Factor Informed Double Deep Learning For Average Treatment Effect Estimation

Jianqing Fan, Soham Jana, Sanjeev Kulkarni, and Qishuo Yin *

August 23, 2025

Abstract

We investigate the problem of estimating the average treatment effect (ATE) under a very general setup where the covariates can be high-dimensional, highly correlated, and can have sparse nonlinear effects on the propensity and outcome models. We present the use of a Double Deep Learning strategy for estimation, which involves combining recently developed factor-augmented deep learning-based estimators, FAST-NN, for both the response functions and propensity scores to achieve our goal. By using FAST-NN, our method can select variables that contribute to propensity and outcome models in a completely nonparametric and algorithmic manner and adaptively learn low-dimensional function structures through neural networks. Our proposed novel estimator, FIDDLE (Factor Informed Double Deep Learning Estimator), estimates ATE based on the framework of augmented inverse propensity weighting AIPW with the FAST-NN-based response and propensity estimates. FIDDLE consistently estimates ATE even under model misspecification, and is flexible to also allow for low-dimensional covariates. Our method achieves semiparametric efficiency under a very flexible family of propensity and outcome models. We present extensive numerical studies on synthetic and real datasets to support our theoretical guarantees and establish the advantages of our methods over other traditional choices, especially when the data dimension is large.

Keywords: Factor models, Deep learning, FAST-NN, AIPW, Average treatment effect.

1 Introduction

Estimating the average treatment effect (ATE) is a central task in causal inference, which has led to significant findings in many disciplines, including economics (Oreopoulos, 2006)

*J.F. and S.K. are with the Department of Operations Research and Financial Engineering and Department of Electric and Computer Engineering, Princeton University, Princeton, NJ, USA email: jqfan@princeton.edu, kulkarni@princeton.edu. Q.Y. is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA email: qy1448@princeton.edu. S.J. is with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA, (correspondence to: soham.jana@nd.edu). J.F.'s research is supported by NSF Grants DMS-2210833 and DMS-2412029 and the ONR Grant N00014-25-1-2317.

and political science (Aronow and Carnegie, 2013). ATE measures the expected difference in responses between units assigned to a treatment and those assigned to a control. In mathematical terms, given an experimental unit with covariate vector $\boldsymbol{x} \in \mathbb{R}^p$ and treatment assignment indicator T ($T = 1$ denotes that the unit receives treatment and $T = 0$ indicates that the unit is in the control group), the population outcome of y is given as

$$\mathbb{E}[y|T, \boldsymbol{x}] = \mu_0^*(\boldsymbol{x})\mathbf{1}_{\{T=0\}} + \mu_1^*(\boldsymbol{x})\mathbf{1}_{\{T=1\}}, \tag{1}$$

where μ_0^*, μ_1^* are unknown outcome functions and the ATE is given by $\mu = \mathbb{E}\left[\mu_1^*(\boldsymbol{x}) - \mu_0^*(\boldsymbol{x})\right]$. In practice, it is common to collect data on many variables deemed important in affecting policy outcomes; however, the treatment effect changes depending only on a handful of covariates, which are usually unknown to statisticians. Our method allows us to select important variables in a completely nonparametric and algorithmic manner. In addition, the covariates affecting the responses and treatments can be high-dimensional and highly correlated, and researchers might have incorrect assumptions about the data-generating models on the outcome and propensity functions. These challenges can be addressed by employing the recently developed neural network method FAST-NN (Fan and Gu, 2024, Factor-Augmented Sparse Throughput Neural Networks).

The ATE estimation problem becomes significantly challenging when the covariate dimension p grows with the sample size, and could be significantly larger than that. It is standard in such scenarios to assume that the output functionals are low-dimensional functions of $\boldsymbol{x}$. Factor modeling is historically considered to be an excellent choice for studying low-dimensional structures in the data and can also account for dependency among data variables (Fama and French, 2015; Fan et al., 2020). Applications exist in various important statistical problems, such as covariance estimation (Fan et al., 2008), dependence modeling (Oh and Patton, 2017), variable selection (Fan et al., 2020), tensor modeling (Zhou et al., 2025) and clustering (Tang et al., 2024). Given a p-dimensional random vector $\boldsymbol{x}$, the factor model assumes

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u}, \quad \boldsymbol{B} \in \mathbb{R}^{p\times r}, \boldsymbol{f} \in \mathbb{R}^r, \boldsymbol{u} \in \mathbb{R}^p, \quad r < p, \tag{2}$$

where the latent factor $\boldsymbol{f}$ and the independent idiosyncratic part $\boldsymbol{u}$ are unobservable random variables and the factor loading matrix $\boldsymbol{B} \in \mathbb{R}^{p\times r}$ is fixed but unknown. The loading matrix $\boldsymbol{B}$ indicates how the covariate vector $\boldsymbol{x}$ depends on the latent factor $\boldsymbol{f}$.

For modeling a function $m(\boldsymbol{x})$ such as the outcome or propensity functions using low-dimensional components, one often assumes the factor-augmented sparse throughput (FAST) model $m(\boldsymbol{x}) \triangleq m(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}})$, where $\mathcal{J}$ is the set of active coordinates of $\boldsymbol{u}$ and $r + |\mathcal{J}|$ (here and below, given any set $\mathcal{J}$, we denote its size by $|\mathcal{J}|$) is significantly smaller than the covariate dimension p. In particular, the factor structure enables us to estimate the effective components $\boldsymbol{u}_{\mathcal{J}}$ and $\boldsymbol{f}$ more accurately as p grows via neural networks (Fan and Gu, 2024) and the FAST-NN in that paper allows us to nonparametrically select the variable set $\mathcal{J}$. As noted in Fan and Gu (2024), the FAST model is very flexible and includes most existing models, from factor regression to sparse models in both parametric and nonparametric forms. Note that given $\boldsymbol{f}$, the FAST model $m(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}})$ can also be written as

$$m(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}}) = \widetilde{m}(\boldsymbol{f}, \boldsymbol{x}_{\mathcal{J}}) = \begin{cases} \widetilde{m}(\boldsymbol{x}_{\mathcal{J}}) & \text{specific case I: sparse regression} \\ \widetilde{m}(\boldsymbol{f}) & \text{specific case II: factor regression} \end{cases}$$

for another function $\widetilde{m}$ under the factor model (2). Therefore, it is a factor-augmented sparse nonparametric regression model. The model includes the nonparametric sparse regression model $\widetilde{m}(\boldsymbol{x}_{\mathcal{J}})$ and nonparametric factor regression model $\widetilde{m}(\boldsymbol{f})$ as two specific examples. It is applicable to the case where there is no factor structure ($r = 0$, covariates are weakly correlated) or low-dimensional setting ($\mathcal{J}$ = all variables).

In our current manuscript, we study the application of Deep Learning (DL) methods for estimating ATE. Efficient estimation of ATE often involves estimating both the responses (corresponding to treatment and control groups) and the propensity score, the conditional probability of receiving treatment given the covariates (Hirano et al., 2003), via the Augmented Inverse Propensity Weighting AIPW. Given an experimental unit with treatment assignment indicator T and covariate $\boldsymbol{x}$, its propensity score is defined as

$$\pi^*(\boldsymbol{x}) = \mathbb{E}\left[T|\boldsymbol{x}\right]. \tag{3}$$

Deep Learning is an extremely useful estimation tool when the structures of the target functions are unknown and possibly nonlinear. We term the strategy of using DL to learn both response and propensity component as the **Double Deep Learning** (DDL) technique, and our proposed estimator of the ATE will combine the benefits of such deep learning strategies. In the current literature, it is unclear whether DL methods are valuable tools for ATE estimation in the presence of strong covariance dependency and sparsity. For applying DL to handle strong covariates dependence, it is sensible to perform a denoising step to capture the independent components of the high-dimensional covariates and use the projected data to perform function estimation. However, the dependency structure is often misspecified (e.g., incorrect knowledge about r), leading to incorrect constructions of the denoising algorithms. It is known in the literature that model misspecifications can hurt propensity estimation significantly and lead to biased estimation of the ATE (Drake, 1993). From a practitioner's perspective, it is desirable to have efficient ATE estimators that can counter the practical issues mentioned above. On the other hand, the ATE estimation strategy should be flexible to tackle the case where the covariates are given to be low-dimensional, and we do not need to estimate the factor structure. In brief, we address the following:

Can using Double Deep Learning for responses and propensity estimation lead to efficient ATE estimation, both in the case of low-dimensional covariates and high-dimensional covariates with or without factor structures, even under model misspecifications?

In this paper we answer this question affirmatively. We show that for high-dimensional covariates, even when an over-specification $\bar{r}$ of the factor dimension r is provided (this includes the useful case that covariates are weakly correlated, but the factor model is used.), we can construct consistent factor augmented and deep learning based ATE estimators. Our results allow the covariate dimension to be significantly larger than the sample size, leading to resolving the problem in high dimensions. The versatility of our inference also provides the option to remove the factor modeling component when dealing with low-dimensional scenarios. Our ATE estimator is asymptotically Gaussian and semiparametrically efficient.

1.1 Our contributions

Methodological contribution

To our knowledge, our work is the first to introduce factor augmented deep learning techniques in the context of ATE estimation and analyze their theoretical guarantees. We propose a double deep learning type estimator called **FIDDLE**, that stands for the **F**actor **I**nformed **D**ouble **D**eep **L**earning **E**stimator. Suppose we have n observations of the response, treatment indicator, and covariate values, given by $\{(y_i, T_i, \boldsymbol{x}_i)\}_{i=1}^n$. Our algorithm consists of three major steps:

- **The pretraining factor augmentation step:** This step aims to estimate the factor components that determine the response and propensity (target) functions. We introduce a novel diversified projection matrix construction to perform factor augmentation. This is the only step where we use an independent pretraining sample of negligible size, vanishing compared to n. When the covariate dimension is low, we remove this factor augmentation step from our method, and the following steps remain the same.

- **The double deep learning step:** We estimate the outcome and propensity functions using factor-augmented deep neural networks. Specifically, we use a newly constructed diversified matrix to construct the FAST-NN (Fan and Gu, 2024) type estimators $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\pi}$.

- **The ATE estimation step:** We use the structure of the *Augmented Inverse Propensity Weighted (AIPW)* estimator (Glynn and Quinn, 2010), to combine the deep-learning-based FAST-NN estimators in the last step, and apply them to the same set of data without any sample splitting to construct the ATE estimator FIDDLE

$$\widehat{\mu}^{\mathsf{FIDDLE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{T_i y_i}{\widehat{\pi}(\boldsymbol{x}_i)} - \frac{(1-T_i)y_i}{1-\widehat{\pi}(\boldsymbol{x}_i)}\right) - (T_i - \widehat{\pi}(\boldsymbol{x}_i))\left(\frac{\widehat{\mu}_1(\boldsymbol{x}_i)}{\widehat{\pi}(\boldsymbol{x}_i)} + \frac{\widehat{\mu}_0(\boldsymbol{x}_i)}{1-\widehat{\pi}(\boldsymbol{x}_i)}\right)\right\}. \tag{4}$$

  The doubly robust structure of the AIPW estimator enables us to combine the consistency of the estimators for response and propensity to produce an efficient estimator of ATE.

Remark 1 (Comparison with previous algorithms)**.** The double deep learning-type strategies to combine deep-learning-based estimators are not new in the ATE estimation literature (Du et al., 2021; Farrell, 2015). However, such constructions often require knowledge of the exact low-dimensional structure of the target functions (Du et al., 2021, Condition 1) and often considers other dependency structures (Farrell, 2015) and sparsity as in the FAST model, which can significantly degrade the performance. In comparison, the construction of FIDDLE employs factor-augmented deep learning strategies to draw inference under dependency assumptions and learn low-dimensional target functions algorithmically. FIDDLE is also able to work with an overspecified number of factors, which significantly extends its applicability. In addition, if practical knowledge suggests that the data-generating response and propensity functions do not depend on the factor components, we apply our algorithm without the factor-augmentation step, thereby avoiding the pretraining step. Such scenarios often arise when the covariate dimension is small.

Contributions to theory and applications

Our paper is the first to show asymptotic normality of the FAST-NN based AIPW estimator for estimating the average treatment effect. We only use a negligible pretraining sample (compared to the sample size n) to construct the diversified projection matrix, and the rest of the deep learning-oriented estimator construction does not involve any further sample splitting. This makes it challenging to provide theoretical guarantees for the corresponding ATE estimator FIDDLE. In particular, our theoretical contributions are threefold:

- **Results for our new diversified projection matrix:** We show that our proposed diversified projection matrix adheres to the requirements in the literature (Fan and Liao, 2022) and that its singular values are large enough to produce strong estimation guarantees for the response and propensity functions. Our construction differs from the previous method in Fan and Gu (2024) where an incoherence condition on the sample variance matrix (Candes and Romberg, 2007; Abbe et al., 2020) is required in order to deduce boundedness of the diversified projection matrix (see Definition 7). In contrast, our new construction is simple and removes such a requirement.

- **Estimation guarantees for response and propensity functions:** We demonstrate that, under the assumption of a hierarchical composition model, factor-augmented neural network estimators can provide optimal guarantees even when the covariate dimension is high. Additionally, these intermediate steps help identify the active components of the covariates in the propensity and response functions, thereby providing interpretable results. This provides valuable information for policy-making, answering questions such as which covariate components influence the assignment of individual units to treatment and control groups, as well as their corresponding outcomes. Our theoretical guarantees for response function estimation using the factor-augmented neural network deviates from the existing work of Fan and Gu (2024), that studied the function estimation problem with a fixed dataset, as we need to use the random subsamples of control and treatment groups to estimate the response functions μ_0^*, μ_1^* respectively. We improve on the above work and provide a detailed analysis of controlling the estimation errors in such random setups.

- **Efficiency guarantees for FIDDLE:** We show that under some broad and relaxed smoothness assumptions on the outcome and propensity functions, FIDDLE enjoys asymptotic normality with semiparametric efficiency for ATE estimation. The analysis comes with significant challenges as we avoid sample splitting to perform the ATE estimation. The semiparametric efficiency is a desirable property in the literature for such tasks (Farrell, 2015; Fan et al., 2022), as this helps to construct confidence intervals for the unknown treatment effects.

- **Contributions in numerical studies:** We also present comparisons of our methods with many classical off-the-shelf ATE estimation techniques and demonstrate how a factor-oriented denoising step helps improve performance in high-dimensional scenarios. Our studies support our theoretical results and show that the accuracy of our estimators increases impressively as the covariate dimensions grow large, even beyond the sample size. The methods we compare against include other regularized neural

networks, Generative Adversarial Networks (GANs), and Causal Forest, among others. In terms of studying semi-synthetic data, we use the CIFAR-10 dataset (Canadian Institute For Advanced Research), which demonstrates the excellent performance of FIDDLE over other state-of-the-art methods for ATE estimation, particularly as the dimensionality of covariates $\boldsymbol{x}$ or sample size increases. Furthermore, we apply FIDDLE and benchmark methods to a real-world dataset from the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) to evaluate the causal effects of different bariatric surgery procedures on weight loss after 30 days of surgery.

