Adjustment for Confounding using Pre-Trained Representations

Rickmer Schulte [1 2] **David Rügamer** [1 2] **Thomas Nagler** [1 2]

Abstract

There is growing interest in extending average treatment effect (ATE) estimation to incorporate non-tabular data, such as images and text, which may act as sources of confounding. Neglecting these effects risks biased results and flawed scientific conclusions. However, incorporating non-tabular data necessitates sophisticated feature extractors, often in combination with ideas of transfer learning. In this work, we investigate how latent features from pre-trained neural networks can be leveraged to adjust for sources of confounding. We formalize conditions under which these latent features enable valid adjustment and statistical inference in ATE estimation, demonstrating results along the example of double machine learning. In this context, we also discuss critical challenges inherent to latent feature learning and downstream parameter estimation using those. As our results are agnostic to the considered data modality, they represent an important first step towards a theoretical foundation for the usage of latent representation from pre-trained models in ATE estimation.

1. Introduction

Causal inference often involves estimating the average treatment effect (ATE), which represents the causal impact of an exposure on an outcome. Under controlled study setups of randomized controlled trials (RCTs), valid inference methods for ATE estimation are well established (Deaton & Cartwright, 2018). However, RCT data is usually scarce and in some cases even impossible to obtain, either due to ethical or economic reasons. This often implies relying on observational data, typically subject to (unmeasured) confounding—(hidden) factors that affect both the exposure and the outcome. To overcome this issue of confounding and to obtain unbiased estimates, several inferential methods have been developed to properly adjust the ATE estimation for confounders. One approach that has garnered significant attention in recent years is the debiased/double machine learning (DML) framework (Chernozhukov et al., 2017; 2018), which allows the incorporation of machine learning methods to adjust for non-linear or complex confounding effects in the ATE estimation. DML is usually applied in the context of tabular features and was introduced for ML methods tailored to such features. However, confounding information might only be present in non-tabular data, such as images or text.

Non-tabular Data as Sources of Confounding Especially in medical domains, imaging is a key component of the diagnostic process. Frequently, CT scans or X-rays are the basis to infer a diagnosis and a suitable treatment for a patient. However, as the information in such medical images often also affects the outcome of the therapy, the information in the image acts as a confounder. Similarly, treatment and health outcomes are often both related to a patient's files, which are typically in text form. Consequently, ATE estimation based on such observational data will likely be biased if the confounder is not adequately accounted for. Typical examples would be the severity of a disease or fracture. The extent of a fracture impacts the likelihood of surgical or conservative therapy, and the severity of a disease may impact the decision for palliative or chemotherapy. In both cases, the severity will likely also impact the outcome of interest, e.g., the patient's recovery rate. Another famous example is the Simpson's Paradox observed in the kidney stone treatment study of Charig et al. (1986). The size of the stone (information inferred from imaging) impacts both the treatment decision and the outcome, which leads to flawed conclusions about the effectiveness of the treatment if confounding is not accounted for (Julious & Mullee, 1994).

Contemporary Applications While the DML framework provides a solution for non-linear confounding, previous examples demonstrate that modern data applications require extending ATE estimation to incorporate non-tabular data. In contrast to traditional statistical methods and classical machine learning approaches, information in non-tabular

[1]Department of Statistics, LMU Munich, Munich, Germany [2]Munich Center for Machine Learning (MCML), Munich, Germany. Correspondence to: Rickmer Schulte <schulte@stat.uni-muenchen.de>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

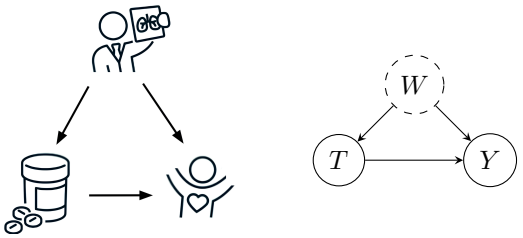

Figure 1. Schematic (left) and DAG visualization (right) of the effect of a treatment T on outcome Y that is confounded by non-tabular data W (e.g. information from medical imaging).

data usually requires additional feature extraction mechanisms to condense high-dimensional inputs to the relevant information in the data. This is usually done by employing neural network-based approaches such as foundation models or other pre-trained neural networks. While it may seem straightforward to use such feature extractors to extract latent features from non-tabular data and use the resulting information in classical DML approaches, we show that this necessitates special caution. In particular, incorporating such features into ATE estimation requires overcoming previously unaddressed theoretical and practical challenges, including non-identifiability, high dimensionality, and the resulting limitations of standard assumptions like sparsity.

Problem setup Given n independent and identically distributed (i.i.d.) observations of (T, W, Y), we are interested in estimating the ATE of a binary variable $T \in \{0, 1\}$ on some outcome of interest $Y \in \mathbb{R}$ while adjusting for some source of confounding $W \in \mathbb{W}$ (cf. Figure 1). W is pre-treatment data from some potentially complex sampling space $\mathbb{W}$ that is assumed to be *sufficient* for adjustment. The definition of sufficiency will be formalized in Section 3.1. Under *positivity* and *consistency* assumption—the standard assumptions in causality—the target parameter of interest can be identified as

$$\text{ATE} := \mathbb{E}[\mathbb{E}[Y|T = 1, W] - \mathbb{E}[Y|T = 0, W]]. \quad (1)$$

While there are many well-known ATE estimators, most require to estimate either the outcome regression function

$$g(t, w) := \mathbb{E}[Y|T = t, W = w] \quad (2)$$

or the propensity score

$$m(t|w) := \mathbb{P}[T = t|W = w] \quad (3)$$

at parametric rate $\sqrt{n}$. Doubly robust estimators such as the Augmented Inverse Probability Weighted, the Targeted Maximum Likelihood Estimation or the DML approach estimate both *nuisance functions* g and m. These methods thus only require the product of their estimation errors to converge at $\sqrt{n}$-rate (Robins & Rotnitzky, 1995; Van Der Laan & Rubin, 2006; Van der Laan & Rose, 2011; Chernozhukov et al., 2017; 2018). However, even this can be hard to achieve, given the *curse of dimensionality* when considering the high-dimensionality of non-tabular data W such as images. Especially given the often limited number of samples available in many medical studies involving images, estimating m and g as a function of W, e.g., via neural networks, might not be feasible or overfit easily. To cope with such issues, a common approach is to adopt ideas from transfer learning and use pre-trained neural networks.

Our Contributions In this paper, we discuss under what conditions *pre-trained representations* $Z := \varphi(W)$ obtained from pre-trained neural networks φ can replace W in the estimation of nuisance functions (2) and (3). Although the dimensionality of Z is usually drastically reduced compared to W, one major obstacle from a theoretical point of view is that representations can only be learned up to invertible linear transformations (e.g., rotations). We argue that common assumptions allowing fast convergence rates, e.g., *sparsity* or *additivity* of the nuisance function, are no longer reasonable in such settings. In contrast, we build on the idea of low *intrinsic dimensionality* of the pre-trained representations. Combining invariance of intrinsic dimensions and functional smoothness with structural sparsity, we establish conditions that allow for sufficiently fast convergence rates of nuisance function estimation and, thus, valid ATE estimation and inference. Our work, therefore, not only advances the theoretical understanding of causal inference in this context but also provides practical insights for integrating modern machine learning tools into ATE estimation.

2. Related Work

The DML framework was initially proposed for tabular features in combination with classical machine learning methods (Chernozhukov et al., 2017; 2018). Several theoretical and practical extensions to incorporate neural networks have been made with a focus on tabular data (Shi et al., 2019; Farrell et al., 2021; Chernozhukov et al., 2022; Zhang & Bradic, 2024). Additionally, there is a growing body of research that aims to incorporate non-tabular data as adjustment into DML (Veitch et al., 2019; 2020; Klaassen et al., 2024). While the latter directly incorporates the non-tabular data in the estimation, none of them discuss conditions that would theoretically justify fast convergence rates necessary for valid inference. A different strand of research instead uses either derived predictions (Zhang et al., 2023; Battaglia et al., 2024; Jerzak et al., 2022a;b; 2023) or proxy variables (Kuroki & Pearl, 2014; Kallus et al., 2018; Miao et al., 2018; Mastouri et al., 2021) in downstream estimation. In contrast to these proposals, we consider the particularly broad setup of using pre-trained representations for confounding adjustment. Given the increasing popularity of pre-trained

models, Dai et al. (2022) and Christgau & Hansen (2024) establish theoretical conditions justifying the use of derived representations in downstream tasks, which we will review in the next section. The idea of a low intrinsic dimensionality of non-tabular data and its latent representations to explain the superior performance of deep neural networks in non-tabular data domains has been explored and validated both empirically (Gong et al., 2019; Ansuini et al., 2019; Pope et al., 2021; Konz & Mazurowski, 2024) and theoretically (Chen et al., 2019; Schmidt-Hieber, 2019; Nakada & Imaizumi, 2020). By connecting several of those theoretical ideas and empirical findings, our work establishes a set of novel theoretical results and conditions that allow to obtain valid inference when using pre-trained representations in adjustment for confounding.

3. Properties of Pre-Trained Representations

Given the high dimensional nature of non-tabular data, together with the often limited number of samples available (especially in medical domains), training feature extractors such as deep neural networks from scratch is often infeasible. This makes the use of latent features from pre-trained neural networks a popular alternative (Erhan et al., 2010). In order to use pre-trained representations for adjustment in the considered ATE setup, certain conditions regarding the representations are required.

3.1. Sufficient Conditions on Pre-Trained Representations

Given any pre-trained model φ, trained independently of W on another dataset, we denote the learned (last-layer) representations as $Z \coloneqq \varphi(W)$. Due to the non-identifiability of Z up to certain orthogonal transformations, further discussed in Section 3.2, we define the following conditions for the induced equivalence class of representations $\mathcal{Z}$ following Christgau & Hansen (2024). For this, we abstract the adjustment as conditioning on information in the ATE estimation, namely conditioning on the uniquely identifiable information contained in the sigma-algebra $\sigma(Z)$ generated by any $Z \in \mathcal{Z}$ (see also Appendix A.1 for a special case).

