tilde{g}(X_i)) \cdot \varepsilon_i]$.

Denote $\delta = (\delta^{\text{appr}}, \delta^{\text{stoc}}) \in [0,3] \times [0,3]$. We define the set of function classes $\mathcal{H}(\delta)$ as all the function classes $\mathcal{F} \subseteq \mathcal{T} := \{f \in L_2(\nu_d) : \|f\|_\infty \le 3\}$ that can be estimated by some black-box machine learning model $\mathcal{G}$ with approximation error $\delta^{\text{appr}}$ and stochastic error $\delta^{\text{stoc}}$:

$$\mathcal{H}_n(\delta) := \Big\{\mathcal{F} \subseteq \mathcal{T} : \exists \mathcal{G} \subseteq \mathcal{T} \; s.t. \; \sup_{f\in\mathcal{F}} \inf_{g\in\mathcal{G}} \|f-g\|_2 \le \delta^{\text{appr}} \text{ and } \mathsf{R}_n(\delta; \partial\mathcal{G}) \le \delta \cdot \delta^{\text{stoc}}, \;\; \forall \delta \ge \delta^{\text{stoc}}\Big\}.$$

Given function class $\mathcal{F}_{\mathsf{PLM}} = (\mathcal{F}_\pi, \mathcal{F}_\mu)$, we define the distribution family as

$$\begin{aligned}\mathcal{P}(\mathcal{F}_{\mathsf{PLM}}) = \Big\{&(X,T,Y) \sim \mathbb{P}_{\mu,\beta,\pi} \text{ where } \pi \in \mathcal{F}_\pi, \mu \in \mathcal{F}_\mu, |\beta| \le 3, |T| \vee |Y| \le 3 : \\ &X \sim \nu_d, \;\; T = \pi(X) + \varepsilon_T \text{ with } \mathbb{E}[\varepsilon_T|X] = 0, \mathbb{E}[\varepsilon_T^2|X] \ge 1/3 \\ &Y = T\cdot\beta + \mu(X) + \varepsilon_Y \text{ with } \mathbb{E}[\varepsilon_Y|X,T] = 0\Big\},\end{aligned} \tag{7}$$

with the target estimand $\theta(\mathbb{P}) = \beta$. The structure-agnostic minimax risk is defined as

$$\mathfrak{m}(n, \delta_\pi, \delta_\mu) := \sup_{\substack{\mathcal{F}_q \in \mathcal{H}_n(\delta_q) \\ \forall q \in \{\pi,\mu\}}} \underbrace{\inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{F}_\pi,\mathcal{F}_\mu)}}_{\text{standard minimax risk over given class } (\mathcal{F}_\mu, \mathcal{F}_\pi)} \mathbb{E}_{\mathbb{P}^{2n}}\left[\left|\hat{\theta} - \theta(\mathbb{P})\right|^2\right], \tag{8}$$

where the randomness in $\mathbb{E}_{\mathbb{P}^{2n}}$ is $2n$ i.i.d. observations from $\mathbb{P}$, and $\hat{\theta}$ is the function of the $2n$ observations and potential black-box machine learning model $\mathcal{G}_\mu$ and $\mathcal{G}_\pi$. Intuitively, the structure-agnostic minimax risk is defined as the worst minimax risk over all the potential function classes with fixed approximation errors $(\delta_\mu^{\text{appr}}, \delta_\pi^{\text{appr}})$ and stochastic errors $(\delta_\mu^{\text{stoc}}, \delta_\pi^{\text{stoc}})$ budget.

The lower bound adopts the construction in Balakrishnan et al. (2023), which picks the hard function class as $L_\infty$ ball with radius $u$. The idea is that this function class itself has local Rademacher complexity $\sqrt{u}$. The attained upper bound follows from the standard proofs of double machine learning (Chernozhukov et al., 2018) with sample-splitting.

**Theorem 1** *There exists a universal constant $C > 0$ such that*

$$\begin{aligned}&C^{-1}\left\{\frac{1}{\sqrt{n}} + \left[\delta_\pi^{\text{appr}} + (\delta_\pi^{\text{stoc}})^2\right]\cdot\left[\delta_\mu^{\text{appr}} + (\delta_\mu^{\text{stoc}})^2\right]\right\} \\ &\qquad \le \; \mathfrak{m}(n, \delta_\pi, \delta_\mu) \le C\left\{\frac{1}{\sqrt{n}} + \left[\delta_\pi^{\text{appr}} + \delta_\pi^{\text{stoc}}\right]\cdot\left[\delta_\mu^{\text{appr}} + \delta_\mu^{\text{stoc}}\right]\right\}\end{aligned} \tag{9}$$

The formal statement of the open problem is: what is the matching upper bound and lower bound of $\mathfrak{m}(n, \delta_\pi, \delta_\mu)$? A potential new algorithm or finer analysis of double machine learning should be proposed for a tighter upper bound. New function classes should be introduced to improve the lower bound, because previous proposed classes like linear basis, well-conditioned sparse linear basis, and $L_\infty$ balls can only give $(\delta^{\text{stoc}})^2$ lower bound.

## References

Sivaraman Balakrishnan, Edward H Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023.

Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Stephen G Donald and Whitney K Newey. Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50(1):30–40, 1994.

Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical foundations of data science*. Chapman and Hall/CRC, 2020.

Jikai Jin and Vasilis Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. *arXiv preprint arXiv:2402.14264*, 2024.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.