Remark 2 (Comparison with similar existing results)**.** To establish the asymptotic distribution of FIDDLE, we apply a proof technique based on the concept of bracketing integral (Vaart and Wellner, 2023) to control the randomness of the estimators without sample splitting, which was inspired by the work of Fan et al. (2022). Still, the analysis differs significantly as we work with (a) an AIPW-type estimator instead of the IPW and (b) deep neural networks instead of the standard non-parametric structural assumption. The use of deep neural networks makes our work model agnostic and algorithmic, providing a more general guarantee. In addition, it is known in the literature (Glynn and Quinn, 2010) that the AIPW estimator obtains ATE estimators with lower variance compared to the IPW estimator, which helps our cause as well. We also establish that even in the presence of the factor structure, which leads to high correlation in the covariates, we can achieve the above results. In particular, our guarantees excel when the covariate dimension is significantly large. The randomness of subsamples $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n],T_i=1}$ and $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n],T_i=0}$ also contribute to the technical proofs.

1.2 Related works

Factor models play a crucial role in uncovering low-dimensional latent structures. Foundational contributions include (Chamberlain and Rothschild, 1982) and (Bai, 2003), which established identification and inference under general factor structures. We learn latent factors based on Diversified Projections (DP) (Fan and Liao, 2022), which allows for low-sample size and purposeful overestimation of latent factors for robustness. Our estimator incorporates DP to recover the shared latent structure and improve both treatment and outcome estimation.

Recent developments in machine learning have further enriched the landscape of causal inference, particularly in high-dimensional or nonlinear regimes. Double Machine Learning (DML) (Chernozhukov et al., 2018) formalizes orthogonalization and sample splitting for inference under ML-based function estimation. Causal Forests (Wager and Athey, 2018) adapt random forests to estimate conditional average treatment effects (CATE) using specialized split criteria. GANITE (Yoon et al., 2018) uses generative adversarial networks to learn counterfactual outcomes and derive individualized treatment effects. Recent methodological reviews (Hoffmann, 2024; Brand et al., 2023) provide comprehensive evaluations of these approaches. Furthermore, recent work on Calibrated Debiased Machine Learning (C-DML) (van der Laan et al., 2024) introduces novel doubly robust estimators that maintain asymptotic linearity even under misspecification. While these methods offer flexibility and

strong empirical performance, many do not explicitly account for latent factor structure in the covariates, and several are sensitive to model misspecification due to reliance on either outcome or treatment models alone. In contrast, our method integrates the strengths of factor structural modeling and modern machine learning by combining factor-based learning, neural network estimation, and the AIPW framework into a unified pipeline.

Deep neural networks (DNNs) (LeCun et al., 2015) have shown state-of-the-art performance in high-dimensional learning tasks and can recover low-dimensional structure (Mousavi et al., 2015; Chen et al., 2025). Recent studies (Yarotsky, 2017; Kohler and Langer, 2021) provide non-asymptotic guarantees across function classes. In nonparametric regression, DNNs mitigate the curse of dimensionality (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Fan et al., 2024; Bhattacharya et al., 2024) with the property of adaptively and algorithmically learning low-dimensional structure. Our method leverages the FAST-NN architecture proposed by Fan and Gu (2024), which is designed to adaptively capture sparse and dense components of the covariate space. This flexibility makes it particularly suitable for estimating both propensity scores and outcomes in AIPW estimation. Furthermore, Farrell et al. (2021) offers theoretical support for using DNNs in semiparametric estimation without sample splitting, aligning with our unified approach for efficient estimation of ATE.

1.3 Organization of the manuscript

The remainder of the paper is organized as follows. Section 2 introduces the setup, notation, and structural definitions of the problem that underpin our framework. Section 3 introduces our proposed estimator FIDDLE, and its components. Section 4 formally describes the model assumptions and the theoretical guarantees for our estimator, including consistency and convergence rates. Section 5 presents simulation studies that compare the performance of our method with existing benchmarks under various simulated and real datasets. Additional supporting results are provided in the appendix.

2 Preparation

We build our model using a fully connected deep neural network with ReLU activation $\bar{\sigma}(\cdot) = \max\{\cdot, 0\}$ similar to Fan and Gu (2024). Before presenting our methodology and results, we provide definitions that we will rely on throughout the manuscript.

Definition 1 (Deep ReLU Networks)**.** Let L be any positive integer and $\boldsymbol{d} = (d_1, ..., d_{L+1}) \in \mathbb{N}^{L+1}$. A deep ReLU network $g : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$ is given as the form

$$g(\boldsymbol{x}) = \mathcal{L}_{L+1} \circ \bar{\sigma} \circ \mathcal{L}_L \circ \bar{\sigma} \circ \cdots \circ \mathcal{L}_2 \circ \bar{\sigma} \circ \mathcal{L}_1(\boldsymbol{x}), \tag{5}$$

where $\mathcal{L}_\ell(\boldsymbol{z}) = \boldsymbol{W}_\ell \boldsymbol{z} + \boldsymbol{b}_\ell$ is a linear transformation with the weight parameters $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \boldsymbol{b}_\ell \in \mathbb{R}^{d_\ell}$, and $\bar{\sigma} : \mathbb{R}^{d_\ell} \mapsto \mathbb{R}^{d_\ell}$ applies the ReLU activation function coordinatewise.

Definition 2 (Deep ReLU network class)**.** For any $L \in \mathbb{N}, \boldsymbol{d} \in \mathbb{N}^{L+1}, B, M \in \mathbb{R}^+ \cup \{\infty\}$, the deep ReLU network family $\mathcal{G}(L, \boldsymbol{d}, M, B)$ with truncation level M, depth L, width vector $\boldsymbol{d}$, and weight bound B is given as

$$\mathcal{G}(L, \boldsymbol{d}, M, B) = \{\mathrm{Tr}_M(g(\boldsymbol{x})) : g \text{ of form (5) with } \|\boldsymbol{W}_\ell\|_{\max} \le B, \|\boldsymbol{b}_\ell\|_{\max} \le B\},$$

where $\mathrm{Tr}_M(\cdot)$ is the coordinatewise truncation operator given by $[\mathrm{Tr}_M(\boldsymbol{z})]_i = \mathrm{sgn}(z_i)(|z_i| \wedge M)$ and $\|\cdot\|_{\max}$ denotes the suppremum norm of a vector. The class of deep ReLU networks with depth L and width N is given by the specific case $\boldsymbol{d} = (d_{in}, N, N, \ldots, N, d_{out})$, and we denote it throughout the text by $\mathcal{G}(L, d_{in}, d_{out}, N, M, B)$.

We will use the following class of hierarchical composition functions to model μ_0^*, μ_1^*, π^*.

Definition 3 ((β, C)-smooth functions)**.** A d-variate function f is called (β, C)-smooth for $\beta, C > 0$ if the following is satisfied. Decompose β into integer part $r \geq 0$ and fraction part $0 < s < 1$. Then given every non-negative sequence $\alpha \in \mathbb{N}^d$ with $\sum_{j=1}^d \alpha_j = r$, the partial derivative $(\partial f)/(\partial x_1^{\alpha_1} \ldots x_d^{\alpha_d})$ exists, and $\left| \frac{\partial^r f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(\boldsymbol{x}) - \frac{\partial^r f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(\boldsymbol{z}) \right| \leq C \|\boldsymbol{x} - \boldsymbol{z}\|_2^s$.

Definition 4 (Hierarchical composition of smooth functions (Kohler and Langer, 2021; Fan and Gu, 2024))**.** Fix a constant $C > 0$. Let $\mathcal{H}(d, l, \mathcal{P})$ denote the class of l-depth and d-variate hierarchical composition of (β, C)-smooth functions for (β, t) in a set $\mathcal{P}$ with $\mathcal{P} \subset [1, \infty) \times \mathbb{N}^+, \sup_{(\beta,t)\in\mathcal{P}}(\beta \vee t) < \infty$

- $(l = 1)$ We have the set of all t-variate functions with (β, C) smoothness

$$\mathcal{H}(d, 1, \mathcal{P}) = \big\{ h : \mathbb{R}^d \mapsto \mathbb{R} : h(\boldsymbol{x}) = g(\boldsymbol{x}_{\mathcal{J}}), \text{ where } g : \mathbb{R}^t \mapsto \mathbb{R} \text{ is } (\beta, C)\text{-smooth for some } (\beta, t) \in \mathcal{P} \text{ and } \mathcal{J} \in [d], |\mathcal{J}| = t \big\}$$

- $(l \geq 2)$ We recursively define $\mathcal{H}(d, l, \mathcal{P})$ as

$$\mathcal{H}(d, l, \mathcal{P}) = \big\{ h : \mathbb{R}^d \mapsto \mathbb{R} : h(\boldsymbol{x}) = g(f_1(\boldsymbol{x}), \ldots, f_t(\boldsymbol{x})), \text{ where } g : \mathbb{R}^t \mapsto \mathbb{R} \text{ is } (\beta, C)\text{-smooth for some } (\beta, t) \in \mathcal{P} \text{ and } f_i \in \mathcal{H}(d, l-1, \mathcal{P}), i \in [t] \big\}$$

Basically, $\mathcal{H}(d, l, \mathcal{P})$ consists of the l time compositions of t-variate functions of (β, C) smoothness for any $(t, \beta) \in \mathcal{P}$. The accuracy of estimating $\mu_0^*, \mu_1^*, \pi^* \in \mathcal{H}(d, l, \mathcal{P})$ will be quantified by the parameter γ^* indicating the hardness of the above composition class.

Definition 5 (Hardness parameter of $\mathcal{H}(d, l, \mathcal{P})$)**.** Given any $\mathcal{P}$ satisfying (4) the hardness quantifier γ^* of the worst case error of approximating any function in $\mathcal{H}(d, l, \mathcal{P})$ by a deep ReLU network is quantified by $\gamma^* = \frac{\beta^*}{d^*}$ with $(\beta^*, d^*) = \mathrm{argmin}_{(\beta,t)\in\mathcal{P}} \frac{\beta}{t}$. In view of Kohler and Langer (2021), we restrict to the case where all the compositions has a smoothness parameter $\beta \geq 1$ to simplify the presentation. The parameter γ^* originates from the following approximation result of Fan and Gu (2024) (Theorem 4 therein), in which β/t reflects the dimension-adjusted degree of smoothness in a component of the hierachical composition model.

Lemma 1 (Approximating $\mathcal{H}(d, l, \mathcal{P})$ via deep ReLU Networks)**.** Let g be a d-variate, (β, C)-smooth function. There exists some universal constants c_1–c_5 depending only on d, β, C, such that for arbitrary $N \in \mathbb{N}^+ \setminus \{1\}$, there exists a deep ReLU network $g^\dagger \in \mathcal{G}(c_1, d, 1, c_2 N, \infty, c_3 N^{c_4})$ satisfying $\|g^\dagger - g\|_{\infty,[0,1]^d} \leq c_5 N^{-2\beta/d}$. Furthermore, if $g \in \mathcal{H}(d, l, \mathcal{P})$ with $\sup_{(\beta,t)\in\mathcal{P}}(\beta \vee t) < \infty$ and g is supported on $[-c_6, c_6]^d$ for some constant c_6. There also exists some universal constants c_7–c_{11} such that for arbitrary $N \in \mathbb{N}^+ \setminus \{1\}$, there exists a deep ReLU network $g^\dagger \in \mathcal{G}(c_7, d, 1, c_8 N, \infty, c_9 N^{c_{10}})$ satisfying $\|g^\dagger - g\|_{\infty,[-c_6,c_6]^d} \leq c_{11} N^{-2\gamma^*}$.

Definition 6 (Bracketing number and integral (Vaart and Wellner, 2023))**.** Given any distribution P, a function class $\mathcal{F}$ and a fraction $\epsilon > 0$, let $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ denote the ϵ-bracketing number of $\mathcal{F}$ under any norm $\|\cdot\|$, i.e., the minimum number of ϵ-brackets needed to cover $\mathcal{F}$ in the $\|\cdot\|$ distance. Denote the bracketing integral as $\widetilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)}\, d\epsilon$. We will pick a suitable norm later to fit our analysis.

Definition 7 (Diversified projection (DP) matrix (Fan and Liao, 2022; Fan and Gu, 2024))**.** Let $\bar{r} \geq r$ and c_1 be a universal positive constant. A $p \times \bar{r}$ matrix $\boldsymbol{W}$ is called a DP matrix if it satisfies (a) Boundedness: $\|\boldsymbol{W}\|_{\max} \leq c_1$, (b) Exogeneity: $\boldsymbol{W}$ is independent of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, (c) Significance: the matrix $\boldsymbol{H} = p^{-1}\boldsymbol{W}^\top \boldsymbol{B} \in \mathbb{R}^{\bar{r} \times r}$ satisfies $\nu_{\min}(\boldsymbol{H}) \gg p^{-1/2}$. Each column of $\boldsymbol{W}$ is called a diversified weight, and $\bar{r}$ is the number of diversified weights.