Definition 3.1. [Christgau & Hansen (2024)] Given the joint distribution P of (T, W, Y), sigma algebra $\sigma(Z)$ of Z, and $t \in \{0, 1\}$, we say that any $Z \in \mathcal{Z}$ is

(i) P**-valid** if (P-$a.s.$):

$$\mathbb{E}_P[\mathbb{E}_P[Y|T = t, \sigma(Z)]] = \mathbb{E}_P[\mathbb{E}_P[Y|T = t, W]]$$

(ii) P**-OMS** (Outcome Mean Sufficient) if (P-$a.s.$):

$$\mathbb{E}_P[Y|T = t, \sigma(Z)] = \mathbb{E}_P[Y|T = t, W]$$

(iii) P**-ODS** (Outcome Distribution Sufficient) if:

$$Y \perp_P W|T, Z.$$

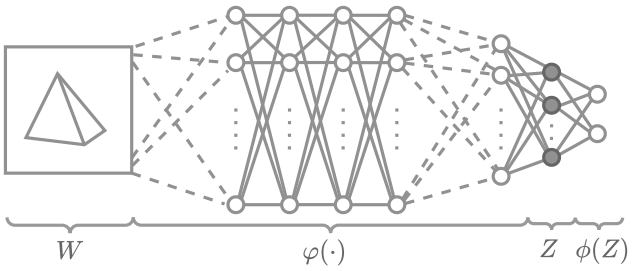

Figure 2. Schematic visualization of a pre-trained neural network $\varphi(\cdot)$ and representations $Z = \varphi(W)$.

Remark 3.2. If $Z \in \mathcal{Z}$ is P-ODS, it is also called a *sufficient embedding* in the literature (Dai et al., 2022).

The three conditions in Definition 3.1 place different restrictions on the nuisance functions (2) and (3). While P-ODS is most restrictive (followed by P-OMS) and thus guarantees valid downstream inference more generally, the strictly weaker condition of P-validity is already sufficient (and in fact necessary) to guarantee that $Z \in \mathcal{Z}$ is a *valid adjustment set* in the ATE estimation (Christgau & Hansen, 2024). Thus, any pre-trained representation Z considered in the following is assumed to be at least P-valid.

3.2. Non-Identifiability under ILTs

In practice, the representation $Z = \varphi(W)$ is extracted from some layer of a pre-trained neural network φ. This information does not change under bijective transformations of Z, so the representation Z itself is not identifiable. We argue that, in this context, non-identifiability with respect to invertible linear transformations (ILTs) is most important. Suppose $Z = \varphi(W)$ is extracted from a deep network's ℓth layer. During pre-training the network further processes Z through a model head $\phi(Z)$, as schematically depicted in Figure 2. The model head usually has the form $\phi^{>\ell}(AZ+b)$ where A, b are the weights and biases of the ℓth layer, and $\phi^{>\ell}$ summarizes all following computations. Due to this structure, any bijective linear transformation $Z \mapsto QZ$ can be reversed by the weights $A \mapsto \tilde{A} = Q^{-1}A$ so that the networks $\phi^{>\ell}(A \cdot +b)$ and $\phi^{>\ell}(\tilde{A}Q \cdot +b)$ have the same output.

Definition 3.3 (Invariance to ILTs)**.** Given a latent representation Z, we say that a model (head) ϕ_ξ with parameters $\xi \in \Xi$ is non-identifiable up to invertible linear transformations if for any invertible matrix $Q \in \mathbb{R}^{d\times d}$ $\exists \tilde{\xi} \in \Xi : \phi_\xi(QZ) = \phi_{\tilde{\xi}}(Z)$.

Important examples of ILTs are rotations, permutations, and scalings of the feature space as well as compositions thereof.

Smoothness	+ Additivity	+ Sparsity & Linearity	Intrinsic Dimension
Stone (1982)	Stone (1985)	Raskutti et al. (2009)	Bickel & Li (2007)
$O(n^{-\frac{s}{2s+d}})$	$O(n^{-\frac{s}{2s+1}})$	$O(\sqrt{p\log(d)/n})$, $p \ll d$	$O(n^{-\frac{s}{2s+d_{\mathcal{M}}}})$, $d_{\mathcal{M}} \ll d$

Table 1. Assumptions and related minimax convergence rates of the estimation error

4. Estimation using Pre-Trained Representations

The previous section discussed sufficient and necessary (information theoretic) conditions for pre-trained representations, justifying their usage for adjustment in downstream tasks. The following section will discuss aspects of the functional estimation in such adjustments. Valid statistical inference in downstream tasks usually requires fast convergence of nuisance function estimators. However, obtaining fast convergence rates in high-dimensional estimation problems is particularly difficult. We argue that some commonly made assumptions are unreasonable due to the non-identifiability of representations. We discuss this in the general setting of nonparametric estimation as described in the following.

The Curse of Dimensionality The general problem in nonparametric regression is to estimate some function f in the regression model

$$Y = f(X) + \epsilon \tag{4}$$

with outcome $Y \in \mathbb{R}$, features $X \in \mathbb{R}^d$, and error $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The minimax rate for estimating Lipschitz functions is known to be $n^{-\frac{1}{2+d}}$ (Stone, 1982). This rate becomes very slow for increasing d, known as the *curse of dimensionality*. Several additional structural and distributional assumptions are commonly encountered to obtain faster convergence rates in high dimensions.

4.1. Structural Assumption I: Smoothness

A common structural assumption is the smoothness of the function f in (4), i.e., the existence of s bounded and continuous derivatives. Most convergence rate results assume at least some level of smoothness (see Table 1). The following lemma verifies that this condition is also preserved under ILTs.

Lemma 4.1 (Smoothness Invariance under ILTs)**.** *Let $D \subseteq \mathbb{R}^d$ be an open set, $f : D \to \mathbb{R}$ be an s-smooth-function on D, and Q by any ILT. Then $h = f \circ Q^{-1} \colon Q(D) \to \mathbb{R}$ is also s-smooth on the transformed domain $Q(D)$.*

The proof of Lemma 4.1 and subsequent lemmas of this section are given in Appendix A.3.

The lemma shows that a certain level of smoothness of a function defined on latent representations may reasonably be assumed due to its invariance to ILTs. If the feature dimension is large, however, an unrealistic amount of smoothness would be required to guarantee fast convergence rates (e.g., of order $n^{-1/4}$). This necessitates additional structural or distributional assumptions.

4.2. Structural Assumptions II: Additivity & Sparsity

The common structural assumption is that f is *additive*, $f(x) = \sum_{j=1}^{d} f_j(x_j)$, i.e., the sum of univariate s-smooth functions. In this case, the minimax convergence rate reduces to $n^{-\frac{s}{2s+1}}$ (Stone, 1985). Another common approach is to rely on the idea of *sparsity*. Assuming that f is p-sparse implies that it only depends on $p < \min(n, d)$ features. In case one further assumes the univariate functions to be linear in each feature, i.e. $f(x) = \sum_{j=1}^{p} \beta_j x_j$ with coefficient $\beta_j \in \mathbb{R}$, the optimal convergence rate reduces to $\sqrt{p\log(d/p)/n}$ (Raskutti et al., 2009).

It can easily be shown that the previously discussed conditions are both preserved under permutation and scaling. But as the following lemma shows, sparsity and additivity of f are (almost surely) not preserved under generic ILTs such as rotations.

Lemma 4.2 (Non-Invariance of Additivity and Sparsity under ILTs)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of $x \in \mathbb{R}^d$. We distinguish between two cases:*

(i) ***Additive****: $f(x) = \sum_{j=1}^{d} f_j(x_j)$, with univariate functions $f_j : \mathbb{R} \to \mathbb{R}$, and at least one f_j being non-linear.*

(ii) ***Sparse Linear****: $f(x) = \sum_{j=1}^{d} \beta_j x_j$, where $\beta_j \in \mathbb{R}$ and at least one (but not all) $\beta_j = 0$.*

Then, for almost every Q drawn from the Haar measure on the set of ILTs, it holds:

(i) If f is additive, then $h = f \circ Q^{-1}$ is not additive.

(ii) If f is sparse linear, then $h = f \circ Q^{-1}$ is not sparse.

Given the non-identifiability of representations with respect to ILTs and the non-invariance result of Lemma 4.2, any additivity or sparsity assumption about the target function f of the latent features seems unjustified. An example of this rotational non-invariance of sparsity is given in Figure 3. This also implies that learners such as the Lasso (with

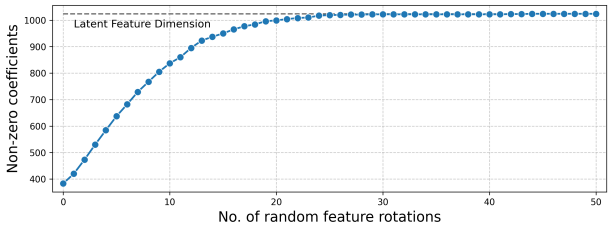

Figure 3. Non-zero coefficients of a linear classifier on latent features, showing that sparsity is lost with an increasing number of random feature rotations.

underlying sparsity assumption), tree-based methods that are based on axis-aligned splits (including corresponding boosting methods), and most feature selection algorithms are not ILT-invariant. Further examples can be found in (Ng, 2004).