3 Methodology: FIDDLE

Our proposed estimator FIDDLE is a *double deep learning estimator* that relies on estimating both the outcome and propensity function using factor-augmented sparse throughput neural networks (FAST-NN) and then applying the AIPW estimator (4). To obtain the above deep learning-based estimators, we use the idea of the FAST estimator introduced in Fan and Gu (2024) that uses a LASSO (Tibshirani, 1996) type penalized loss function. We describe the estimator below. Let $\boldsymbol{W} \in \mathbb{R}^{\bar{r} \times p}$ be a given diversified projection matrix as defined in Definition 7 (a construction of $\boldsymbol{W}$ used in our work is outlined below). Suppose that we have the data $\{(y_i, T_i, \boldsymbol{x}_i)\}_{i=1}^n$. Then estimate the factor component of $\boldsymbol{x}_i$ as

$$\widetilde{\boldsymbol{f}}_i = \frac{1}{p}\boldsymbol{W}^\top \boldsymbol{x}_i, \quad i = 1, \ldots, n. \tag{6}$$

To describe our objective functions to construct the deep learning estimators, define the clipped-L_1 function $\psi_\tau(x)$ with the clipping threshold $\tau > 0$ as $\psi_\tau(x) = \frac{|x|}{\tau} \wedge 1$. Define $n_0 = \sum_{i=1}^n (1 - T_i)$, $\quad n_1 = \sum_{i=1}^n T_i$. Then the penalized mean squared error objectives $\widehat{R}_0, \widehat{R}_1, \widehat{R}_2$ corresponding to estimating μ_0^*, μ_1^*, π^* are defined as (the choice of the tuning parameters $\lambda_0, \lambda_1, \lambda_2, \tau_0, \tau_1, \tau_2, B, M, \bar{r}$ to guarantee our results will be described later)

$$\begin{aligned}
\widehat{R}_t(g, \boldsymbol{\Theta}) &= \frac{1}{n_t} \sum_{i=1, T_i = t}^n \left\{ y_i - g\left(\left[\widetilde{\boldsymbol{f}}_i, \mathrm{Tr}_M(\boldsymbol{\Theta}^\top \boldsymbol{x}_i) \right] \right) \right\}^2 + \lambda_t \sum_{i,j} \psi_{\tau_t}(\Theta_{i,j}), \quad t = 0, 1 \\
\widehat{R}_2(g, \boldsymbol{\Theta}) &= \frac{1}{n} \sum_{i=1}^n \left\{ T_i - g\left(\left[\widetilde{\boldsymbol{f}}_i, \mathrm{Tr}_M(\boldsymbol{\Theta}^\top \boldsymbol{x}_i) \right] \right) \right\}^2 + \lambda_2 \sum_{i,j} \psi_{\tau_2}(\Theta_{i,j})
\end{aligned} \tag{7}$$

where $[x, y]$ denotes the concatenation of two vectors $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$ to form a $(d_1 + d_2)$-dimensional vector, $\mathrm{Tr}_M(\cdot)$ is the truncation operator defined in Definition 2. Following Fan and Gu (2024), we optimize the above loss functions over $g \in \mathcal{G}(L, \bar{r} + N, 1, N, M, B)$, the ReLU deep network class given via Definition 2, and $\boldsymbol{\Theta} \in \mathbb{R}^{p \times N}$. Given any estimators $\widehat{g}, \widehat{\boldsymbol{\Theta}}$ originating from the above optimization, denote the corresponding FAST-NN estimator as

$$m^{\mathsf{FAST}}(\boldsymbol{x}; \boldsymbol{W}, \widehat{g}, \widehat{\boldsymbol{\Theta}}) = \widehat{g}\left(\left[\widetilde{\boldsymbol{f}}, \mathrm{Tr}_M(\widehat{\boldsymbol{\Theta}}^\top \boldsymbol{x}) \right] \right). \tag{8}$$

In light of the above, we are now ready to present our primary estimators.

3.1 Constructing a diversified projection matrix

To construct a diversified projection matrix $\boldsymbol{W}$ we first randomly pick $\{i_1, \ldots, i_m\} \subset [n]$ and consider the spectral decomposition of the corresponding variance covariance matrix $\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{x}_{i_j}\boldsymbol{x}_{i_j}^\top$ to obtain the eigenvalues $\{\widehat{\lambda}_j\}$ and eigenvectors $\{\widehat{\boldsymbol{v}}\}$ so that

$$\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{x}_{i_j}\boldsymbol{x}_{i_j}^\top = \sum_{j=1}^{m} \widehat{\lambda}_j \widehat{\boldsymbol{v}}_j \widehat{\boldsymbol{v}}_j^\top, \quad \lambda_1 \geq \lambda_2 \geq \ldots, \lambda_m \geq 0.$$

Then we propose the following novel construction of a diversified projection matrix

$$\boldsymbol{W} = \left[\sqrt{\widehat{\boldsymbol{\lambda}}_1} \cdot \widehat{\boldsymbol{v}}_1, \ldots, \sqrt{\widehat{\boldsymbol{\lambda}}_{\bar{r}}} \cdot \widehat{\boldsymbol{v}}_{\bar{r}}\right] \tag{9}$$

We will show later in Theorem 1 that $\boldsymbol{W}$ satisfies the requirements of a diversified projection matrix with a constant-order smallest singular value. For showing theoretical guarantees, we can use $m = n^{1-\gamma}$ for some constant $\gamma > 0$ and use $\{(y_i, T_i, \boldsymbol{x}_i) : i \in [n]/\{i_1, \ldots, i_m\}\}$ for ATE estimation. Therefore, the pretraining sample size m is negligible. For the convenience of notations, we will assume from this point onward an access to a $\boldsymbol{W}$ that is independent of the data, whose size is indexed by n. As m is negligible with respect to n, our theoretical results presented later will remain the same in view of the construction of $\boldsymbol{W}$ above.

Remark 3 (Comparison with the previous construction of DP matrix)**.** Fan and Gu (2024) uses the matrix $\widetilde{\boldsymbol{W}} = \sqrt{p}[\widehat{\boldsymbol{v}}_1, \ldots \widehat{\boldsymbol{v}}_{\bar{r}}]$ as their choice of the DP matrix. Notably, showing the boundedness requirement for $\widetilde{\boldsymbol{W}}$ as in Definition 7 is challenging, and requires the incoherence assumption in Abbe et al. (2020). For example, it is challenging to satisfy the boundedness requirement of Definition 7 for the submatrix $\sqrt{p}[\widehat{\boldsymbol{v}}_{r+1}, \ldots, \widehat{\boldsymbol{v}}_{\bar{r}}]$ of $\widetilde{\boldsymbol{W}}$, as the usual argument based on Weyl's Theorem (Chen et al., 2021, Lemma 2.2) provides significantly weaker controls on the magnitudes of eigenvalues $\widehat{\lambda}_{r+1}, \ldots, \widehat{\lambda}_{\bar{r}}$ when the data generating $\boldsymbol{B}$ matrix in (2) is of rank $r < \bar{r}$. Our modifications for constructing $\boldsymbol{W}$ directly guarantee the boundedness requirements and provide a more natural candidate for the DP matrix compared to $\widetilde{\boldsymbol{W}}$.

3.2 Response function estimation

To estimate the outcome functions corresponding to the control and treatment groups, we run two separate FAST-NN on the data $\{(y_i, \boldsymbol{x}_i) : T_i = 0, i \in [n]\}$ and $\{(y_i, \boldsymbol{x}_i) : T_i = 1, i \in [n]\}$ respectively, and define the FAST-NN estimators for estimating μ_0^*, μ_1^* as

$$\widehat{g}_i(\cdot), \widehat{\boldsymbol{\Theta}}_i \in \underset{\substack{\boldsymbol{\Theta} \in \mathbb{R}^{p \times N} \\ g \in \mathcal{G}(L, \bar{r}+N, 1, N, M, B)}}{\operatorname{argmin}} \widehat{R}_i(g, \boldsymbol{\Theta}), \quad \widehat{\mu}_i^{\mathsf{FAST}}(\cdot) = m^{\mathsf{FAST}}(\cdot; \boldsymbol{W}, \widehat{g}_i, \widehat{\boldsymbol{\Theta}}_i), \quad i = 0, 1. \tag{10}$$

3.3 Propensity function estimation

To estimate the propensity function π^* given in (3) we construct the FAST-NN estimator using the treatment indicators for the experimental units:

$$\widehat{g}_2(\cdot), \widehat{\boldsymbol{\Theta}}_2 \in \underset{\substack{\boldsymbol{\Theta}_2 \in \mathbb{R}^{p \times N} \\ g_2 \in \mathcal{G}(L, \bar{r}+N, 1, N, M, B)}}{\operatorname{argmin}} \widehat{R}_2(g, \boldsymbol{\Theta}), \quad \widetilde{\pi}(\cdot) = m^{\mathsf{FAST}}(\cdot; \boldsymbol{W}, \widehat{g}_2, \widehat{\boldsymbol{\Theta}}_2). \tag{11}$$

Note that to aid the theoretical analysis later on, we do not initially impose any restrictions to ensure that $m^{\mathsf{FAST}}(\cdot; \boldsymbol{W}, \widehat{g}_2, \widehat{\boldsymbol{\Theta}}_2)$ will lie within $[0,1]$, which is the range for the true propensity score. We perform a subsequent truncation step to obtain the final propensity estimator $\widehat{\pi}^{\mathsf{FAST}} = \max\left\{\alpha_n, \min\left\{\widetilde{\pi}^{\mathsf{FAST}}, 1-\alpha_n\right\}\right\}$ for a suitable $\alpha_n \in [0,1]$ to be chosen.

3.4 ATE estimation

Ultimately, the *double deep learning*-type ATE estimator FIDDLE is given as

$$\widehat{\mu}^{\mathsf{FIDDLE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{T_i y_i}{\widehat{\pi}^{\mathsf{FAST}}(\boldsymbol{x}_i)} - \frac{(1-T_i)y_i}{1-\widehat{\pi}^{\mathsf{FAST}}(\boldsymbol{x}_i)}\right) - (T_i - \widehat{\pi}^{\mathsf{FAST}}(\boldsymbol{x}_i))\left(\frac{\widehat{\mu}_1^{\mathsf{FAST}}(\boldsymbol{x}_i)}{\widehat{\pi}^{\mathsf{FAST}}(\boldsymbol{x}_i)} + \frac{\widehat{\mu}_0^{\mathsf{FAST}}(\boldsymbol{x}_i)}{1-\widehat{\pi}^{\mathsf{FAST}}(\boldsymbol{x}_i)}\right)\right\}. \tag{12}$$

Remark 4 (Modifying our algorithm for low-dimensional covariates)**.** When the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ have a low dimension, the factor augmentation step becomes redundant. In that case, we modify our algorithm by replacing $\widetilde{\boldsymbol{f}}_i$ in (6) with $\boldsymbol{x}_i$ for all $i \in \{1, \ldots, n\}$ and set $\boldsymbol{\Theta} = 0$ in (7). For simplicity of presentation, the reference to the FIDDLE method will also include such modifications. The proof of the theoretical results presented later accommodates this specific scenario in the case $r = 0$.

4 Theory

4.1 Model

We assume that data $\{(y_i, T_i, \boldsymbol{x}_i)\}_{i=1}^n$ are independently and identically distributed realizations of random variables $(y, T, \boldsymbol{x})$. The model that generates $(y, T, \boldsymbol{x})$ is given by

$$y(t) = \mu_t^*(\boldsymbol{x}) + \varepsilon(t), t \in \{0,1\}, \quad \mathbb{P}[T=1|\boldsymbol{x}] = 1 - \mathbb{P}[T=0|\boldsymbol{x}] = \pi^*(\boldsymbol{x}), \tag{13}$$

where $\varepsilon(0), \varepsilon(1)$ are mean zero random variables. The goal is to estimate $\mu = \mathbb{E}[\mu_1^*(\boldsymbol{x}) - \mu_0^*(\boldsymbol{x})]$. We assume the factor model $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{f} + \boldsymbol{u}$ as in (2), and model the functions μ_0^*, μ_1^*, π^* as

$$\mu_0^*(\boldsymbol{x}) = \mu_0^*(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}_0}), \mu_1^*(\boldsymbol{x}) = \mu_1^*(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}_1}), \pi^*(\boldsymbol{x}) = \pi^*(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}_2}), \mathcal{J}_0, \mathcal{J}_1, \mathcal{J}_2 \subset \{1, \ldots, p\}. \tag{14}$$

4.2 Assumptions

Assumption 1 (Low dimensionality)**.** $r, |\mathcal{J}_0|, |\mathcal{J}_1|, |\mathcal{J}_2|$ are at most finite constants.

Assumption 2 (Response and propensity bounds)**.** $\|\mu_0^*\|_\infty, \|\mu_1^*\|_\infty \leq M^*, \pi^* \in (\alpha_*, 1-\alpha_*)$ for constants $M^*, \alpha_* \in (0,1)$ and μ_0^*, μ_1^*, π^* are c_1-Lipschitz for some constants $c_1 > 0$. We further assume that $1 \leq M^* \leq M \leq c_2 M^*$ for some constant $c_2 > 1$, where M is the trimming parameter used in constructing the FAST-NN estimators in Section 3.

Assumption 3 (Unconfoundedness)**.** T is independent of $(y(0), y(1))$ given $\boldsymbol{x}$.

Assumption 4 (Boundedness)**.** For the factor model (2.2), there exist universal constants c_1 and b such that (a) the factor loading matrix satisfies $\|\boldsymbol{B}\|_{\max} \leq c_1$, (b) the factor component $\boldsymbol{f}$ of $\boldsymbol{x}$ is zero-mean and supported on $[-b, b]^r$, and (c) the idiosyncratic component $\boldsymbol{u}$ of $\boldsymbol{x}$ is zero-mean and supported on $[-b, b]^p$. This also implies that covariates are bounded in each coordinate and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in [-K, K]^p$ for some constant $K > 0$.

Assumption 5 (Weak dependence)**.** $\sum_{j,k\in[p], j\neq k} |\mathbb{E}[u_j u_k]| \leq c_1 \cdot p$ for some constant c_1.

Assumption 6 (Sub-Gaussian noise)**.** There exists a universal constant c_1 such that

$$\mathbb{P}\left[|\varepsilon(0)| \geq t | \boldsymbol{f}, \boldsymbol{u}\right], \mathbb{P}\left[|\varepsilon(1)| \geq t | \boldsymbol{f}, \boldsymbol{u}\right] \leq 2e^{-c_1 t^2}$$

for all the $t > 0$ almost surely.

Assumption 7 (Pervasiveness)**.** $\frac{p}{c_1} < \lambda_{\min}(\boldsymbol{B}^\top \boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{B}^\top \boldsymbol{B}) \leq c_1 p$ for a constant c_1.

Assumption 8 (Weak dependence between $\boldsymbol{f}$ and $\boldsymbol{u}$)**.** $\|\boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{f},\boldsymbol{u}}\|_F \leq c_1\sqrt{p}$ for a constant $c_1 > 0$, where $\boldsymbol{\Sigma}_{\boldsymbol{f},\boldsymbol{u}} = \mathbb{E}[\boldsymbol{f}\boldsymbol{u}^T] \in \mathbb{R}^{r\times p}$ is the covariance matrix between $\boldsymbol{f}$ and $\boldsymbol{u}$.

Assumption 9 (Function Class and ReLU Hyperparameters)**.** The true response and propensity functions satisfy $\mu_0^* \in \mathcal{H}(r + |\mathcal{J}_0|, l, \mathcal{P}), \mu_1^* \in \mathcal{H}(r + |\mathcal{J}_1|, l, \mathcal{P}), \pi^* \in \mathcal{H}(r + |\mathcal{J}_2|, l, \mathcal{P})$ for some bounded constants $r, l, |\mathcal{J}_0|, |\mathcal{J}_1|, |\mathcal{J}_2|$, and $\mathcal{P}$ has the dimension-adjusted smoothness $\gamma^* > \frac{1}{2} + c_0$ for a constant $c_0 > 0$, where γ^* is given by (5)). The following conditions on the deep ReLU network hyperparameters hold for constants $c_1, \ldots, c_6$ which only depend on l and $\mathcal{P}$ of $\{\mathcal{H}(r + |\mathcal{J}_j|, l, \mathcal{P})\}_{j=0,1,2}$.

$$\begin{gathered} c_1 \leq L \leq c_2, \quad c_3 \log n \leq \log B \leq c_4 \log n, \quad r \leq \bar{r} \lesssim c_3 \\ c_5 (n/\log n)^{\frac{1}{4\gamma^*+2}} \leq N \leq c_6 (n/\log n)^{\frac{1}{4\gamma^*+2}}. \end{gathered} \tag{15}$$

Remark 5 (Discussion of the assumptions)**.** Assumption 1 and Assumption 2 are standard in the deep learning literature (Kohler and Langer, 2021; Fan et al., 2024). Assumption 3 is also standard in the Causal Inference literature (Hirano et al., 2003), which ensures that there are no unmeasured confounders. Assumption 4 through Assumption 7 are also borrowed from the factor modeling literature (Fan and Gu, 2024). Assumption 8 subsumes the standard assumption of independence of $\boldsymbol{f}$ and $\boldsymbol{u}$ in the factor modeling literature, which is usually needed for identifiability of the model (Fan et al., 2021). Assumption 9 provides necessary constraints on the complexity of the outcome and propensity models to guarantee asymptotic normality of FIDDLE. This is also standard in the literature of nonparametric regressions via deep neural networks for achieving optimal mean squared errors (Fan and Gu, 2024) and the class of functions is indeed very broad, including additive models or more generally the compositions of low-dimensional functions.

4.3 Main results

We begin with a guarantee for our proposed diversified projection matrix, as shown in (9).

Theorem 1. The diversified projection matrix $\boldsymbol{W}$ constructed in (9) is a valid diversified projection under Assumption 4 through Assumption 8. In addition, there exist universal constants c_1, c_2, c_3 independent of $m, p, t, r, \bar{r}$ such that

$$c_1 - c_2\left(r\sqrt{\frac{\log p + t}{m}} + r^2\sqrt{\frac{\log r + t}{m}} + \frac{1}{\sqrt{p}}\right) \le \nu_{\min}(p^{-1}\boldsymbol{W}^\top \boldsymbol{B}) \le \nu_{\max}(p^{-1}\boldsymbol{W}^\top \boldsymbol{B}) \le c_3.$$

Note that due to our slight modification of the construction of $\boldsymbol{W}$, we do not require incoherent type of conditions. We now present function estimation guarantees. For $\widehat{R}_0, \widehat{R}_1, \widehat{R}_2$ in (7) and optimization error δ_{opt}, define $\{(\widehat{g}_t, \widehat{\boldsymbol{\Theta}}_t)\}_{t=0}^2$ as

$$\widehat{R}_t\left(\widehat{g}_t, \widehat{\Theta}_t\right) \le \inf_{\substack{\boldsymbol{\Theta}\in\mathbb{R}^{p\times N} \\ g\in\mathcal{G}(L,\bar{r}+N,1,N,M,B)}} \widehat{R}_t\left(g, \boldsymbol{\Theta}\right) + \delta_{\mathsf{opt}}, \quad t = 0, 1, 2. \tag{16}$$

Consider the FAST-NN estimators $\widehat{\mu}_0^{\mathsf{FAST}}, \widehat{\mu}_1^{\mathsf{FAST}}, \widehat{\pi}^{\mathsf{FAST}}$ defined as

$$\begin{aligned} &\widehat{\mu}_0^{\mathsf{FAST}}(\boldsymbol{x}) = m^{\mathsf{FAST}}(\boldsymbol{x}; \boldsymbol{W}, \widehat{g}_0, \widehat{\boldsymbol{\Theta}}_0), \quad \widehat{\mu}_1^{\mathsf{FAST}}(\boldsymbol{x}) = m^{\mathsf{FAST}}(\boldsymbol{x}; \boldsymbol{W}, \widehat{g}_1, \widehat{\boldsymbol{\Theta}}_1), \\ &\widehat{\pi}^{\mathsf{FAST}}(\boldsymbol{x}) = \max\left\{1/\log n, \min\left\{m^{\mathsf{FAST}}(\boldsymbol{x}; \boldsymbol{W}, \widehat{g}_2, \widehat{\boldsymbol{\Theta}}_2), 1 - 1/\log n\right\}\right\}, \end{aligned} \tag{17}$$

Given any function h and $j = 0, 1$, define $\|h\|_{n,j}^2 = \frac{1}{n_j}\sum_{i:T_i=j} h^2(\boldsymbol{x}_i)$, $\|h\|_n^2 = \frac{1}{n}\sum_{i=1}^n h^2(\boldsymbol{x}_i)$, and $\|h\|_2^2 = \int h^2(\boldsymbol{x}) dP(\boldsymbol{x})$, where P is the law of $\boldsymbol{x}$. Then we have the following result.

Theorem 2 (Oracle-type inequality for FAST-NN estimator)**.** Suppose that all the assumptions in Section 4.2 hold, except for Assumption 9, which is not used in this context. Consider the FAST model (14) and the FAST-NN estimator obtained by solving (10) and (11) with N, B large enough such that $N \ge 2(r + \max\{|\mathcal{J}_j| : j = 0, 1, 2\}), B \ge c_1 r \max\{|\mathcal{J}_j|\}_{j=0}^2$,

$$\begin{aligned} &\lambda_j \ge c_2\frac{\log(n_j p(N+r)) + L\log(BN)}{n_j}, \qquad \tau_j^{-1} \ge c_3(r+1)(BN)^{L+1}(N+\bar{r})pn_j, \quad j = 0, 1 \\ &\lambda_2 \ge c_2\frac{\log(np(N+r)) + L\log(BN)}{n}, \qquad \tau_2^{-1} \ge c_3(r+1)(BN)^{L+1}(N+\bar{r})pn \end{aligned}$$

for some constants c_1, c_2, c_3, and the number of diversified projections $\bar{r} \ge r$. Define

$$\begin{aligned} \delta_{i,a} &= \inf_{g\in\mathcal{G}(L,\bar{r}+N,1,N,M,B)} \|g - \mu_i^*\|_\infty^2, \quad i = 0, 1, \quad \delta_{2,a} = \inf_{g\in\mathcal{G}(L,\bar{r}+N,1,N,M,B)} \|g - \pi^*\|_\infty^2, \\ \delta_{i,s} &= \frac{(N^2L + N\bar{r})L\log(BNn)}{n} + \lambda_i|\mathcal{J}_i|, \quad \delta_{i,f} = \frac{|\mathcal{J}_i| r\cdot\bar{r}}{\nu_{\min}^2(\boldsymbol{H})\cdot p}, \quad i = 0, 1, 2. \end{aligned}$$

Then, with probability at least $1 - 3e^{-t}$, the following holds, for n large enough,

$$\|\widehat{\mu}_j^{\mathsf{FAST}} - \mu_j^*\|_2^2 + \|\widehat{\mu}_j^{\mathsf{FAST}} - \mu_j^*\|_{n,j}^2 \le \frac{c_4}{\alpha_*^2}\left\{\delta_{\mathsf{opt}} + \delta_{j,a} + \delta_{j,s} + \delta_{j,f} + \frac{t}{n}\right\}, \quad j = 0, 1 \tag{18}$$

$$\|\widehat{\pi}^{\mathsf{FAST}} - \pi^*\|_2^2 + \|\widehat{\pi}^{\mathsf{FAST}} - \pi^*\|_n^2 \le c_4\left\{\delta_{\mathsf{opt}} + \delta_{2,a} + \delta_{2,s} + \delta_{2,f} + \frac{t}{n}\right\}, \tag{19}$$

where c_4 is a constant and α_* is as given in Assumption 4.

The results (18),(19) give the mean squared errors for estimating the outcome functions and propensity score function in terms optimization error, approximation error, complexity of neural networks, penalization biases, as well as the estimation error of latent factors. The proof of the above results do not follow from the standard estimation guarantees for FAST-NN with fixed datasets as in Fan and Gu (2024), as our estimators $\widehat{\mu}_0^{\mathsf{FAST}}$ and $\widehat{\mu}_1^{\mathsf{FAST}}$ are constructed using the random sub-samples corresponding to the treatment group and control groups. The mean squared errors are measured with respected to the probability measure of the covariate $\boldsymbol{x}$ and its empirical version.

Next, we present asymptotic and efficiency guarantees of $\widehat{\mu}^{\mathsf{FIDDLE}}$ for estimating μ.

Theorem 3 (Asymptotic normality of FIDDLE)**.** Assume that all the assumptions in Section 4.2 hold and that $(n/\log n)^{\frac{1}{2}+c_1} < p < n^{100}$ for some $c_1 \in (0, \frac{1}{2})$ depending on c_0 in Assumption 9. Let $\widehat{\mu}_0 = \widehat{\mu}_0^{\mathsf{FAST}}, \widehat{\mu}_1 = \widehat{\mu}_1^{\mathsf{FAST}}, \widehat{\pi} = \widehat{\pi}^{\mathsf{FAST}}$ be as in (17) with $\delta_{\mathsf{opt}} < (n/\log n)^{-\frac{\gamma^*}{2\gamma^*+1}}$ and tuning parameters

$$\lambda_j = c_2 \frac{\log(n_j p(N+r)) + L\log(BN)}{n_j}, \qquad \tau_j^{-1} = c_3(r+1)(BN)^{L+1}(N+\bar{r})pn_j n, \quad j = 0, 1,$$
$$\lambda_2 = c_2 \frac{\log(np(N+r)) + L\log(BN)}{n}, \qquad \tau_2^{-1} = c_3(r+1)(BN)^{L+1}(N+\bar{r})pn^2,$$

where c_2, c_3 are constants as in Theorem 3. Then the ATE estimator FIDDLE (12) satisfies

$$\sqrt{n}(\widehat{\mu}^{\mathsf{FIDDLE}} - \mu) \to N(0, \sigma^2),$$
$$\sigma^2 = \mathbb{E}\left[(\mu_1^*(\boldsymbol{x}) - \mu_0^*(\boldsymbol{x}) - \mu)^2 + \frac{\mathrm{Var}[y(1)|\boldsymbol{x}]}{\pi^*(\boldsymbol{x})} + \frac{\mathrm{Var}[y(0)|\boldsymbol{x}]}{1-\pi^*(\boldsymbol{x})}\right],$$

and σ^2 attains the semiparametric efficiency bound (Hahn, 1998, Theorem 1). In addition, if $r = 0$, i.e., $\boldsymbol{x}$ does not have a factor component, then the result holds for any $p \le n^{100}$.

Remark 6 (Discussion of the results)**.** Theorem 3 establishes asymptotic normality and semiparametric efficiency of FIDDLE even when there is strong dependency among the covariates. We require only that the dimensionality-adjusted degree of smoothness γ^* satisfies $\gamma^* > 1/2$ in Assumption 9. Additionally, our proof shows that if μ_0^*, μ_1^*, π^* have different dimensionality-adjusted degree of smoothness $\gamma_0^*, \gamma_1^*, \gamma_2^*$ then we can establish the above result by only requiring $\gamma_0^*\gamma_2^* > \frac{1}{4}$ and $\gamma_1^*\gamma_2^* > \frac{1}{4}$, rather than having $\gamma_i^* > \frac{1}{2}, i = 0, 1, 2$. This establishes the doubly-robustness of FIDDLE, which relaxes the complexities of outcome and propensity models. Under the factor model assumptions, the requirement of a large covariate dimension p is essential to consistently estimating the factor components, see, e.g., (Fan and Gu, 2024, Theorem 3). The additional requirement of $p > (n/\log n)^{\frac{1}{2}+c_1}$ is imposed to guarantee the stronger result of asymptotic normality of the AIPW estimator, which can be removed in the absence of the factor component. This is captured in the second half of Theorem 3 with the case $r = 0$. In addition, the asymptotic normality and semiparametric efficiency hold even when our algorithm uses an overspecified number of factors $\bar{r}$. Our proof involves applying empirical process theory to establish a uniform error bound over the set of possible estimators within complex neural network classes.

5 Numerical studies

5.1 Candidate methods

We will compare the performance of FIDDLE with the following alternative approaches for estimating the average treatment effect. See Appendix D for implementation details.

- **Vanilla Neural Networks (Vanilla-NN):** A baseline variant that replaces FAST-NN with the trained fully connected neural networks, and with everything else the same.
- **Generative Adversarial Nets for Individualized Treatment Effects (GAN-ITE):** A GAN-based approach (Yoon et al., 2018) that first generates counterfactual outcomes via a dedicated generator, followed by training an individualized treatment effect (ITE) estimator. The ATE is estimated by the sample average of the estimated ITEs.
- **Double Robust Forest (DR):** A forest-based implementation of the doubly robust estimator (Bang and Robins, 2005). It jointly estimates the propensity scores and outcome models using random forests and computes the ATE via the AIPW estimator.
- **Double Machine Learning Forest (DML):** A double machine learning framework of Chernozhukov et al. (2018), which employs cross-fitting procedures to estimate nuisance parameters and eliminate regularization bias. It utilizes forest learners for both propensity score and response estimation, then applies double machine learning for ATE estimation.
- **Causal Forest (CF):** A nonparametric method based on generalized random forests (Wager and Athey, 2018; Athey et al., 2019). It estimates the responses by ensembling classification and regression trees (CART), and computes the ATE by their weighted difference.
- **Causal Forest on Latent Factors (Factor-CF):** A variant of the Causal Forest method applied exclusively to the latent factors $\boldsymbol{f}$ extracted from the covariates $\boldsymbol{x}$.
- **Oracle Inverse Propensity Weighting (Oracle-IPW):** Oracle benchmark using the ground truth propensity scores and corresponding IPW estimator Robins et al. (2000).
- **Oracle Augmented Inverse Propensity Weighting (Oracle-AIPW):** Oracle benchmark using the true response and propensity functions for AIPW (Robins et al., 1994).