4.3. Distributional Assumption: Intrinsic Dimension

While the previous conditions are structural assumptions regarding the function f itself, faster convergence rates can also be achieved by making distribution assumptions about the support of f. A popular belief is that the d-dimensional data $X \in \mathbb{R}^d$ lie on or close to a low-dimensional manifold $\mathcal{M}$ with intrinsic dimension $d_{\mathcal{M}}$. This relates to the famous *manifold hypothesis* that many high-dimensional data concentrate on low-dimensional manifolds (Fefferman et al., 2016, e.g.,). There is strong empirical support for this assumption, especially for non-tabular modalities such as text and images, see Appendix B.1. Given that $d_{\mathcal{M}} \ll d$, and again assuming f to be s-smooth, this can lead to a much faster convergence rate of $n^{-\frac{s}{2s+d_{\mathcal{M}}}}$ (Bickel & Li, 2007), as it is independent of the dimension d of the ambient space.

Similarly to Lemma 4.1, the following lemma shows the invariance of the intrinsic dimension of a manifold with respect to any ILT of the coordinates in the d-dimensional ambient space.

Lemma 4.3 (Intrinsic Dimension Invariance under ILTs)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a smooth manifold of dimension $d_{\mathcal{M}} \leq d$. For any ILT Q, the transformed set*

$$Q(\mathcal{M}) \;=\; \{\, Qx \;\mid\; x \in \mathcal{M} \}\,.$$

is also a smooth manifold of dimension $d_{\mathcal{M}}$.

Remark 4.4. Put differently, in case the latent representations $Z \in \mathbb{R}^d$ lie on a $d_{\mathcal{M}}$-dimensional smooth manifold $\mathcal{M}$, then the IL-transformed representations $Q(Z)$ also lie on a smooth manifold $Q(\mathcal{M})$ of dimension $d_{\mathcal{M}}$.

Summarizing previous results, the structural and distribution assumptions of smoothness and low intrinsic dimensionality are invariant with respect to any ILT of the features. Hence, as opposed to additivity or sparsity, the two conditions hold not only for a particular instantiation of a latent representation Z but for the entire equivalence class of latent representations induced by the class of ILTs. This is crucial given the non-identifiability of latent representations, highlighting the importance of low intrinsic dimensions (IDs).

Deep Networks Can Adapt to Intrinsic Dimensions Recently, several theoretical works have shown that DNNs can adapt to the low intrinsic dimension of the data and thereby attain the optimal rate of $n^{-\frac{s}{2s+d_{\mathcal{M}}}}$ (Chen et al., 2019; Schmidt-Hieber, 2019; Nakada & Imaizumi, 2020; Kohler et al., 2023). In Section 5, we present a new convergence rate result that builds on the ideas of low ID and a hierarchical composition of functions particularly suited for DNNs.

5. Downstream Inference

The manifold assumption alone, however, cannot guarantee sufficient approximation rates in our setting. Even if the manifold dimension $d_{\mathcal{M}}$ is much smaller than the ambient dimension d (for example, $d_{\mathcal{M}} \approx 30$), an unreasonably high degree of smoothness would need to be assumed to allow for convergence rates below $n^{-1/4}$. In what follows, we give a more realistic assumption to achieve such rates. In particular, we combine the low-dimensional manifold structure in the feature space with a structural smoothness and sparsity assumption on the target function.

5.1. Structural Sparsity on the Manifold

Kohler & Langer (2021) recently derived convergence rates based on the following assumption.

Definition 5.1 (Hierarchical composition model, HCM)**.**

(a) We say that $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfies a HCM of level 0, if $f(x) = x_j$ for some $j \in \{1, \dots, d\}$.

(b) We say that f satisfies a HCM of level $k \geq 1$, if there is a s-smooth function $h\colon \mathbb{R}^p \to \mathbb{R}$ such that

$$f(x) = h\big(h_1(x), \dots, h_p(x)\big),$$

where $h_1, \dots, h_p : \mathbb{R}^d \to \mathbb{R}$ are HCMs of level $k-1$.

The collection $\mathcal{P}$ of all pairs $(s, p) \in \mathbb{R} \times \mathbb{N}$ appearing in the specification is called the constraint set *of the HCM.*

The assumption includes the case of sparse linear and (generalized) additive models as a special case but is much more general. Kohler & Langer (2021) and Schmidt-Hieber (2020) exploit such a structure to show that neural networks can approximate the target function at a rate that is only determined by the worst-case pair (s, p) appearing in the constraint set. It already follows from Lemma 4.2 that the constraint set of such a model is not invariant to ILTs of

the input space. Furthermore, the assumption does not exploit the potentially low intrinsic dimensionality of the input space. To overcome these limitations, we propose a new assumption combining the input space's manifold structure with the hierarchical composition model.

Assumption 5.2. The target function f_0 can be decomposed as $f_0 = f \circ \psi$, where $\mathcal{M}$ is a smooth, compact, $d_{\mathcal{M}}$-dimensional manifold, $\psi \colon \mathcal{M} \to \mathbb{R}^p$ is s_ψ-smooth, and f is a HCM of level $k \in \mathbb{N}$ with constraint set $\mathcal{P}$.

Whitney's embedding theorem (e.g., Lee & Lee, 2012, Chapter 6) allows any smooth manifold to be smoothly embedded into $\mathbb{R}^{2d_{\mathcal{M}}}$. This corresponds to a mapping ψ with $s_\psi = \infty$ and $p = 2d_{\mathcal{M}}$ in the assumption above. If not all information in the pre-trained representation Z is relevant, however, p can be much smaller. Importantly, Assumption 5.2 is not affected by ILTs.

Lemma 5.3 (Invariance of Assumption 5.2 under ILTs)**.** *Let Q be any ILT. If f_0 satisfies Assumption 5.2 for a given $\mathcal{P}$ and $(s_\psi, d_{\mathcal{M}})$, then $\tilde{f}_0 = f_0 \circ Q^{-1}$ satisfies Assumption 5.2 with the same $\mathcal{P}$ and $(s_\psi, d_{\mathcal{M}})$,*

5.2. Convergence Rate of DNNs

We now show that DNNs can efficiently exploit this structure. Let $(Y_i, Z_i)_{i=1}^n$ be i.i.d. observations and ℓ be a loss function. Define

$$f_0 = \operatorname*{arg\,min}_{f \colon \mathbb{R}^d \to \mathbb{R}} \mathbb{E}[\ell(f(Z), Y)],$$

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}(L_n, \nu_n)} \frac{1}{n} \sum_{i=1}^n \ell(f(Z_i), Y_i),$$

where $\mathcal{F}(L, \nu)$ is the set of feed-forward neural networks with L layers and ν neurons per layer. Let $Z \sim P_Z$ and define the $L_2(P_Z)$-norm of a function f as $\|f\|^2_{L_2(P_Z)} = \int f(z)^2 dP(z)$. We make the following assumption on the loss function ℓ.

Assumption 5.4. There is $a, b \in (0, \infty)$ such that

$$\frac{\mathbb{E}[\ell(f(Z), Y)] - \mathbb{E}[\ell(f_0(Z), Y)]}{\|f - f_0\|^2_{L_2(P_Z)}} \in [a, b].$$

Assumption 5.4 is satisfied for the squared and logistic loss, among others (e.g., Farrell et al., 2021, Lemma 8).

Theorem 5.5. *Suppose Assumption 5.2 and Assumption 5.4 hold. There are sequences L_n, ν_n and a corresponding sequence of neural network architectures $\mathcal{F}(L_n, \nu_n)$ such that (up to $\log n$ factors)*

$$\|\hat{f} - f_0\|_{L_2(P_Z)} = O_p\left(\max_{(s,p) \in \mathcal{P} \cup (s_\psi, d_{\mathcal{M}})} n^{-\frac{s}{2s+p}} \right).$$

The result shows that the convergence rate of the neural networks is only determined by the worst-case pair (s, p) appearing in the constraint set of the HCM and the embedding map ψ. The theorem extends the results of Kohler & Langer (2021) in two ways. First, it allows for more general loss functions than the square loss. This is important since classification methods are often used to adjust for confounding effects. Second, it explicitly exploits the manifold structure of the input space, which may lead to much sparser HCM specifications and dramatically improved rates.

5.3. Validity of DML Inference

In the previous sections, we explored plausible conditions under which the ATE is identifiable, and DNNs can estimate the nuisance functions with fast rates. We now combine our findings to give a general result for the validity of DML from pre-trained representations.

For binary treatment $T \in \{0, 1\}$ and pre-trained representations Z, we define the outcome regression function

$$g(t, z) \coloneqq \mathbb{E}[Y | T = t, Z = z],$$

and the propensity score

$$m(z) \coloneqq \mathbb{P}[T = 1 | Z = z].$$

Suppose we are given an i.i.d. sample $(Y_i, Z_i, T_i)_{i=1}^n$. DML estimators of the ATE are typically based on a cross-fitting procedure. Specifically, let $\bigcup_{k=1}^K I_k = \{1, \dots, n\}$ be a partition of the sample indices such that $|I_k|/n \to 1/K$. Let $\hat{g}^{(k)}$ and $\hat{m}^{(k)}$ denote estimators of g and m computed only from the samples $(Y_i, Z_i, T_i)_{i \notin I_k}$. Defining

$$\widehat{\mathrm{ATE}}^{(k)} = \frac{1}{|I_k|} \sum_{i \in I_k} \rho(T_i, Y_i, Z_i; \hat{g}^{(k)}, \hat{m}^{(k)}),$$

with orthogonalized score

$$\rho(T_i, Y_i, Z_i; g, m) = g(1, Z_i) - g(0, Z_i) + \frac{T_i(Y_i - g(1, Z_i))}{m(Z_i)} + \frac{(1 - T_i)(Y_i - g(0, Z_i))}{1 - m(Z_i)},$$

the final DML estimate of ATE is given by

$$\widehat{\mathrm{ATE}} = \frac{1}{K} \sum_{k=1}^K \widehat{\mathrm{ATE}}^{(k)}.$$

We need the following additional conditions.

Assumption 5.6. It holds

$$\max_{t \in \{0,1\}} \mathbb{E}[|g(t, Z)|^5] < \infty, \quad \mathbb{E}[|Y|^5] < \infty,$$

$$\mathbb{E}[|Y - g(T, Z)|^2] > 0, \quad \Pr(m(Z) \in (\varepsilon, 1 - \varepsilon)) = 1,$$

for some $\varepsilon > 0$.

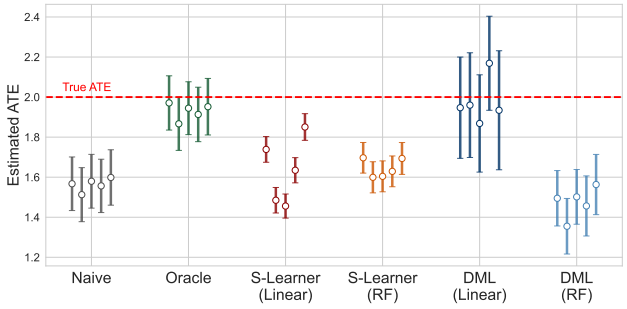

Figure 4. Label Confounding: Comparison of ATE estimators on the IMDb dataset. DML and S-Learner use pre-trained representations. Point estimates and 95% CI's are depicted.

The first two conditions ensure that the tails of Y and $g(t, Z)$ are not too heavy. The second two conditions are required for the ATE to be identifiable.

Theorem 5.7. *Suppose the pre-trained representation is P-valid, Assumption 5.6 holds, and the outcome regression and propensity score functions g and m satisfy Assumption 5.2 with constraints $\mathcal{P}_g \cup (s_\psi, d_\mathcal{M})$ and $\mathcal{P}_m \cup (s'_\psi, d_\mathcal{M})$, respectively. Suppose further*

$$\min_{(s,p)\in\mathcal{P}_g\cup(s_\psi,d_\mathcal{M})} \frac{s}{p} \times \min_{(s',p')\in\mathcal{P}_m\cup(s'_\psi,d_\mathcal{M})} \frac{s'}{p'} > \frac{1}{4}, \quad (5)$$

and the estimators $\hat{g}^{(k)}$ and $\hat{m}^{(k)}$ are DNNs as specified in Theorem 5.5 with the restriction that $\hat{m}^{(k)}$ is clipped away from 0 and 1. Then

$$\sqrt{n}(\widehat{\text{ATE}} - \text{ATE}) \to \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \mathbb{E}[\rho(T_i, Y_i, Z_i; g, m)^2]$.

Condition (5) is our primary regularity condition, ensuring sufficiently fast convergence for valid DML inference. It characterizes the necessary trade-off between smoothness and dimensionality of the components in the HCM. In particular, it is satisfied when each component function in the model has input dimension less than twice its smoothness.

6. Experiments

In the following, we will complement our theoretical results from the previous section with empirical evidence from several experiments. The experiments include both images and text as non-tabular data, which act as the source of confounding in the ATE setting. Further experiments can be found in Appendix D.

6.1. Validity of ATE Inference from Pre-Trained Representations

Text Data We utilize the IMDb Movie Reviews dataset from Lhoest et al. (2021) consisting of 50,000 movie reviews labeled for sentiment analysis. The latent features Z as representations of the movie reviews are computed using the last hidden layer of the pre-trained Transformer-based model BERT (Devlin et al., 2019). More specifically, each review results in a 768-dimensional latent variable Z by extracting the [CLS] token that summarizes the entire sequence. For this, each review is tokenized using BERT's subword tokenizer (bert-base-uncased), truncated to a maximum length of 128 tokens, and padded where necessary.

Image Data We further use the dataset from Kermany et al. (2018) that contains 5,863 chest X-ray images of children. Each image is labeled according to whether the lung disease pneumonia is present or not. The latent features are obtained by passing the images through a pre-trained convolutional neural network and extracting the 1024-dimensional last hidden layer features of the model. We use the pre-trained Densenet-121 model from the TorchXRayVision library (Cohen et al., 2022), which was trained on a large publicly available chest X-rays dataset (Cohen et al., 2020).

Confounding Setup For both data applications, we simulate treatment and outcome variables while inducing confounding based on the labels. As an example, for the modified image dataset, children with pneumonia have a higher chance of receiving treatment compared to healthy children. In contrast, pneumonia negatively impacts the outcome variable. The same confounding is present in our modified text dataset. Hence, the label creates a negative bias in both ATE settings if not properly accounted for. Further details about the confounding setups are provided in Appendix C.

ATE Estimators We compare the performance of DML using both linear and random forest (RF) based nuisance function estimators. For comparison, we also include another common causal estimator, called S-Leaner, which only estimates the outcome function (2) (details in Appendix B.1). In each of the simulations, estimators facilitate the information contained in the non-tabular data to adjust for confounding by using the latent features from the pre-trained models in the estimation. As a benchmark, we compare the estimate to the ones of a *Naive* estimator (unadjusted estimation) and the *Oracle* estimator (adjusts for the true label).

Label Confounding Results The results for the IMDb simulation over 5 simulations are depicted in Figure 4. As expected, the naive estimator shows a strong negative bias. The same can be observed for the S-Learner (for both nuisance estimators) and for DML using random forest. In contrast, DML with linear nuisance estimator yields unbiased estimates with good coverage, as can be seen by the confidence intervals. First, these results indicate that DML seems to benefit from the double robust estimation. Second, DML fails when using random forest nuisance estimators. A random forest cannot achieve sufficiently fast convergence

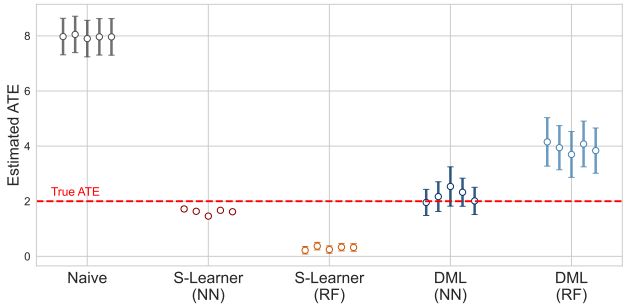

Figure 5. Complex Confounding: Comparison of ATE estimators on the X-ray dataset. DML and S-Learner use pre-trained representations. Point estimates and 95% CI's are depicted.

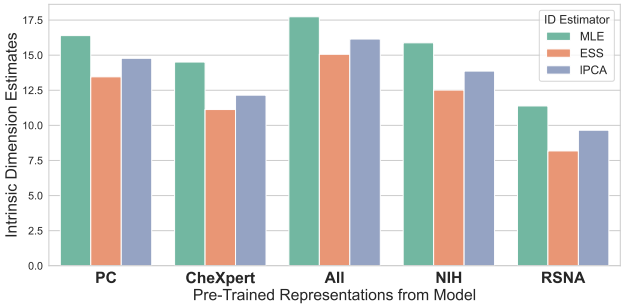

Figure 6. Different Intrinsic Dimension (ID) Estimates of Pre-Trained Representations obtained from different pre-trained models. Representations are based on the X-Ray dataset.

rates without structural sparsity assumptions. Such assumptions are unlikely to hold due to their sensitivity to ILTs. The results for image-based simulation are given in Appendix D, where the same phenomenon can be observed.

6.2. Neural Networks Adapt to Functions on Low Dimensional Manifolds

In a second line of experiments, we investigate the ability of neural networks to adapt to low intrinsic dimensions. The features in our data sets already concentrate on a low-dimensional manifold. For example, Figure 6 shows that the intrinsic dimension of the X-ray images is around $d_{\mathcal{M}} = 12$, whereas the ambient dimension is $d = 1024$. To simulate complex confounding with structural smoothness and sparsity, we first train an autoencoder (AE) with 5-dimensional latent space on the pre-trained representations. These AE-encodings are then used to simulate confounding. The encoder-then-linear function is a multi-layered hierarchical composition as in Assumption 5.2. We refer to this as *complex* confounding.

Complex Confounding Results We again compare DML to the S-Learner. In contrast to the previous section, we now use a neural network (with ReLU activation, 100 hidden layers with 50 neurons each) instead of a linear model in the outcome regression nuisance estimation. Similar to the previous experiments, we find that the naive estimate is strongly biased similar to the random forest-based estimators. In contrast, the neural network-based estimators exhibit much less bias. While the S-Learner's confidence intervals are too optimistic, the DML estimator shows high coverage and is therefore the only estimator that enables valid inference. The results for the IMDb dataset with complex confounding are given in Appendix D.

Low Intrinsic Dimension We also investigate the low intrinsic dimension hypothesis about pre-trained representations. Using different intrinsic dimension (ID) estimators such as the Maximum Likelihood (MLE) (Levina & Bickel, 2004), the Expected Simplex Skewness (ESS), and the Local Principal Component Analysis (lPCA) we estimate the ID of different pre-trained representations of the X-ray dataset obtained from different pre-trained models from the TorchXRayVision library (Cohen et al., 2022). The results in Figure 6 indicate that the intrinsic dimension of the pre-trained representations is much smaller than the dimension of the ambient space (1024). A finding that is in line with previous research further discussed in Appendix D.

7. Discussion

In this work, we explore ATE estimation under confounding induced by non-tabular data. We investigate conditions under which pre-trained neural representations can effectively be used to adjust for such kind of confounding. While the representations typically have lower dimensionality, their invariance under orthogonal transformations challenges common assumptions to obtain fast nuisance function convergence rates, like sparsity and additivity. Instead, the study leverages the concept of low intrinsic dimensionality, combining it with invariance properties and structural sparsity to establish conditions for fast convergence rates in nuisance estimation. This ensures valid ATE estimation and inference, contributing both theoretical insights and practical guidance for integrating machine learning into causal inference.