5.2 Analysis with simulated data

We conduct two Monte Carlo experiments using synthetic datasets to evaluate the empirical performances. The first experiment benchmarks FIDDLE against a range of state-of-the-art machine learning estimators for the ATE, focusing on performance across varying covariate

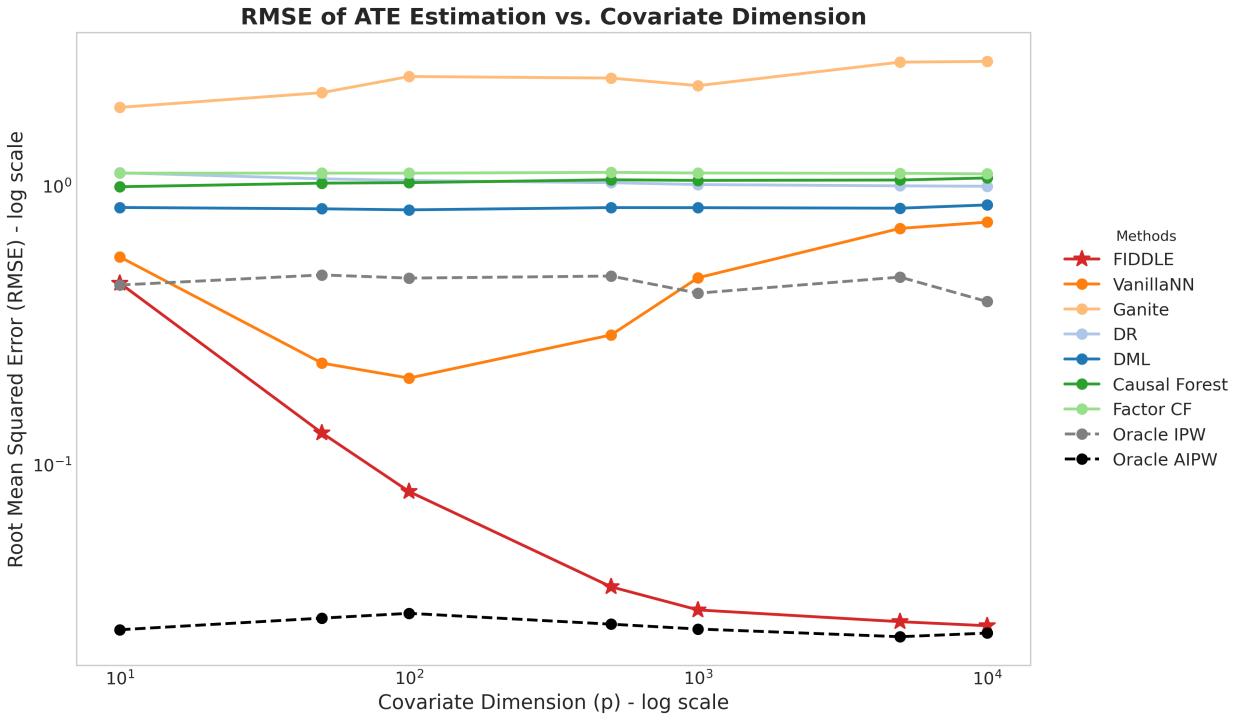

Figure 1: RMSE for different numbers of covariates by candidate ATE estimation methods.

dimensions. The second experiment investigates how FIDDLE behaves as the sample size increases. In both experiments, the data-generating process incorporates latent factor structures and nonlinear treatment and outcome models, enabling us to assess the robustness of estimators in high-dimensional, complex settings. Each experiment is replicated 100 times. Denote $\sigma(x) = \frac{e^x}{1+e^x}$, $\text{trun}\{z\} = 0.8z + 0.1$ for the rest of the paper.

Data Generating Process. We assume that the covariate vector $\boldsymbol{x}$ follows a factor model with a loading matrix $\boldsymbol{B} = (b_{ij})_{p\times r} \in \mathbb{R}^{p\times r}$, a vector of latent factors $\boldsymbol{f} = (f_i)_r \in \mathbb{R}^r$, and an idiosyncratic component $\boldsymbol{u} = (u_i)_p \in \mathbb{R}^p$, with $b_{ij} \overset{\text{i.i.d}}{\sim} \text{Unif}(-\sqrt{3}, \sqrt{3})$, $f_i, u_i \overset{\text{i.i.d}}{\sim} \text{Unif}(-1, 1)$. The number of factors is fixed at $r = 4$, and we evaluate performance for $p \in \{10, 50, 100, 500, 1000, 5000, 10000\}$. The sample size is set to $n = 5000$. The propensity and response models presented below incorporate nonlinear interactions of $\boldsymbol{f}$ and $\boldsymbol{u}$. We model $\pi^*(\boldsymbol{x})$ as $\pi^*(\boldsymbol{f}, \boldsymbol{u}) = \text{trun}\{\sigma(\sin(f_1) + \tan(f_2) + f_3 + f_4 + \sum_{j=1}^5 u_j),\}$ the outcome y as $y = \mu^*(\boldsymbol{f}, \boldsymbol{u}) + T\,\tau^*(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}}) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 1/4)$ independent of $\boldsymbol{f}, \boldsymbol{u}$ and

$$\mu^*(\boldsymbol{f}, \boldsymbol{u}) = 10 + f_1 + f_2 f_3 + \sin(f_4) + \log(5 + u_1 + u_2 u_3) + \tan(u_4) + u_5$$
$$\tau^*(\boldsymbol{f}, \boldsymbol{u}) = 5 + f_1 + f_2 + \sin(f_3) + \tan(f_4) + u_1 + u_2 + \sin(u_3 + u_4) + \tan(u_5).$$

The ground truth ATE with the specified model below is $\mathbb{E}[\tau^*(\boldsymbol{f}, \boldsymbol{u})] = 5$.

Results with varying covariate dimensions. Fig. 1 shows that the proposed estimator FIDDLE demonstrates a remarkable advantage over all other non-oracle methods in root mean squared error (RMSE), achieving consistently superior performance (see Table 1 for

Model	$p = 10$	$p = 50$	$p = 100$	$p = 500$	$p = 1000$	$p = 5000$	$p = 10000$
Oracle AIPW	0.0255	0.0280	0.0292	0.0267	0.0256	0.0240	0.0248
	(0.0021)	(0.0020)	(0.0020)	(0.0018)	(0.0019)	(0.0017)	(0.0016)
FIDDLE	0.4467	0.1295	0.0799	0.0363	0.0300	0.0272	0.0263
	(0.0258)	(0.0076)	(0.0055)	(0.0027)	(0.0021)	(0.0020)	(0.0019)
VanillaNN	0.5548	0.2308	0.2039	0.2912	0.4673	0.7028	0.7399
	(0.0248)	(0.0076)	(0.0069)	(0.0090)	(0.0108)	(0.0148)	(0.0154)
Ganite	1.9107	2.1582	2.4663	2.4348	2.2871	2.7776	2.7944
	(0.0176)	(0.0204)	(0.0174)	(0.0427)	(0.0355)	(0.0858)	(0.0765)
DR	1.1123	1.0588	1.0441	1.0252	1.0095	0.9990	0.9953
	(0.0148)	(0.0078)	(0.0081)	(0.0061)	(0.0078)	(0.0063)	(0.0061)
DML	0.8355	0.8263	0.8191	0.8346	0.8340	0.8305	0.8532
	(0.0198)	(0.0129)	(0.0127)	(0.0103)	(0.0112)	(0.0114)	(0.0090)
CF	0.9914	1.0214	1.0258	1.0501	1.0449	1.0482	1.0663
	(0.0186)	(0.0128)	(0.0112)	(0.0096)	(0.0121)	(0.0108)	(0.0090)
Factor-CF	1.1094	1.1096	1.1094	1.1165	1.1108	1.1080	1.1030
	(0.0049)	(0.0054)	(0.0050)	(0.0049)	(0.0059)	(0.0046)	(0.0052)
Oracle IPW	0.4401	0.4783	0.4660	0.4739	0.4114	0.4699	0.3838
	(0.0260)	(0.0300)	(0.0325)	(0.0353)	(0.0280)	(0.0276)	(0.0275)

Table 1: RMSE and standard error (in parentheses) of candidate ATE estimation methods across different covariate dimensions p with fixed sample size $n = 5000$ and 100 replications.

the exact values and standard errors (SE)). This is due to improved accuracy of FIDDLE in estimating latent factors and highlights our method's robustness and ability to leverage latent factor structure. Remarkably, when p is sufficiently large, FIDDLE attains performance comparable to that of the oracle-AIPW estimator. These empirical patterns align with the theoretical guarantees of FIDDLE. Moreover, FIDDLE achieves performance comparable to the oracle-IPW estimator when the covariate dimension is relatively low and surpasses it as p increases. In comparison, competing estimators—whether factor-based or neural network-based—fail to realize similar gains, with no appreciable decline in RMSE as p grows, underscoring their limitations in handling complex, high-dimensional covariate structures and nonparametric function learning.

Results for different sample sizes We run the experiments with the previous values of p and $n \in \{1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$. As shown in Fig. 2 (see Table 4 for the standard errors (SE) and Fig. 3 for a $\log - \log$ plot), FIDDLE demonstrates a pronounced and consistent reduction in both root mean square error (RMSE) as n increases, across all covariate dimensions p. This trend is particularly evident in high dimensions, where traditional methods often suffer from instability or bias. The rapid convergence of FIDDLE with increasing n highlights its statistical efficiency, affirming that the method scales well even with large dimensionality of the covariates.

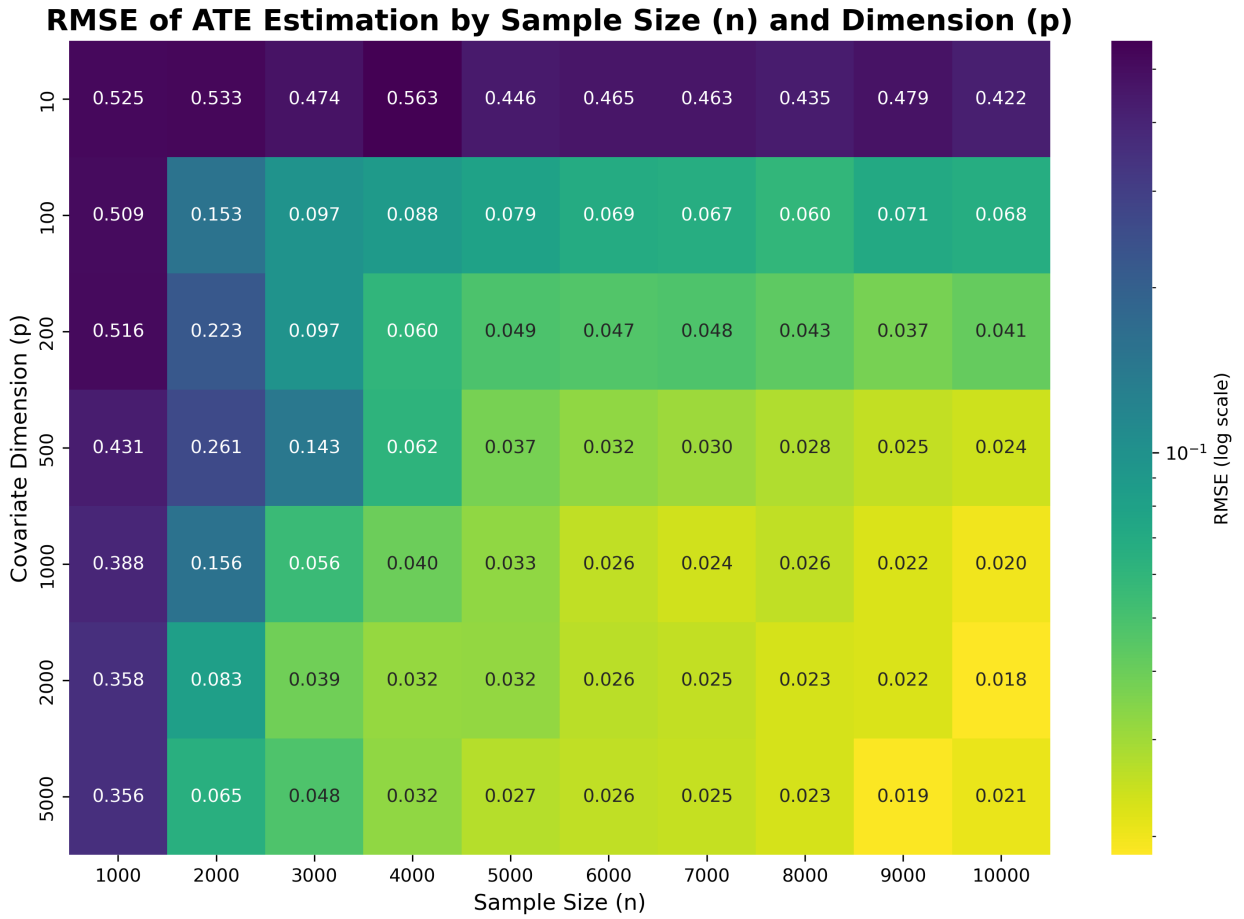

Figure 2: RMSE of FIDDLE for ATE estimation across different sample sizes and covariate dimensions. Dark regions and light regions indicate high and low RMSE, respectively.

5.3 Application to semi-synthetic image data

In this section, we demonstrate the practical utility of FIDDLE through a semi-synthetic image-based simulation derived from the Canadian Institute for Advanced Research (CIFAR-10) dataset. The CIFAR-10 dataset (Canadian Institute For Advanced Research) (Krizhevsky, 2009) is a widely used benchmark in machine learning and computer vision. It consists of 60,000 color images of size 32×32 pixels, categorized into 10 different classes. To facilitate analysis, we reshape the multidimensional array into a two-dimensional matrix $\boldsymbol{X}$, where $n = 60,000$ observations and $p = 32 \times 32 \times 3 = 3,072$ covariates. The covariate matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is normalized and decomposed into its factor structure $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{B}^T + \boldsymbol{U}$, where $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^T \in \mathbb{R}^{p \times r}$ represents the loading matrix, $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^T \in \mathbb{R}^{n \times r}$ denotes the latent factors, and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)^T \in \mathbb{R}^{n \times p}$ is the residual component with the number of factors $r = 4$. The unknown factors are estimated via a least-squares optimization algorithm: $\underset{\boldsymbol{B} \in \mathbb{R}^{p \times r}, \boldsymbol{F} \in \mathbb{R}^{n \times r}}{\text{minimize}} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{B}\boldsymbol{f}_i\|^2 = \|\boldsymbol{X} - \boldsymbol{F}\boldsymbol{B}^T\|_F^2$. We randomly sample $n' = 5,000$ observations from the entire dataset as $\boldsymbol{x}$ to replicate the semi-synthetic data-generating process. The factor $\widetilde{\boldsymbol{f}}$ is selected from the solution $\widetilde{\boldsymbol{F}}$,

corresponding to $\boldsymbol{x}$, and the residuals are computed as $\widetilde{\boldsymbol{u}} = \boldsymbol{x} - \widetilde{\boldsymbol{f}}\widetilde{\boldsymbol{B}}^T$. The treatment T is generated by Bernoulli sampling with probability $\pi^*(\boldsymbol{x})$, where $\pi^*(\boldsymbol{x})$ is modeled as $\pi^*(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{u}}) = \text{trun}\{\sigma(\sin\left(\widetilde{f}_1\right) + \sum_{i=2}^{4} \widetilde{f}_i + \sin(\widetilde{u}_1) + \sum_{j=2}^{5} \widetilde{u}_j)\}$ with $\sigma(\cdot), \text{trun}(\cdot)$ are as in Section 5.2. We model the outcomes as $y = \mu^*(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{u}}) + T\,\tau^*(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{u}}) + \varepsilon$, where

$$\mu^*(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{u}}) = 10 + \widetilde{f}_1 + \sin\left(\widetilde{f}_2\right) + \widetilde{f}_3\widetilde{f}_4 + \widetilde{u}_1(\widetilde{u}_2 + \sin(\widetilde{u}_3)) + \widetilde{u}_4 + \widetilde{u}_5,$$
$$\tau^*(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{u}}) = \widetilde{f}_1(\widetilde{f}_2 + 3) + \widetilde{f}_3 + \sin\left(\widetilde{f}_4\right) + \sin(\widetilde{u}_1) + \widetilde{u}_2 + \widetilde{u}_3\widetilde{u}_4\widetilde{u}_5,$$
$$\varepsilon \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1/4), \quad \text{independent of } \widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{u}}.$$

The ground truth ATE is empirically estimated for each simulation, and all candidate methods are implemented identically to Section 5.1.