Limitations and Future Research In this work, we focus on a single source of confounding from a non-tabular data modality. A potential future research direction is to study the influence of multiple modalities on ATE estimation. In particular, having multiple modalities requires further causal and structural assumptions on the interplay of the modalities. This could, e.g., mean that each modality is best processed by a separate network or that the confounding information can only be extracted through a joint network that correctly fuses modalities at some point. We note, however, that this is more of a technical aspect and a matter of domain knowledge, and thus being of minor relevance for the discussion and theoretical contributions of our study.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. Doubleml-an object-oriented implementation of double machine learning in python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. Inference for regression with variables generated from unstructured data. *arXiv preprint arXiv:2402.15585*, 2024.

Bickel, P. J. and Li, B. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pp. 177–186, 2007.

Charig, C. R., Webb, D. R., Payne, S. R., and Wickham, J. E. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, 292 (6524):879–882, 1986.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., and Zhao, Z. Causalml: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*, 2020.

Chen, M., Jiang, H., Liao, W., and Zhao, T. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Chernozhukov, V., Newey, W., Quintas-Martınez, V. M., and Syrgkanis, V. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pp. 3901–3914. PMLR, 2022.

Christgau, A. M. and Hansen, N. R. Efficient adjustment for complex covariates: Gaining efficiency with dope. *arXiv preprint arXiv:2402.12980*, 2024.

Cohen, J. P., Hashir, M., Brooks, R., and Bertrand, H. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, pp. 136–155. PMLR, 2020.

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pp. 231–249. PMLR, 2022.

Dai, B., Shen, X., and Wang, J. Embedding learning. *Journal of the American Statistical Association*, 117(537): 307–319, 2022.

Deaton, A. and Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Erhan, D., Courville, A., Bengio, Y., and Vincent, P. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 201–208. JMLR Workshop and Conference Proceedings, 2010.

Farrell, M. H., Liang, T., and Misra, S. Deep neural networks for estimation and inference. *Econometrica*, 89(1): 181–213, 2021.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Gong, S., Boddeti, V. N., and Jain, A. K. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3987–3996, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jerzak, C. T., Johansson, F., and Daoud, A. Estimating causal effects under image confounding bias with an application to poverty in africa. *arXiv preprint arXiv:2206.06410*, 2022a.

Jerzak, C. T., Johansson, F., and Daoud, A. Image-based treatment effect heterogeneity. *arXiv preprint arXiv:2206.06417*, 2022b.

Jerzak, C. T., Johansson, F., and Daoud, A. Integrating earth observation data into causal inference: challenges and opportunities. *arXiv preprint arXiv:2301.12985*, 2023.

Julious, S. A. and Mullee, M. A. Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481, 1994.

Kallus, N., Mao, X., and Udell, M. Causal inference with noisy and missing covariates via matrix factorization. *Advances in neural information processing systems*, 31, 2018.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Klaassen, S., Teichert-Kluge, J., Bach, P., Chernozhukov, V., Spindler, M., and Vijaykumar, S. Doublemldeep: Estimation of causal effects with multimodal data. *arXiv preprint arXiv:2402.01785*, 2024.

Kohler, M. and Langer, S. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Kohler, M., Langer, S., and Reif, U. Estimation of a regression function on a manifold by fully connected deep neural networks. *Journal of Statistical Planning and Inference*, 222:160–181, 2023. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2022.05.008.

Konz, N. and Mazurowski, M. A. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *International Conference on Learning Representations*, 2024.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Lee, J. M. and Lee, J. M. *Smooth manifolds*. Springer, 2012.

Levina, E. and Bickel, P. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pp. 7512–7523. PMLR, 2021.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Nakada, R. and Imaizumi, M. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

Ng, A. Y. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first International Conference on Machine Learning*, pp. 78, 2004.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Raskutti, G., Yu, B., and Wainwright, M. J. Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. *Advances in Neural Information Processing Systems*, 22, 2009.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Schmidt-Hieber, J. Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 2020.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pp. 1040–1053, 1982.

Stone, C. J. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

Van der Laan, M. J. and Rose, S. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Van Der Laan, M. J. and Rubin, D. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes: With applications to statistics*. Springer Nature, 2023.

Veitch, V., Wang, Y., and Blei, D. Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Veitch, V., Sridhar, D., and Blei, D. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pp. 919–928. PMLR, 2020.

Zhang, J., Xue, W., Yu, Y., and Tan, Y. Debiasing machine-learning-or ai-generated regressors in partial linear models. *Available at SSRN*, 2023.

Zhang, Y. and Bradic, J. Causal inference through multi-stage learning and doubly robust deep neural networks. *arXiv preprint arXiv:2407.08560*, 2024.

A. Proofs and Additional Results

A.1. Equivalence Class of Representations

Lemma A.1 (Equivalence Class of Representations)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $Z : \Omega \to \mathbb{R}^d$ be a measurable map (a random* representation*). Then for each ILT Q the random variable $Q(Z)$ satisfies*

$$\sigma\big(Q(Z)\big) \;=\; \sigma(Z),$$

where $\sigma(Z)$ denotes the σ-algebra generated by the random variable Z. Consequently,

$$\mathcal{Z} \;=\; \big\{Q(Z) \;\mid\; Q \in \mathcal{Q}\big\}$$

forms an equivalence class of representations that are indistinguishable from the viewpoint of measurable information.

Proof. Each $Q \in \mathcal{Q}$ is an invertible linear transformation. Consequently, Q is a Borel measurable bijection with a Borel measurable inverse. To show $\sigma(Q(Z)) = \sigma(Z)$, consider any Borel set $B \subseteq \mathbb{R}^d$. We have

$$\{\omega \in \Omega : Q(Z(\omega)) \in B\} = \{\omega \in \Omega : Z(\omega) \in Q^{-1}(B)\}.$$

Since $Q^{-1}(B)$ is Borel (as Q is a Borel isomorphism), the pre-image $\{\omega : Z(\omega) \in Q^{-1}(B)\}$ belongs to $\sigma(Z)$. Similarly, for any Borel set $A \subseteq \mathbb{R}^d$,

$$\{\omega \in \Omega : Z(\omega) \in A\} = \{\omega \in \Omega : Q(Z(\omega)) \in Q(A)\},$$

which belongs to $\sigma(Q(Z))$. Therefore, $\sigma(Q(Z)) = \sigma(Z)$. □

A.2. Proof of Lemma 4.1

Proof. We consider f being C^s on the open domain $D \subseteq \mathbb{R}^d$, so by definition, all partial derivatives of f up to order s exist and are continuous on D. Further, we consider any invetible matrix Q. Such linear transformations are known to be infinitely smooth (as all their partial derivatives of any order exist and are constant, hence continuous). Hence, the function $h = f \circ Q^{-1}$ is the composition of a C^s function f with a linear and thus C^∞ map Q^{-1}.

Applying the multivariate chain rule, we can easily verify that the differentiability properties of h are inherited from those of f and the linear transformation Q^{-1}. Specifically, since Q^{-1} is C^∞, and f is C^s, their composition h retains the C^s smoothness. Lastly, the (transformed) domain $Q(D)$ is also open as linear (and thus continuous) transformations preserve the openness of sets in $\mathbb{R}^d$. Therefore, h is well-defined and C^s on $Q(D)$. □

A.3. Proof of Lemma 4.2

Proof. Suppose that Q is an invertible matrix representing the linear map $z \mapsto Q(z)$. Denote by $\tilde{Q} = Q^{-1}$ its inverse and its rows by $\tilde{q}_1, \ldots, \tilde{q}_d$.

(i) Additivity

Assume that f is additive, i.e.,

$$f(x) = \sum_{j=1}^{d} f_j(x_j),$$

and that at least one f_j is nonlinear. Define $h(\tilde{x}) = h(Q^{-1}\tilde{x})$. We have

$$h(\tilde{x}) = \sum_{j=1}^{d} f_j(\tilde{q}_j^\top \tilde{x}).$$

Assume without loss of generality that f_1 is nonlinear. The set of invertible matrices where $\tilde{q}_1$ equals a multiple of a standard basis vector has Haar measure 0. Hence, $f_1(\tilde{q}_1^\top \tilde{x})$ is a nonlinear function of multiple coordinates of $\tilde{x}$, implying that h is not additive.

(ii) Sparsity

Assume f is sparse linear of the form $f(x) = \beta^\top x$ with $1 \leq \|\beta\|_0 < d$. We have $h(\tilde{x}) = f(Q^{-1}\tilde{x}) = \beta^\top Q^{-1}\tilde{x} =: \tilde{\beta}^\top \tilde{x}$. While the map h is still linear, the set of matrices Q such that $\|\tilde{\beta}\|_0 = \|\beta^\top Q^{-1}\|_0 \neq d$ has Haar measure zero. □

A.4. Proof of Lemma 4.3

Proof. As in the previous proof in Appendix A.2, it is essential to note that ILTs Q are linear, invertible maps that are C^∞ (infinitely differentiable) with inverses that are likewise C^∞. Specifically, Q serves as a global diffeomorphism on $\mathbb{R}^d$, ensuring that both Q and Q^{-1} are smooth (C^∞) functions.

Given that M is a $d_{\mathcal{M}}$-dimensional smooth manifold, for each point x on the manifold ($x \in M$), there exists a neighborhood $U \subseteq M$ and a smooth chart $\varphi : U \to \mathbb{R}^{d_{\mathcal{M}}}$ that is a diffeomorphism onto its image. Applying the orthogonal transformation Q to M results in the set $Q(M)$, and correspondingly, the image $Q(U) \subseteq Q(M)$. To construct a smooth chart for $Q(M)$, we can consider the map

$$\tilde{\varphi} : Q(U) \to \mathbb{R}^{d_{\mathcal{M}}}, \quad \tilde{\varphi}(Q(x)) = \varphi(x),$$

where $x \in U$. Since Q is a diffeomorphism, the composition $\tilde{\varphi} = \varphi \circ Q^{-1}$ restricted to $Q(U)$ remains a smooth diffeomorphism onto its image. Hence, this defines a valid smooth chart for $Q(M)$. Covering $Q(M)$ with such transformed charts derived from those of M ensures that $Q(M)$ inherits a smooth manifold structure. Each chart $\tilde{\varphi}$ smoothly maps an open subset of $Q(M)$ to an open subset of $\mathbb{R}^{d_{\mathcal{M}}}$, preserving the intrinsic dimension. Therefore, the intrinsic dimension $d_{\mathcal{M}}$ of the manifold M is preserved under any orthogonal transformation Q, and $Q(M)$ remains a $d_{\mathcal{M}}$-dimensional smooth manifold in $\mathbb{R}^d$. □

A.5. Proof of Lemma 5.3

Proof. Recall that Q is an invertible linear map, $f_0 = f \circ \psi \colon \mathcal{M} \to \mathbb{R}$, and $\tilde{f}_0 = f_0 \circ \psi \circ Q^{-1} \colon Q(\mathcal{M}) \to \mathbb{R}$. Write $\tilde{f} = f \circ \tilde{\psi}$ with $\tilde{\psi} = \psi \circ Q^{-1} \colon Q(\mathcal{M}) \to \mathbb{R}$. Since $\mathcal{M}$ is a smooth manifold, $Q(\mathcal{M})$ is a smooth manifold with the same intrinsic dimension $d_{\mathcal{M}}$ by Lemma 4.3. Since $z \mapsto Q^{-1}$ is continuous and $\mathcal{M}$ is compact, $Q(\mathcal{M})$ is also compact. Next, since ψ is s_ψ-smooth by assumption, $\tilde{\psi}$ is also s_ψ-smooth by Lemma 4.1. Finally, the HCM part f in the two models f_0 and $\tilde{f}_0$ is the same, so they share the same constraint set $\mathcal{P}$. This concludes the proof. □

A.6. Proof of Theorem 5.5

We will use Theorem 3.4.1 of Van der Vaart & Wellner (2023) to show that the neural network $\hat{f}$ converges at the rate stated in the theorem. For ease of reference we re-state a slightly simplified version of the theorem adapted to the notation used in our paper. Here and in the following, we write $a \lesssim b$ to indicate $a \leq Cb$ for a constant $C \in (0, \infty)$ not depending on n.

Proposition A.2. *Let $\mathcal{F}_n$ be a sequence of function classes, ℓ be some loss function, f_0 the estimation target, and*

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \ell(f(Z_i), Y_i).$$

Define $\mathcal{F}_{n,\delta} = \{f \in \mathcal{F}_n \colon \|f - f_0\|_{L_2(P_Z)} \leq \delta\}$ and suppose that for every $\delta > 0$, it holds

$$\inf_{f \in \mathcal{F}_{n,\delta} \setminus \mathcal{F}_{n,\delta/2}} \mathbb{E}[\ell(f(Z), Y)] - \mathbb{E}[\ell(f_0(Z), Y)] \gtrsim \delta^2, \tag{A.2.1}$$

and, writing $\bar{\ell}_f(z, y) = \ell(f(z), y) - \ell(f_0(z), y)$, that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_{n,\delta}} \left| \frac{1}{n} \sum_{i=1}^{n} \bar{\ell}_f(Z_i, Y_i) - \mathbb{E}[\bar{\ell}_f(Z, Y)] \right|\right] \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}, \tag{A.2.2}$$

for functions $\phi_n(\delta)$ such that $\delta \mapsto \phi_n(\delta)/\delta^{2-\varepsilon}$ is decreasing for some $\varepsilon > 0$. If there are $\tilde{f}_0 \in \mathcal{F}_n$ and $\varepsilon_n \geq 0$ such that

$$\varepsilon_n^2 \gtrsim \mathbb{E}[\ell(\tilde{f}_0(Z), Y)] - \mathbb{E}[\ell(f_0(Z), Y)], \tag{A.2.3}$$

$$\phi_n(\varepsilon_n) \lesssim \sqrt{n}\varepsilon_n^2, \tag{A.2.4}$$

it holds $\|\hat{f} - f_0\|_{L_2(P_Z)} = O_p(\varepsilon_n)$.

Proof of Theorem 5.5. Define $(s^*, d^*) = \arg\min_{(s,p)\in\mathcal{P}\cup(s_\psi, d_\mathcal{M})} s/p$ and denote the targeted rate of convergence by

$$\varepsilon_n = \max_{(s,p)\in\mathcal{P}\cup(s_\psi,d_\mathcal{M})} n^{-\frac{s}{2s+p}} (\log n)^4 = n^{-\frac{s^*}{2s^*+d^*}} (\log n)^4.$$

We now check the conditions of Proposition A.2.

Condition (A.2.1)**:** Follows from Assumption 5.4, since

$$\inf_{f\in\mathcal{F}_{n,\delta}\setminus\mathcal{F}_{n,\delta/2}} \mathbb{E}[\ell(f(Z),Y)] - \mathbb{E}[\ell(f_0(Z),Y)] \geq \inf_{f\in\mathcal{F}_{n,\delta}\setminus\mathcal{F}_{n,\delta/2}} a\|f - f_0\|^2_{L_2(P_Z)} \geq \frac{a}{4}\delta^2.$$

Condition (A.2.2)**:** Let $N(\varepsilon, \mathcal{F}, L_2(Q))$ be the minimal number of ε-balls required to cover $\mathcal{F}$ in the $L_2(Q)$-norm. Theorem 2.14.2 of Van der Vaart & Wellner (2023) states that eq. (A.2.2) holds with

$$\phi_n(\delta) = J_n(\delta)\left(1 + \frac{J_n(\delta)}{\delta^2\sqrt{n}}\right),$$

where

$$J_n(\delta) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon, \mathcal{F}(L,\nu), L_2(Q))}d\epsilon,$$

with the supremum taken over all probability measures Q. Lemma A.3 in Appendix A.7 gives

$$J_n(\delta) \lesssim \delta\sqrt{\log(1/\delta)}L\nu\sqrt{\log(L\nu)},$$

which implies that $\delta \mapsto \phi_n(\delta)/\delta^{2-1/2}$ is decreasing, so the condition is satisfied.

Condition (A.2.3)**:** According to Lemma A.4 in Appendix A.7 there are sequences $L_n = O(\log \varepsilon_n^{-1})$, $\nu_n = O(\varepsilon_n^{-d^*/2s^*})$ such that there is a neural network $\widetilde{f}_0 \in \mathcal{F}(L_n, \nu_n)$ with

$$\sup_{z\in\mathcal{M}} |\widetilde{f}_0(z) - f_0(z)| = O(\varepsilon_n).$$

Together with Assumption 5.4, this implies

$$\mathbb{E}[\ell(\tilde{f}_0(Z),Y)] - \mathbb{E}[\ell(f_0(Z),Y)] \leq b\|\tilde{f}_0 - f_0\|^2_{L_2(P_Z)} \leq b \sup_{z\in\mathcal{M}} |\widetilde{f}_0(z) - f_0(z)|^2 \lesssim \varepsilon_n^2,$$

as required.

Condition (A.2.4)**:** Using $L_n = O(\log \varepsilon_n^{-1})$, $\nu_n = O(\varepsilon_n^{-d^*/2s^*})$ and our bound on $J_n(\delta)$ from Lemma A.3, we get

$$J_n(\delta) \lesssim \delta \log^{1/2}(\delta^{-1})\varepsilon_n^{-\frac{d^*}{2s^*}} \log^{3/2}(\varepsilon_n^{-1}).$$

Now observe that

$$\begin{aligned}
\frac{\phi_n(\varepsilon_n)}{\varepsilon_n^2} &\lesssim \varepsilon_n^{-\frac{d^*}{s^*}-1}\log^2(\varepsilon_n^{-1}) + \frac{\varepsilon_n^{-\frac{d^*}{s^*}-2}\log^4(\varepsilon_n^{-1})}{\sqrt{n}} \\
&= \varepsilon_n^{-\frac{2s^*+d^*}{2s^*}}\log^2(\varepsilon_n^{-1}) + \varepsilon_n^{-\frac{2s^*+d^*}{s^*}}\log^4(\varepsilon_n^{-1})n^{-1/2} \\
&\lesssim n^{1/2}(\log n)^{-2} + n^{1/2},
\end{aligned}$$

where the last step follows from our definition of ε_n and the fact that $\log(\varepsilon_n^{-1}) \lesssim \log n$. In particular, ε_n satisfies $\phi_n(\varepsilon_n) \lesssim \sqrt{n}\varepsilon_n^2$, which concludes the proof of the theorem. □

A.7. Auxiliary results

Lemma A.3. *Let $\mathcal{F}(L,\nu)$ be a set of neural networks with $\sup_{f\in\mathcal{F}(\mathcal{L},\nu)}\|f\|_\infty<\infty$. For all $\delta>0$ sufficiently small, it holds*

$$\sup_Q\int_0^\delta\sqrt{1+\log N(\epsilon,\mathcal{F}(L,\nu),L_2(Q))}d\epsilon\lesssim\delta\sqrt{\log(1/\delta)}L\nu\sqrt{\log(L\nu)}.$$

Proof. Denote by $\mathrm{VC}(\mathcal{F})$ the Vapnik-Chervonenkis dimension of the set $\mathcal{F}$. By Theorem 2.6.7 in Van der Vaart & Wellner (2023), it holds

$$\sup_Q\log N(\varepsilon,\mathcal{F},L_2(Q))\lesssim\log(1/\varepsilon)\mathrm{VC}(\mathcal{F}),$$

for $\varepsilon>0$ sufficiently small. By Theorem 7 of Bartlett et al. (2019), we have

$$\mathrm{VC}(\mathcal{F}(L,\nu))\lesssim L^2\nu^2\log(L\nu).$$

For small ε, this gives

$$\sup_Q\sqrt{1+\log N(\varepsilon,\mathcal{F}(L,\nu),L_2(Q))}\lesssim\sqrt{\log(1/\varepsilon)}L\nu\sqrt{\log(L\nu)},$$