Results. Table 2 reports the root mean squared error (RMSE) and standard error (SE) of each candidate method on the semi-synthetic CIFAR-10 dataset over 100 replications. FIDDLE achieves the best performance among all non-Oracle methods. The remarkable closeness of FIDDLE's performance to Oracle-AIPW supports our semiparametric efficiency claim. In contrast, the Vanilla Neural Network (Vanilla-NN) has a substantially higher MSE, and other methods perform considerably worse. FIDDLE also outperforms the Oracle-IPW estimator, highlighting the added stability gained through the doubly robust AIPW framework.

Oracle AIPW	FIDDLE	Vanilla NN	GANITE	DR	DML	CF	Factor CF	Oracle IPW
0.009	0.030	0.282	1.389	1.664	1.427	1.878	1.990	0.448
(0.001)	(0.003)	(0.012)	(0.032)	(0.007)	(0.007)	(0.007)	(0.006)	(0.030)

Table 2: RMSE and its standard error (in parentheses) of candidate ATE estimation methods on the semi-synthetic dataset based on CIFAR-10 over 100 replications.

5.4 Application with real dataset from bariatric surgery

We evaluate the causal effect of different bariatric surgery procedures on short-term weight loss using different candidate ATE estimators. Bariatric surgery remains the most effective treatment for morbid obesity, achieved by reducing stomach size or altering nutrient absorption pathways. Our analysis is based on the 2017 Participant Use File (PUF) from the Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) (American College of Surgeons, 2025), which collects high-quality, nationwide data on bariatric surgeries. After preprocessing, the dataset includes 174,013 patient records with 42 pretreatment covariates, surgery type, and 30-day BMI reduction as the outcome. We select Sleeve Gastrectomy (Sleeve)—the most widely performed procedure—as the control, and compare it against four common alternatives as the treatment: Roux-en-Y Gastric Bypass (RYGB), Adjustable Gastric Band (AGB), Biliopancreatic Diversion with Duodenal Switch (BPD/DS), and Single Anastomosis Duodeno-Ileal Bypass with Sleeve Gastrectomy (SADI-S). Table 3 reports the estimated ATE on 30-day BMI reduction, and associated 95% confidence intervals over 100 replications.

Surgery	RYGB	AGB	BPD/DS	SADI-S
FIDDLE	-0.0111	-1.0391	0.3364	-0.5020
	(-0.0116, -0.0106)	(-1.0516, -1.0266)	(0.3109, 0.3620)	(-0.5187, -0.4854)
Vanilla NN	-0.3814	-1.6807	-0.3670	-1.5438
	(-0.3880, -0.3749)	(-1.6913, -1.6701)	(-0.3860, -0.3479)	(-1.5573, -1.5304)
GANITE	-0.4511	-2.4651	-1.5018	-2.0452
	(-0.5161, -0.3861)	(-2.6590, -2.2712)	(-1.6552, -1.3484)	(-2.1686, -1.9218)
DR	-0.0310	-0.7343	0.0307	-0.5118
	(-0.0313, -0.0308)	(-0.7412, -0.7273)	(0.0305, 0.0309)	(-0.5178, -0.5058)
DML	-0.0154	-0.2403	0.6306	-0.5457
	(-0.0160, -0.0148)	(-0.2725, -0.2080)	(0.6060, 0.6552)	(-0.5498, -0.5416)
CF	-0.0225	-1.0804	0.2135	-0.6469
	(-0.0229, -0.0222)	(-1.0815, -1.0792)	(0.2114, 0.2157)	(-0.6488, -0.6451)
Factor-CF	-0.0569	-1.0924	0.1749	-0.7773
	(-0.0572, -0.0565)	(-1.0936, -1.0912)	(0.1730, 0.1769)	(-0.7787, -0.7759)

Table 3: Estimated ATE and 95% confidence intervals for 30-day BMI reduction by different procedures, compared with Sleeve as the control.

As shown in Table 3, FIDDLE yields relatively robust results that support the clinical understanding of the mechanism of each surgical procedure and the expected impact on short-term weight loss. RYGB shows a small negative ATE, consistent with equivalent short-term outcomes (Arterburn et al., 2018). AGB demonstrates a substantially negative ATE, reflecting its restrictive mechanism that produces slower weight loss requiring behavioral adaptation (Hady et al., 2012). BPD/DS yields a positive ATE, indicating a superior early reduction in BMI through its combined restrictive-malabsorptive approach (Hutter et al., 2013). SADI-S shows a moderately negative ATE, providing the first quantitative evidence that its staged design prioritizes long-term metabolic benefits over immediate weight loss enhancement (Pereira et al., 2024). Taken together, FIDDLE provides ATE estimates that support existing knowledge in the medical community– less aggressive procedures produce smaller short-term benefits, while more invasive techniques correspond to greater reductions in BMI.

Data availability statement: The data that support the findings of this study is provided in the 2017 Participant Use File (PUF) from the website of Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP) and the file is available upon request at their website (American College of Surgeons, 2025).

A Proof of Theorem 1

Define $\widehat{\boldsymbol{\Sigma}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top$ and consider the spectral decomposition of $\widehat{\boldsymbol{\Sigma}}$

$$\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^{p} \widehat{\lambda}_i \cdot \widehat{\boldsymbol{v}}_i \widehat{\boldsymbol{v}}_i^\top.$$

Using the coordinatewise boundedness of $\{\boldsymbol{x}_i\}_{i=1}^m$ we get that each coordinate of $\widehat{\boldsymbol{\Sigma}}$ is bounded. Then we use

$$\widehat{\boldsymbol{\Sigma}}_{jj} = \sum_{i=1}^{p} \widehat{\lambda}_i (\widehat{v}_{ij})^2, \quad j = 1, \dots p,$$

which implies $\sqrt{\widehat{\lambda}_i} \cdot \widehat{\boldsymbol{v}}_i$ has bounded coordinates for all $i = 1, \dots, p$. Hence, to show that $\boldsymbol{W} = \left[\sqrt{\widehat{\lambda}_1} \cdot \widehat{\boldsymbol{v}}_1, \dots, \sqrt{\widehat{\lambda}_{\bar{r}}} \cdot \widehat{\boldsymbol{v}}_{\bar{r}}\right]$ is a valid diversified projection matrix, it sufficies to ensure that it is independent of the data that we project using $\boldsymbol{W}$ and that the smallest singular value of $\frac{1}{p}\boldsymbol{W}^\top \boldsymbol{B}$ is large enough. As we use data splitting to construct the diversified projection matrix, and then use $\boldsymbol{W}$ to project the second half of the data, independence comes for free. Hence, it is only left to prove the singular value bounds mentioned in the theorem statement.

To this end, using Weyl's theorem (Chen et al., 2021, Lemma 2.2) it follows that

$$|\widehat{\lambda}_i - \lambda_i(\boldsymbol{B}\boldsymbol{B}^\top)| \le \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{B}\boldsymbol{B}^\top\|_F, \quad i = 1, \dots, p, \tag{20}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In view of (Fan and Gu, 2024, Lemma 5) we note that under Assumption 4, Assumption 5, Assumption 8

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{B}\boldsymbol{B}^\top\|_F \le c_1 p \left(r\sqrt{\frac{\log p + t}{m}} + r^2 \sqrt{\frac{\log r + t}{m}} + \frac{1}{\sqrt{p}} \right) \tag{21}$$

for a universal constant c_1. Note that from Assumption 7 we have that

$$\lambda_i(\boldsymbol{B}\boldsymbol{B}^\top) \in (\frac{p}{c_2}, c_2 p), \quad i = 1, \dots, r$$

for a large constant c_2. Hence, for $m \ge c_3 \log p$ with a large constant $c_3 > 0$ we combine (20) and (21) with the last display to get

$$\widehat{\lambda}_i \in (\frac{p}{c_4}, c_4 p), \quad i = 1, \dots, r$$

for a constant $c_4 > 0$. Next, denoting $\boldsymbol{W}_r = \left[\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{v}}_1, \dots, \sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{v}}_r\right], \widehat{\boldsymbol{V}}_r = [\widehat{\boldsymbol{v}}_1, \dots, \widehat{\boldsymbol{v}}_r]$ we get

$$\nu_{\min}(p^{-1}\boldsymbol{W}^\top \boldsymbol{B}) \ge \nu_{\min}(p^{-1}\boldsymbol{W}_r^\top \boldsymbol{B}) \ge \frac{1}{\sqrt{c_4}} \nu_{\min}(p^{-1/2}\widehat{\boldsymbol{V}}_r \boldsymbol{B}),$$

where the above inequalities followed using the Courant-Fischer minimax characterization of the smallest singular value $\nu_{\min}(\boldsymbol{A}) = \min_{\dim(\mathcal{U})=1} \max_{\boldsymbol{x}\in\mathcal{U}:\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$ (Dax, 2013, Theorem 1). The rest of the proof follows the strategy of (Fan and Gu, 2024, Proposition 1, Equation (F.23)), and is omitted here.

B Proof of Theorem 3

For the section below, we use the following notation. Let P denote the law of the covariate $\boldsymbol{x}$. Given any function $h = h(\boldsymbol{x})$ define

$$\mathbb{E}_P[h] = \int h \, dP(\boldsymbol{x}), \quad \|h\|_{L_2(P)} = \sqrt{\mathbb{E}_P[h^2]}, \quad \mathbb{E}_n[h] = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i).$$

We first provide a general result on the asymptotic normality of the AIPW estimator, given response and propensity estimators $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\pi}$

$$\widehat{\mu}^{\mathsf{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{T_i y_i}{\widehat{\pi}(\boldsymbol{x}_i)} - \frac{(1-T_i)y_i}{1-\widehat{\pi}(\boldsymbol{x}_i)}\right) - (T_i - \widehat{\pi}(\boldsymbol{x}_i))\left(\frac{\widehat{\mu}_1(\boldsymbol{x}_i)}{\widehat{\pi}(\boldsymbol{x}_i)} + \frac{\widehat{\mu}_0(\boldsymbol{x}_i)}{1-\widehat{\pi}(\boldsymbol{x}_i)}\right)\right\}. \tag{22}$$

The following result provides conditions on the complexity of possible function classes for $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\pi}$, that guarantees the asymptotic normality of $\widehat{\mu}^{\mathsf{AIPW}}$. The proof of Theorem 3 will rely on verifying these assumptions for the neural network classes of the FAST estimators.

Theorem 4. Consider estimators $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\pi} \in \mathcal{F}$ from function class $\mathcal{F}$, that are constructed from $\{(y_i, T_i, \boldsymbol{x}_i)\}_{i=1}^n$. Suppose that with probability $1-\xi$, the estimators satisfy

$$\mathbb{E}_P[(\widehat{\mu}_0 - \mu_0^*)^2] \le \delta_0^2, \quad \mathbb{E}_P\left[(\widehat{\mu}_1 - \mu_1^*)^2\right] \le \delta_1^2, \quad \mathbb{E}_P\left[(\widehat{\pi} - \pi^*)^2\right] \le \delta_2^2, \tag{23}$$

for the true response and propensity functions μ_0^*, μ_1^*, π^*, where the expectation is taken with respect to new $\boldsymbol{x}$ given the data. Assume that T_i, π_i are independent of y_i, conditional on $\boldsymbol{x}_i$. In addition, suppose that the following also holds true

(i) $\sup_{f\in\mathcal{F}} \|f\|_\infty \le \widetilde{B}$ for some constant $\widetilde{B} > 0$,

(ii) $\alpha < \pi^*, \widehat{\pi} < 1-\alpha$ for some $\alpha \in (0, 1/2)$,

(iii) $\frac{\sqrt{n}\delta_1\delta_2}{\alpha^2} + \frac{\sqrt{n}\delta_0\delta_2}{\alpha^2} + \sqrt{n}\xi/\alpha \to 0$, as $n \to \infty$,

(iv) $\lim_{n\to\infty} \frac{(\widetilde{J}_{[]}(\delta, \mathcal{F}, L_2(P)))^2}{\mathbf{1}_{\{\alpha\delta^2\sqrt{n}<1\}}\alpha\delta^2\sqrt{n} + \mathbf{1}_{\{\alpha\delta^2\sqrt{n}\ge 1\}}} = 0$ for $\delta \in \left\{\frac{\delta_0}{\alpha}, \frac{\delta_1}{\alpha}, \frac{\delta_2}{\alpha^2}\right\}$.

Then we have that $\sqrt{n}\left(\widehat{\mu}^{\mathsf{AIPW}} - \mu\right) \to N(0, \sigma^2)$, where

$$\sigma^2 = \mathbb{E}\left[(\mu_1^*(\boldsymbol{x}) - \mu_0^*(\boldsymbol{x}) - \mu)^2 + \frac{\mathrm{Var}[y(1)|\boldsymbol{x}]}{\pi^*(\boldsymbol{x})} + \frac{\mathrm{Var}[y(0)|\boldsymbol{x}]}{1-\pi^*(\boldsymbol{x})}\right].$$

Remark 7. The result also outlines the *doubly-robust* property of the AIPW estimator. Condition (iii) above shows that even when the estimation guarantees of the response functions are poor, i.e., δ_0, δ_1 converges to zero slowly, we can still guarantee the above asymptotic result as long as we have strong estimation guarantees for the propensity score, meaning $\delta_2\delta_0$ and $\delta_2\delta_1$ converges to zero at a rate $n^{-(1+c)}$ for some constant $c > 0$. In particular, if the target functions μ_0^*, μ_1^*, π^* belong to the classes $\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2$ with different dimensionality-adjusted degree of smoothness parameters $\gamma_0^*, \gamma_1^*, \gamma_2^*$ (as in Definition 5) respectively, and

the estimators $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\pi}$ achieves the nonparametric rates $\delta_i^2 = n^{-\frac{2\gamma_i^*}{2\gamma_i^*+1}}, i = 0, 1, 2$, then we can establish the above result by requiring $\gamma_i^* \gamma_2^* > \frac{1}{4} + c$, rather than individually requiring $\gamma_i^* > \frac{1}{2} + c, i = 0, 1, 2$, for some $c > 0$. The is essentially the doubly robustness property in terms of the hardness of the target function classes.

Proof of Theorem 4. Let $\mathcal{R}$ denote the event in which (23) holds. Then note that showing $\sqrt{n}(\widehat{\mu}^{\mathsf{AIPW}} - \mu)$ converges in distribution is equivalent to showing $\sqrt{n}(\widehat{\mu}^{\mathsf{AIPW}} - \mu)\mathbf{1}_{\{\mathcal{R}\}}$ converges in distribution. This is because $\sqrt{n}(\widehat{\mu}^{\mathsf{AIPW}} - \mu)$ is bounded by $O(\sqrt{n}/\alpha)$ (in view of the boundedness of the estimators, the response functions and the outputs), and the difference of the above terms satisfies

$$\sqrt{n}\mathbb{E}\left[|\widehat{\mu}^{\mathsf{AIPW}} - \mu| \cdot \mathbf{1}_{\{\mathcal{R}^c\}}\right] \leq O(\sqrt{n}/\alpha)\mathbb{P}[\mathcal{R}^c] = O(\sqrt{n}\xi/\alpha) \to 0.$$

Hence, we will assume without a loss of generality that the event $\mathcal{R}$ holds.

For simplicity of notations, we note the following definitions

$$\pi_i^* = \pi^*(\boldsymbol{x}_i), \quad \widehat{\pi}_i = \widehat{\pi}(\boldsymbol{x}_i).$$

We first note the following decomposition

$$\widehat{\mu}^{\mathsf{AIPW}} - \mu = \frac{1}{n}\sum_{i=1}^{n} S_i + R_0 + R_1 + R_2 + R_3,$$

where

$$S_i = \frac{T_i}{\pi_i^*}\left[y_i(1) - \mu_1^*(\boldsymbol{x}_i)\right] - \frac{1 - T_i}{1 - \pi_i^*}\left[y_i(0) - \mu_0^*(\boldsymbol{x}_i)\right] + \mu_1^*(\boldsymbol{x}_i) - \mu_0^*(\boldsymbol{x}_i) - \mu,$$

$$R_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i(y_i(1) - \mu_1^*(\boldsymbol{x}_i))}{\widehat{\pi}_i \pi_i^*}(\pi_i^* - \widehat{\pi}_i), \quad R_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - T_i)(y_i(0) - \mu_0^*(\boldsymbol{x}_i))}{(1 - \widehat{\pi}_i)(1 - \pi_i^*)}(\pi_i^* - \widehat{\pi}_i)$$

$$R_2 = \frac{1}{n\widehat{\pi}_i}\sum_{i=1}^{n}(\widehat{\pi}_i - T_i)(\widehat{\mu}_1(\boldsymbol{x}_i) - \mu_1^*(\boldsymbol{x}_i)), \quad R_3 = \frac{1}{n(1 - \widehat{\pi}_i)}\sum_{i=1}^{n}(\widehat{\pi}_i - T_i)(\widehat{\mu}_0(\boldsymbol{x}_i) - \mu_0^*(\boldsymbol{x}_i)).$$

We will show below that $\sqrt{n}R_i, i = 0, 1, 2, 3$, converges to zero in probability when the assumptions in the theorem statement are satisfied. Thus, the asymptotic normality of $\sqrt{n}(\widehat{\mu}^{\mathsf{AIPW}} - \mu)$ follows from the previous decomposition. We will use the following result.

Lemma 2. (Vaart and Wellner, 2023, Theorem 2.14.17') Let $\mathcal{F}$ be a class of measurable functions such that $\mathbb{E}_P[h^2] \leq \delta^2, \|h\|_\infty \leq Q$ for every $h \in \mathcal{F}$, and $\mathbb{G}_n[h] = \sqrt{n}(\mathbb{E}_n[h] - \mathbb{E}_P[h])$ based on samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{iid}{\sim} P$. Then there is a constant $c > 0$ such that

$$\mathbb{E}\left[\sup_{h \in \mathcal{F}} |\mathbb{G}_n[h]|\right] \leq c\widetilde{J}_{[]}\left(\delta, \mathcal{F}, L_2(P)\right)\left(1 + \frac{\widetilde{J}_{[]}\left(\delta, \mathcal{F}, L_2(P)\right)}{\delta^2\sqrt{n}}Q\right).$$

In view of the above theorem, we will proceed the proofs in the following way.

1. We first show $\sqrt{n}R_0$ converges to zero in probability and note that the convergence of $\sqrt{n}R_1$ can be shown similarly.

2. Next we show that $\sqrt{n}R_2$ converges to zero in probability if $\sqrt{n}\delta_1\delta_2$ converges to zero and note that we can use a similar strategy to show that $\sqrt{n}R_3$ converges to zero in probability whenever $\sqrt{n}\delta_0\delta_2$ converges to zero.

We now jump into the details of the above two steps.

1. Note that $\sqrt{n}R_0 = \mathbb{G}_n[h_{\widehat{\pi}}]$, where

$$h_\pi(\boldsymbol{x}) = \frac{T(y(1) - \mu_1^*(\boldsymbol{x}))}{\pi(\boldsymbol{x})\pi^*(\boldsymbol{x})}\left(\pi^*(\boldsymbol{x}) - \pi(\boldsymbol{x})\right)$$

and $\widehat{\pi} \in \mathcal{F}_2 \equiv \{h_\pi : \mathbb{E}_P\left[(\pi - \pi^*)^2\right] \leq \delta_2^2, \pi \in \mathcal{F}, \alpha < \pi < 1 - \alpha\}$. Then for each fixed $h_\pi \in \mathcal{F}_2$ (given by a fixed $\pi \in \mathcal{F}$)

   (a) $\mathbb{E}_P[h_\pi] = \mathbb{E}_P\left[\mathbb{E}_P\left[h_\pi|\,\boldsymbol{x}\right]\right] = 0$, as given $\boldsymbol{x}$, $y(1) - \mu_1^*(\boldsymbol{x})$ is uncorrelated with T (unconfoundedness)

   (b) $\mathbb{E}_P\left[h_\pi^2\right] \leq \frac{C\delta_2^2}{\alpha^4}$ for a constant C, as $y(1) - \mu_1^*(\boldsymbol{x})$ is uniformly bounded over all $\boldsymbol{x}$ and we have assumed that $\alpha < \pi^*, \pi < 1 - \alpha$.

   (c) $\|h_\pi\|_\infty \leq \frac{C}{\alpha}$ for a constant $C > 0$.

   Then using Theorem 2 with $\delta^2 = \frac{C\delta_2^2}{\alpha^4}, Q = \frac{C}{\alpha}$ we get that $\mathbb{E}\left[\sup_{h_\pi \in \mathcal{F}_2}|\mathbb{G}_n(h_\pi)|\right]$ converges to zero if the following are satisfied as $n \to \infty$

$$\widetilde{J}_{[]}\left(\delta, \mathcal{F}, L_2(P)\right) \to 0, \quad \frac{\widetilde{J}_{[]}\left(\delta, \mathcal{F}, L_2(P)\right)}{\alpha\delta^2\sqrt{n}} \to 0$$

   which is guaranteed by our assumptions. Consequently,

$$|\sqrt{n}R_0| = \left|\mathbb{G}_n[h_{\widehat{\pi}}]\right| \leq \sup_{h_\pi \in \mathcal{F}_2}|\mathbb{G}_n(h_\pi)| \xrightarrow{P} 0.$$

2. We next bound $\sqrt{n}R_2$. Denote

$$\widetilde{h}_{\mu_1,\pi}(\boldsymbol{x}) = \frac{(\pi^*(\boldsymbol{x}) - T)}{\pi(\boldsymbol{x})}(\mu_1(\boldsymbol{x}) - \mu_1^*(\boldsymbol{x})), \quad \breve{h}_{\mu_1,\pi}(\boldsymbol{x}) = \frac{(\pi(\boldsymbol{x}) - \pi^*(\boldsymbol{x}))}{\pi(\boldsymbol{x})}(\mu_1(\boldsymbol{x}) - \mu_1^*(\boldsymbol{x}))$$

   and note that $\sqrt{n}R_2 = \mathbb{G}_n\left[\widetilde{h}_{\widehat{\mu}_1,\widehat{\pi}}\right] + \mathbb{G}_n\left[\breve{h}_{\widehat{\mu}_1,\widehat{\pi}}\right]$. Denote

$$\mathcal{F}_{11} = \left\{\widetilde{h}_{\mu_1,\pi} : \mathbb{E}_P\left[(\mu_1 - \mu_1^*)^2\right] \leq \delta_1^2, \alpha < \pi < 1 - \alpha, \|\mu_1\|_\infty < \widetilde{B}, \mu_1, \pi \in \mathcal{F}\right\},$$

$$\mathcal{F}_{12} = \left\{\breve{h}_{\mu_1,\pi} : \mathbb{E}_P\left[(\mu_1 - \mu_1^*)^2\right] \leq \delta_1^2, \mathbb{E}_P\left[(\pi - \pi^*)^2\right] \leq \delta_2^2,\right.$$
$$\left.\|\mu_1\|_\infty < \widetilde{B}, \alpha < \pi < 1 - \alpha, \mu_1, \pi \in \mathcal{F}\right\}.$$

   Then, it suffices to separately show that

$$\lim_{n\to\infty}\sup_{\widetilde{h}_{\mu_1,\pi} \in \mathcal{F}_{11}}\left|\mathbb{G}_n\left[\widetilde{h}_{\mu_1,\pi}\right]\right| = 0, \quad \lim_{n\to\infty}\sup_{\breve{h}_{\mu_1,\pi} \in \mathcal{F}_{12}}\left|\mathbb{G}_n\left[\breve{h}_{\mu_1,\pi}\right]\right| = 0.$$

(a) We first show $\lim_{n\to\infty} \sup_{\widetilde{h}_{\mu_1} \in \mathcal{F}_{11}} \left|\mathbb{G}_n\left[\widetilde{h}_{\mu_1,\pi}\right]\right| = 0$. For each fixed $\widetilde{h}_{\mu_1,\pi} \in \mathcal{F}_{11}$ (given by fixed $\mu_1, \pi \in \mathcal{F}$)

i. $\mathbb{E}_P[\widetilde{h}_{\mu_1,\pi}] = \mathbb{E}_P\left[\mathbb{E}_P\left[\widetilde{h}_{\mu_1,\pi}\middle|\boldsymbol{x}\right]\right] = 0$, as $\mathbb{E}[T - \pi^*(\boldsymbol{x})|\boldsymbol{x}]$ has expectation zero.

ii. $\mathbb{E}_P\left[\widetilde{h}^2_{\mu_1,\pi}\right] \leq \frac{C\delta_1^2}{\alpha^2}$ for a constant C, as we assumed that $\alpha < \pi^*, \pi < 1-\alpha$.

iii. $\|\widetilde{h}_{\mu_1,\pi}\|_\infty \leq \frac{C}{\alpha}$ for a constant $C > 0$ as μ_1, μ^* are bounded.

Then we apply Theorem 2 with $\delta^2 = \frac{C\delta_1^2}{\alpha^2}, Q = \frac{C}{\alpha}$ to get $\mathbb{E}\left[\sup_{\widetilde{h}_{\mu_1,\pi}\in\mathcal{F}_1}\left|\mathbb{E}_n(\widetilde{h}_{\mu_1,\pi})\right|\right]$ converges to zero if the following are satisfied as $n \to \infty$

$$\widetilde{J}_{[]}\left(\delta, \mathcal{F}\times\mathcal{F}, L_2(P)\right) \to 0, \quad \frac{\widetilde{J}_{[]}\left(\delta, \mathcal{F}\times\mathcal{F}, L_2(P)\right)}{\alpha\delta^2\sqrt{n}} \to 0 \tag{24}$$

which is guaranteed by our assumptions as for any function classes $\mathcal{F}_1, \mathcal{F}_2$

$$\widetilde{J}_{[]}\left(\delta, \mathcal{F}_1\times\mathcal{F}_2, L_2(P)\right) \leq \widetilde{J}_{[]}\left(\delta, \mathcal{F}_1, L_2(P)\right) + \widetilde{J}_{[]}\left(\delta, \mathcal{F}_2, L_2(P)\right).$$

(b) We establish $\lim_{n\to\infty} \sup_{\breve{h}_{\mu_1,\pi}\in\mathcal{F}_{12}} \left|\mathbb{G}_n\left[\breve{h}_{\mu_1,\pi}\right]\right| = 0$ using the following subparts

$$\lim_{n\to\infty} \sup_{\breve{h}_{\mu_1,\pi}\in\mathcal{F}_{12}} \left|\mathbb{G}_n\left[\breve{h}_{\mu_1,\pi}\right]\right| = 0, \quad \lim_{n\to\infty} \sqrt{n}\cdot \sup_{\breve{h}_{\mu_1,\pi}\in\mathcal{F}_{12}} \left|\mathbb{E}_P\left[\breve{h}_{\mu_1,\pi}\right]\right| = 0 \tag{25}$$

For showing the first part, we note

i. $\mathbb{E}_P\left[\breve{h}^2_{\mu_1,\pi}\right] \leq \frac{C\delta_1^2}{\alpha^2}$ for a constant C, as we assumed that $\alpha < \pi^*, \pi < 1-\alpha$.

ii. $\|\breve{h}_{\mu_1,\pi}\|_\infty \leq \frac{C}{\alpha}$ for a constant $C > 0$ as μ_1, μ^* are bounded.