Integrating the right-hand side gives the desired result. □

Lemma A.4. *Suppose f_0 satisfies Assumption 5.2 for a given constraint set $\mathcal{P}$ and $(s_\psi,d_\mathcal{M})$. Define $(s^*,d^*)=\arg\min_{(s,p)\in\mathcal{P}\cup(s_\psi,d_\mathcal{M})}s/p$. Then for any $\varepsilon>0$ sufficiently small, there is a neural network architecture $\mathcal{F}(L,\nu)$ with $L=O(\log\varepsilon^{-1})$, $\nu=O(\varepsilon^{-d^*/2s^*})$ such that there is $\widetilde{f}_0\in\mathcal{F}(L,\nu)$ with*

$$\sup_{z\in\mathcal{M}}|\widetilde{f}_0(z)-f_0(z)|=O(\varepsilon).$$

Proof. The proof proceeds in three steps. We first approximate the embedding component ψ by a neural network $\widetilde{\psi}$, then the HCM component f by a neural network $\widetilde{f}$. Finally, we concatenate the networks to approximate the composition $f_0=f\circ\psi$ by $\widetilde{f}_0=\widetilde{f}\circ\widetilde{\psi}$.

Approximation of the embedding component. Recall that $\psi\colon\mathcal{M}\to\mathbb{R}^d$ is a s_ψ-smooth mapping. Write $\psi(z)=(\psi_1(z),\dots,\psi_d(z))$ and note that each $\psi_j\colon\mathcal{M}\to\mathbb{R}$ is also s_ψ-smooth. Since $\mathcal{M}$ is a smooth $d_\mathcal{M}$-dimensional manifold, it has Minkowski dimension $d_\mathcal{M}$. Then Theorem 2 of Kohler et al. (2023) (setting $M=\varepsilon^{-1/2s_\psi}$ in their notation) implies that there is a neural network $\widetilde{\psi}_j\in\mathcal{F}(L_\psi,\nu_\psi)$ with $L_\psi=O(\log\varepsilon^{-1})$ and $\nu_\psi=O(\varepsilon^{-d_\mathcal{M}/2s_\psi})$ such that

$$\sup_{z\in\mathcal{M}}|\widetilde{\psi}_j(z)-\psi_j(z)|=O(\varepsilon).$$

Parallelize the networks $\widetilde{\psi}_j$ into a single network $\widetilde{\psi}:=(\widetilde{\psi}_1,\dots,\widetilde{\psi}_d)\colon\mathcal{M}\to\mathbb{R}^d$. By construction, the parallelized network $\widetilde{\psi}$ has L_ψ layers, width $d\times\nu_\psi=O(\nu_\psi)$, and satisfies

$$\sup_{z\in\mathcal{M}}\|\widetilde{\psi}(z)-\psi(z)\|=O(\varepsilon).$$

Approximation of the HCM component. Let $a\in(0,\infty)$ be arbitrary. By Theorem 3(a) of Kohler & Langer (2021) (setting $M_{i,j}=\varepsilon^{-1/2p_j^{(i)}}$ in their notation), there is a neural network $\widetilde{f}\in\mathcal{F}(L_f,\nu_f)$ with $L_f=O(\log\varepsilon^{-1})$ and $\nu_f=O(\varepsilon^{-d^*/2s^*})$ such that

$$\sup_{x\in[-a,a]^d}|\widetilde{f}(x)-f(x)|=O(\varepsilon),$$

Combined approximation. Now concatenate the networks $\widetilde{\psi}$ and $\widetilde{f}$ to obtain the network $\widetilde{f}_0 = \widetilde{f} \circ \widetilde{\psi} \in \mathcal{F}(L_\psi + L_f, \max\{\nu_\psi, \nu_f\})$. Observe that $L_\psi + L_f = O(\log \varepsilon^{-1})$ and $\nu_\psi + \nu_f = O(\varepsilon^{-d^*/2s^*})$, so the network has the right size. It remains to show that its approximation error is sufficiently small. Define

$$\gamma := \sup_{z \in \mathcal{M}} \|\widetilde{\psi}(z) - \psi(z)\|,$$

which is $O(\varepsilon)$ by the construction of $\widetilde{\psi}$,

$$a := \sup_{z \in \mathcal{M}} \|\psi(z)\| + \gamma,$$

which is $O(1)$ by assumption, and

$$K := \sup_{x, x'} \frac{|f(x) - f(x')|}{\|x - x'\|},$$

which is finite since f is Lipschitz due to $\min_{(s,d) \in \mathcal{P}} s \geq 1$ and the fact that finite compositions of Lipschitz functions are Lipschitz. By the triangle inequality, we have

$$\begin{aligned}
\sup_{z \in \mathcal{M}} |\widetilde{f}_0(z) - f_0(z)| &\leq \sup_{z \in \mathcal{M}} |\widetilde{f}(\widetilde{\psi}(z)) - f(\widetilde{\psi}(z))| + \sup_{z \in \mathcal{M}} \|f(\widetilde{\psi}(z)) - f(\psi(z))\| \\
&\leq \sup_{x \in [-a,a]^d} |\widetilde{f}(x) - f(x)| + K \\
&= O(\varepsilon),
\end{aligned}$$

as claimed. □

A.8. Proof of Theorem 5.7

Proof. We validate the conditions of Theorem II.1 of Chernozhukov et al. (2017). Our Assumption 5.6 covers all their moment and boundedness conditions on g and m. By Theorem 5.5, we further know that

$$\|\hat{m}^{(k)} - m\|_{L_2(P_Z)} + \|\hat{g}^{(k)} - g\|_{L_2(P_Z)} = o_p(1).$$

Further, Theorem 5.5 yields

$$\begin{aligned}
\|\hat{m}^{(k)} - m\|_{L_2(P_Z)} \times \|\hat{g}^{(k)} - g\|_{L_2(P_Z)} &= O_p\left(\max_{(s,p) \in \mathcal{P}_g \cup (s_\psi, d_\mathcal{M})} n^{-\frac{s}{2s+p}} \times \max_{(s',p') \in \mathcal{P}_m \cup (s'_\psi, d_\mathcal{M})} n^{-\frac{s'}{2s'+p'}} \right) \\
&= O_p\left(\max_{(s,p) \in \mathcal{P}_g \cup (s_\psi, d_\mathcal{M})} \max_{(s',p') \in \mathcal{P}_m \cup (s'_\psi, d_\mathcal{M})} n^{-\left(\frac{s}{2s+p} + \frac{s'}{2s'+p'}\right)} \right).
\end{aligned}$$

We have to show that the term on the right is of order $o_p(n^{-1/2})$. Observe that

$$\begin{aligned}
\frac{s}{2s+p} + \frac{s'}{2s'+p'} > \frac{1}{2} \quad &\Leftrightarrow \quad \frac{1}{2+p/s} + \frac{1}{2+p'/s'} > \frac{1}{2} \\
&\Leftrightarrow \quad \frac{4 + p/s + p'/s'}{(2+p/s)(2+p'/s')} > \frac{1}{2} \\
&\Leftrightarrow \quad 4 + p/s + p'/s' > 2 + p/s + p'/s' + \frac{pp'}{2ss'} \\
&\Leftrightarrow \quad 4 > \frac{pp'}{ss'}.
\end{aligned}$$

Thus, our condition

$$\min_{(s,p) \in \mathcal{P}_g \cup (s_\psi, d_\mathcal{M})} \frac{s}{p} \times \min_{(s',p') \in \mathcal{P}_m \cup (s'_\psi, d_\mathcal{M})} \frac{s'}{p'} > \frac{1}{4},$$

implies

$$\|\hat{m}^{(k)} - m\|_{L_2(P_Z)} \times \|\hat{g}^{(k)} - g\|_{L_2(P_Z)} = o_p(n^{-1/2}),$$

as required. □

B. Additional Related Literature

B.1. Empirical Evidence of Low Intrinsic Dimensions

Using different ID estimators such as the maximum likelihood estimator (MLE; Levina & Bickel, 2004) on popular image datasets such as ImageNet (Deng et al., 2009), several works find clear empirical evidence for low ID of both the image data and related latent features obtained from pre-trained NNs (Gong et al., 2019; Ansuini et al., 2019; Pope et al., 2021). The existence of the phenomenon of low intrinsic dimensions was also verified in the medical imaging (Konz & Mazurowski, 2024) and text-domain (Aghajanyan et al., 2020). All of the mentioned research finds a striking inverse relation between intrinsic dimensions and (state-of-the-art) model performance, which nicely matches the previously introduced theory about ID-related convergence rates.

C. Experimental Details and Computing Environment

Simulation Setup

We conduct several simulation studies to investigate the performance of different Average Treatment Effect (ATE) estimators of a binary treatment on some outcome in the presence of a confounding induced by non-tabular data. In the experiments, the confounding is induced by the labels, i.e., the pneumonia status or the review as well as more complex functions of the pre-trained features. Nuisance function estimation is based on the pre-trained representations that are obtained from passing the non-tabular data through the pre-trained neural models and extracting the last hidden layer features.

Data and Pre-trained Models For the text data, we utilize the IMDb Movie Reviews dataset from Lhoest et al. (2021) consisting of 50,000 movie reviews labeled for sentiment analysis. For each review, we extract the [CLS] token, a 768-dimensional vector per review entry, of the pre-trained Transformer-based model BERT (Devlin et al., 2019). To process the text, we use BERT's subword tokenizer (bert-base-uncased) and truncate sequences to a maximum length of 128 tokens. We use padding if necessary. After preprocessing and extraction of pre-trained representations, we sub-sampled 1,000 and 4,000 pre-trained representations for the two confounding setups to make the simulation study tractable. For the image data simulation, we use the dataset from Kermany et al. (2018) that originally contains 5,863 chest X-ray images of children that were obtained from routine clinical care in the Guangzhou Women and Children's Medical Center, Guangzhou. We preprocess the data such that each patient appears only once in the dataset. This reduces the effective sample size to 3,769 chest X-rays. Each image is labeled according to whether the lung disease pneumonia is present or not. The latent features are obtained by passing the images through a pre-trained convolutional neural network and extracting the 1024-dimensional last hidden layer features of the model. We use the pre-trained Densenet-121 model from the TorchXRayVision library (Cohen et al., 2022), which was trained on a large publicly available chest X-rays dataset (Cohen et al., 2020). Specifically, we use the Densenet-121 with resolution 224×224 and the training data it was trained on (all). The representations are taken from this model as it showed superior performance in benchmark studies (Cohen et al., 2020). Note that the dataset from the Guangzhou Women and Children's Medical Center that we use, was not used during the training of the model. This is important from a theoretical and practical viewpoint, as the confounding simulation via labels might otherwise be too easy to adjust for given that the model could have memorized the input data. However, using this kind of data we rule out this possibility.

Confounding As introduced in the main text, we simulate confounding both on the true labels of the non-tabular data as well as encodings from a trained autoencoder. While this induces a different degree of complexity for the confounding, the simulated confounding is somewhat similar in both settings. We first discuss the simpler setting of *Label Confounding*. In all of the experiments, the true average treatment effect was chosen to be two.

Label Confounding Label confounding was induced by simulating treatment and outcome both dependent on the binary label. In the case of the label being one (so in case of pneumonia or in case of a positive review), the probability of treatment is 0.7 compared to 0.3 when the label is zero. The chosen probabilities guaranteed a sufficient amount of overlap between the two groups. The outcome is simulated by a linear model of the treatment times the ATE. We then add a linear term for the label as well as Gaussian noise. The linear term of the label is the label times a negative coefficient in order to induce a negative bias to the average treatment setup compared to a randomized setting. Overall, the simulated confounding matches the setup of the partial linear model. Given that the confounding simulation is only based on the labels, the study was in fact randomized with respect to any other source of confounding.

Complex Confounding To simulate complex confounding with structural smoothness and sparsity, we first train an autoencoder (AE) with 5-dimensional latent space on the pre-trained representations, both in the case of the text and image representations. These AE-encodings are then used to simulate confounding similarly as in the previous experiment. The only difference is that we now sample the coefficients for the 5-dimensional AE-encodings. For the propensity score, these are sampled from a normal distribution, while the sampled coefficients for outcome regression are restricted to be negative, to ensure a sufficiently larger confounding effect, that biases naive estimation. We choose a 5-dimensional latent space to allow for sufficiently good recovery of the original pre-trained representations.

Estimators We estimate the ATE using multiple methods across 5 simulation iterations. In each of these, we estimate a *Naive* estimator that simply regresses the outcome on treatment while not adjusting for confounding. The *Oracle* estimator uses a linear regression of outcome on both treatment and the pneumonia label that was used to induce confounding. The S-Learner estimates the outcome regression function $g(t, z) = \mathbb{E}[Y \mid T = t, Z = z]$ by fitting a single model $\hat{g}(t, z)$ to all data, treating the treatment indicator as a feature. The average treatment effect estimate of the S-Learner is then given by

$$\widehat{ATE}_S = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(1, z_i) - \hat{g}(0, z_i).$$

In contrast, the Double Machine Learning estimators estimates both the outcome regression function and the propensity score to obtain its double robustness property. In the experiments both the S-Learner and DML estimators are used in combination with linear and random forest-based nuisance estimators. DML (Linear) uses standard linear regression and logistic regression for the outcome and propensity score estimation respectively, while the S-Learner (Linear) only uses linear regression for the outcome regression. For the random forest-based estimation, a standard random forest implementation from *scikit-learn* is used. The number of estimated trees is varied in certain experiments to improve numerical stability. For the neural network-based estimators, we use neural networks with a depth of 100 and width of 50 while using ReLU activation and Adam for optimization. Besides using such a neural network nuisance estimator for the outcome regression, the DML (NN) estimator uses logistic regression to improve numerical stability. Generally, DML was used with sample splitting and with two folds for cross-validation. For the S-Learner and the DML Learner the Python packages *CausalML* (Chen et al., 2020) and *DoubleML* (Bach et al., 2022) are used, respectively.

Intrinsic Dimensions of Pre-trained Representations In Section 6.2 we also provide empirical evidence that validates the hypothesis of low intrinsic dimensions of pre-trained representations. For this, we use different pre-trained models from the from the TorchXRayVision library (Cohen et al., 2022). All of these are trained on chest X-rays and use a Densenet-121 (Huang et al., 2017) architecture. Given the same architecture of the models, the dimension of the last layer hidden features is 1024 for all models. The different names of the models on the x-axis indicate the dataset they were trained on. We use the 3,769 chest X-rays from the X-rays dataset described above and pass these through each pre-trained model to extract the last layer features of each model, which we call the pre-trained representations of the data. Subsequently, we use standard intrinsic dimension estimators such as the MLE (Levina & Bickel, 2004), the Expected Simplex Skewness estimator, and the Local Principal Component Analysis estimator, with a choice of number of neighbors set to 5, 25 and 50, respectively. While the intrinsic dimension estimates vary by the pre-trained model and the intrinsic dimension estimator used, the results indicated that the intrinsic dimension of the pre-trained representations is much smaller than the dimension of the ambient space (1024).

Computational environment All computations were performed on a user PC with Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz, 8 cores, and 16 GB RAM. Run times of each experiment do not exceed one hour.

D. Further Experiments

This section provides additional results from further experiments. The results depicted in Figure 7 and Figure 8 complement Figure 4 and Figure 5 that are discussed in Section 6. The results for the pneumonia simulation with label confounding over 5 simulations are depicted in Figure 7. As before, the naive estimator shows a strong negative bias. Similarly, the S-Learner (for both nuisance estimators) and for DML using random forest exhibit a negative bias and too narrow confidence intervals. In contrast, DML with linear nuisance estimator yields less biased estimates with good coverage due to its properly adapted confidence intervals. A similar pattern can be observed for the complex confounding setting in the IMDb data depicted in Figure 8. The naive estimator and both of the random forest-based ATE estimators exhibit strong bias. In contrast, both

neural network-based estimators show very little bias. This provides further evidence that neural networks can adapt to the low intrinsic dimension of the data. However, in contrast to the DML estimator, the S-Learner still shows too narrow confidence intervals and has thus poor coverage. As it was in the example discussed in the main body of the test, the DML (NN) estimator is the only estimator that yields unbiased estimates and valid inference.

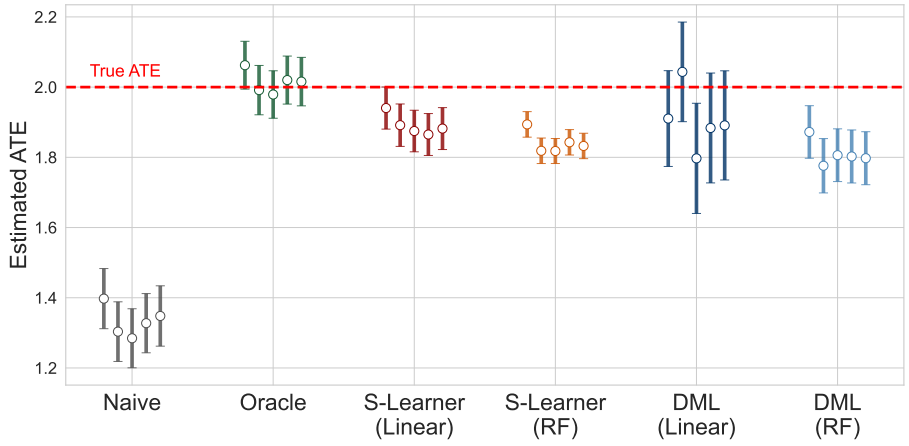

Figure 7. Label Confounding: Comparison of ATE estimators on the X-Ray dataset. DML & S-Learner use pre-trained representations.

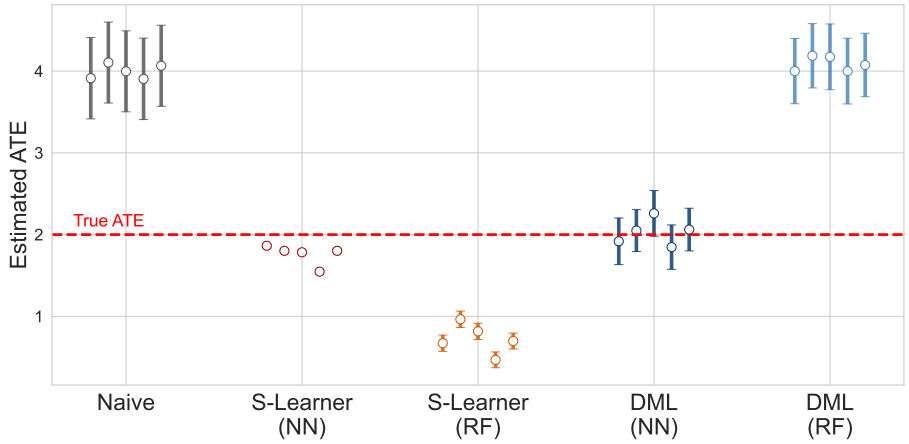

Figure 8. Complex Confounding: Comparison of ATE estimators on the IMDb dataset. DML & S-Learner use pre-trained representations.