Then we apply Theorem 2 with $\delta^2 = \frac{C\delta_1^2}{\alpha^2}, Q = \frac{C}{\alpha}$ and (24) to get

$$\mathbb{E}_P\left[\sup_{\breve{h}_{\mu_1,\pi}\in\mathcal{F}_{12}}\left|\mathbb{G}_n\left[\breve{h}_{\mu_1,\pi}\right]\right|\right] \to 0.$$

For the second expression in (25) using the Cauchy-Schwarz inequality we get

$$\mathbb{E}_P[\breve{h}_{\mu_1,\pi}] \leq \frac{1}{\alpha}\sqrt{\mathbb{E}_P\left[(\pi-\pi^*)^2\right]\mathbb{E}_P\left[(\mu_1-\mu_1^*)^2\right]}$$

for every fixed μ_1, π such that $\breve{h}_{\mu_1,\pi} \in \mathcal{F}_{12}$. As all such μ_1, π satisfy

$$\mathbb{E}_P\left[(\mu_1-\mu_1^*)^2\right] \leq \delta_1^2, \mathbb{E}_P\left[(\pi-\pi^*)^2\right] \leq \delta_2^2,$$

we continue the last display to get $\sqrt{n}\sup_{\breve{h}_{\mu_1,\pi}\in\mathcal{F}_b}\left|\mathbb{E}\left[\breve{h}_{\mu_1,\pi}\right]\right| \leq \sqrt{n}\delta_1\delta_2/\alpha$, which converges to zero in view of our assumptions.

□

Proof of Theorem 3. We will verify the conditions in Theorem 4. Note that the conditions (i) and (ii) are satisfied in view of Assumption 4. Next we check condition (iii) in Theorem 4. In view of (Fan and Gu, 2024, Theorem 4) we note that there exists a constant c_1 such that

$$\delta_{i,a} \leq \sup_{g \in \mathcal{H}(r+|\mathcal{J}_i|,l,\mathcal{P})} \inf_{\widehat{g} \in \mathcal{G}(L,\bar{r}+N,1,N,M,B)} \|g - \widehat{g}\|_\infty^2 \leq c_1 (n/\log n)^{-\frac{2\gamma^*}{2\gamma^*+1}}, \quad i = 0,1,2.$$

In view of the definition of $\delta_{i,s}, i = 0,1,2$ in Theorem 2 we get that there exists a constant c_2 such that

$$\delta_{i,s} \leq c_2 (n/\log n)^{-\frac{2\gamma^*}{2\gamma^*+1}}, \quad i = 0,1,2.$$

On the other hand, if $r \geq 1$, we use the assumption $p > (n/\log n)^{\frac{1}{2}+c}$ for some constant $c > 0$. This implies for constants $c_3 > 0$

$$\delta_{i,f} \leq c_3 (n/\log n)^{-(\frac{1}{2}+c)}, \quad i = 0,1,2.$$

Note that from the definition of $\delta_{i,f}$ in Theorem 2, the above error becomes zero if $r = 0$. As we do not require $p > (n/\log n)^{\frac{1}{2}+c}$ for any other aspects of our proof, we can remove this requirement when $r = 0$ and our proof for this specific case follows the remainder of the arguments. Hence, assuming $\gamma^* > \frac{1}{2} + c_4$ for some constant $c_4 > 0$, we get from Theorem 2 that there is an event $\mathcal{E}$ with $\mathbb{P}[\mathcal{E}] \geq 1 - n^{-2}$ and constant $c_5 \in (0, \frac{1}{2})$ such that

$$\begin{aligned}
\mathbb{E}_P[(\widehat{\mu}_j^{\mathsf{FAST}} - \mu_j^*)^2] &\leq (n/\log n)^{-(\frac{1}{2}+c_5)}, \quad j = 0,1, \\
\mathbb{E}_P[(\widehat{\pi}^{\mathsf{FAST}} - \pi^*)^2] &\leq (n/\log n)^{-(\frac{1}{2}+c_5)}.
\end{aligned}$$

Hence, we get that on the event $\mathcal{E}$, our estimators satisfy the requirement (23) with $\delta_0^2 = \delta_1^2 = \delta_2^2 = (n/\log n)^{-(\frac{1}{2}+c_5)}$. Call this common value to be $\widetilde{\delta}^2$. Then, the requirement (iii) in Theorem 4 is satisfied. It remains to prove the condition (iv) in Theorem 4. In view of our choice of $\left\{\tau_k, \widehat{\mathbf{\Theta}}_k\right\}_{k=0}^2$ we first show that with a high probability

$$\sum_{i,j} \psi_{\tau_k}(\widehat{\Theta}_{k,i,j}) \leq c_6 (n/\log n)^{\frac{1}{2}-c_5}, \quad k = 0,1,2. \tag{26}$$

We prove the case for τ_1 as the other cases can be proven in a similar way. In view of (45), choosing $t = 2\log n$ we get that on an event $\mathcal{E}_1$ with $\mathbb{P}[\mathcal{E}_1] \geq 1 - \frac{1}{n^2}$, for a constant $\bar{c} > 0$

$$\begin{aligned}
\sum_{i,j} \psi_{\tau_1}(\widehat{\Theta}_{1,i,j}) \leq \bar{c} \Bigg(|\mathcal{J}_1| + \frac{1}{\lambda_1} \Bigg\{ (n/\log n)^{-\frac{2\gamma^*}{2\gamma^*+1}} + (N^2 L + N\bar{r}) \frac{L \log(BNn)}{n} \\
+ \frac{\log(np(N+\bar{r})) + L\log(BN)}{n} \Bigg\} \Bigg).
\end{aligned} \tag{27}$$

As our choice of λ_1 guarantees $\lambda_1 > \frac{\log(np)}{n}$, we get the desired result. Here we used $N^2 \asymp (n/\log n)^{\frac{1}{2\gamma^*+1}} < \sqrt{\frac{n}{\log n}}$ from Assumption 9. Now we are ready to verify the bracketing integral requirements in Theorem 4. We first note the following results for deep neural networks class that we use. A proof is provided later in this section.

Lemma 3. Consider the set $\mathcal{G}_{m,s}$ defined as

$$\mathcal{G}_{m,s} = \left\{ \mu_1 = m^{\mathsf{FAST}}\ (\cdot; \boldsymbol{W}, g, \boldsymbol{\Theta}) \in \mathcal{G}_m : \sum_{i,j} \psi_\tau(\Theta_{ij}) \le s, \|\boldsymbol{\Theta}\|_{\max} \le B \right\}. \tag{28}$$

and denote

$$A = LN^2 + (L+1)N + N\bar{r} + 1,$$
$$\widetilde{C} = (M \vee K\|\boldsymbol{W}\|_{\max})(L+1)B^L(N+1)^{L+1} + KB^L N^L(N+\bar{r})p.$$

Then for all $\delta > 0$

$$\widetilde{J}_{[]}(\delta, \mathcal{G}_{m,s}, L_2(P)) \le 8\sqrt{\log\left(B\widetilde{C}\right)}\left(\tau\log(1/\tau)\sqrt{A+Np} + \mathbf{1}_{\{\delta\ge\tau\}}\delta\log(1/\delta)\sqrt{(A+2s)}\right).$$

We now apply Lemma 3 with $\alpha = \frac{1}{\log n}, \delta = \frac{\widetilde{\delta}}{\alpha}, s = c_6(n/\log n)^{\frac{1}{2}-c_5}$ and $\log p \le (n/\log n)^{\frac{1}{2}-\widetilde{c}}$ for some constant $\widetilde{c} \in (0, \frac{1}{2})$ and obtain

$$(\log n)\frac{(\widetilde{J}_{[]}(\delta, \mathcal{G}_{m,s}, L_2(P)))^2}{\delta^2\sqrt{n}} \le c_8(\log n)^4 \log\left(B\widetilde{C}\right)\left\{\frac{1}{\sqrt{n}} + \frac{(A+2s)\log n}{\sqrt{n}}\right\} \to 0 \text{ as } n \to \infty,$$

and

$$\widetilde{J}_{[]}(\delta, \mathcal{G}_{m,s}, L_2(P)) \le c_9(\log n)^2\sqrt{\log\left(B\widetilde{C}\right)}\left\{\delta + \sqrt{\delta^2(A+2s)\log n}\right\} \to 0 \text{ as } n \to \infty.$$

where we used that $A \asymp N^2 \asymp (n/\log n)^{\frac{1}{2\gamma^*+1}} < \sqrt{\frac{n}{\log n}}$, given $\gamma^* > \frac{1}{2}$.

□

Proof of Lemma 3. The proof strategy is as follows. Consider the parameter space

$$\mathcal{U} = \left\{(\boldsymbol{\Theta}, \{\boldsymbol{W}_\ell, \boldsymbol{b}_\ell\}_{\ell=1}^{L+1}) \in [-B, B]^{Np+A} : m(\cdot; \boldsymbol{W}, g, \boldsymbol{\Theta}) \in \mathcal{G}_{m,s}\right\}.$$

We will first show a bound on $\log \mathcal{N}(\epsilon_1, \mathcal{U}, \|\cdot\|_\infty)$ separately for the case $\epsilon_1 > \tau$ and $\epsilon_1 \le \tau$. Then in view of (Vaart and Wellner, 2023, Theorem 2.7.17) we can use the bound

$$\mathcal{N}_{[]}(\epsilon_1, \mathcal{G}_{m,s}, L_2(\mathcal{P})) \le N\left(\frac{\epsilon_1}{2\widetilde{C}}, \mathcal{U}, \|\cdot\|_\infty\right), \tag{29}$$

where $\widetilde{C}$ is as defined in the result statement, and satisfies (Fan and Gu, 2024, Lemma 8)

$$\sup_{\boldsymbol{x}\in[-K,K]^p} |m(\boldsymbol{x}) - \breve{m}(\boldsymbol{x})| \le \widetilde{C}\|\boldsymbol{\theta}(m) - \boldsymbol{\theta}(\breve{m})\|_\infty, \quad m, \breve{m} \in \mathcal{G}_{m,s}, \boldsymbol{\theta}(m) = \{\boldsymbol{\Theta}, (\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)_{\ell=1}^{L+1}\}.$$

1. Consider the case $\epsilon_1 \ge \tau$: We will use the cover

$$\mathcal{U}(\epsilon_1) = \cup_{S\subset[N]\times[p]:|S|=s}\mathcal{U}(\epsilon_1, S)$$

where

$$\mathcal{U}(\epsilon_1, S) = \Bigg\{ [\boldsymbol{W}_\ell]_{i,j}, [\boldsymbol{b}_\ell]_j \in \left\{ -B + \epsilon_1, \cdots, -B + \epsilon_1 \cdot \left\lceil \frac{2B}{\epsilon_1} \right\rceil \right\}, \\ \boldsymbol{\Theta}_S \in \left\{ -B + \epsilon_1, \cdots, -B + \epsilon_1 \cdot \left\lceil \frac{2B}{\epsilon_1} \right\rceil \right\}^s, \boldsymbol{\Theta}_{S^c} = 0 \Bigg\}$$

Then we show that $\mathcal{U}(\epsilon_1)$ is a valid ϵ_1-cover of $\mathcal{U}$ in the $\|\cdot\|_\infty$ norm. Note that

$$\sum_{i,j} \mathbf{1}_{\{|\Theta_{i,j}| > \tau\}} \leq \sum_{i,j} \psi_\tau(\Theta_{ij}) \leq s.$$

The above implies, given any $T \in \mathcal{U}$, there is a set $S \subset [n] \times [p]$ with $|S| \leq s$ such that $\boldsymbol{\Theta}_{S^c}$ has all entries with absolute value bounded by τ. As $\epsilon_1 \geq \tau$, we can find a $\widetilde{T} \in \mathcal{U}(\epsilon_1, S)$ such that $\|\widetilde{T} - T\|_\infty \leq \epsilon_1$. Hence, $\mathcal{U}(\epsilon_1)$ gives us an ϵ_1-cover of $\mathcal{U}$. Note that the total number of parameters in $\{\boldsymbol{W}_\ell, \boldsymbol{b}_\ell\}_{\ell=1}^{L+1}$ is $N(N+\bar{r})+N+\sum_{\ell=2}^{L} N(N+1)+N+1$, which is defined as A in the lemma statement. Then, it is straightforward to check that $\mathcal{U}(\epsilon_1)$ has at most $\left\lceil \frac{2B}{\epsilon_1} \right\rceil^{s+A}$ entries. Next, note that the number of choices for $S \subset [N] \times [p]$ such that $|S| = s$ is $\binom{Np}{s} \leq (Np)^s$. Hence

$$\mathcal{N}(\epsilon_1, \mathcal{U}, \|\cdot\|_\infty) \leq \left\lceil \frac{2B}{\epsilon_1} \right\rceil^{s+A} (Np)^s \leq \left\lceil \frac{2B}{\epsilon_1} \right\rceil^{s+A} (\widetilde{C})^s, \quad \epsilon_1 \geq \tau.$$

2. Next, we consider $\epsilon_1 < \tau$:

$$\mathcal{U}(\epsilon_1) = \left\{ [\boldsymbol{W}_\ell]_{i,j}, [\boldsymbol{b}_\ell]_j, [\boldsymbol{\Theta}]_{i,j} \in \left\{ -B + \epsilon_1, \cdots, -B + \epsilon_1 \cdot \left\lceil \frac{2B}{\epsilon_1} \right\rceil \right\} \right\}$$

It is straightforward to show that $\mathcal{U}(\epsilon_1)$ gives an ϵ_1-cover of $\mathcal{U}$. As the total number of parameters in $(\boldsymbol{\Theta}, \{\boldsymbol{W}_\ell, \boldsymbol{b}_\ell\}_{\ell=1}^{L+1})$ is $Np+N(N+\bar{r})+N+\sum_{\ell=2}^{L} N(N+1)+N+1 = Np+A$, $\mathcal{U}(\epsilon_1)$ has at most $\left\lceil \frac{2B}{\epsilon_1} \right\rceil^{Np+A}$ entries. Hence

$$\mathcal{N}(\epsilon_1, \mathcal{U}, \|\cdot\|_\infty) \leq \left\lceil \frac{2B}{\epsilon_1} \right\rceil^{Np+A}, \quad \epsilon_1 < \tau.$$

Combining the above, in view of (29) we get that

$$\log\left(\mathcal{N}_{[]}(\epsilon_1, \mathcal{G}_{m,s}, L_2(\mathcal{P}))\right) \leq \left\{ \mathbf{1}_{\{\epsilon_1 < \tau\}} (A + Np) + \mathbf{1}_{\{\epsilon_1 \geq \tau\}} (A + 2s) \right\} \log\left(\frac{1 + 4B\widetilde{C}}{\epsilon_1} \right).$$

As $B\widetilde{C}$ is large, for all $\epsilon_1 < 1$, we can use the inequality $\log\left(\frac{1+4B\widetilde{C}}{\epsilon_1}\right) \leq 4\log\left(B\widetilde{C}\right)\log(1/\epsilon_1)$. Noting the definition of the bracketing integral in Definition 6 we get

$$\widetilde{J}_{[]}(\delta, \mathcal{G}_{m,s}, L_2(P)) \\ \leq 4\sqrt{\log\left(B\widetilde{C}\right)} \left(\mathbf{1}_{\{\delta < \tau\}} \sqrt{A + Np} \int_0^\tau \sqrt{\log(1/\varepsilon)} d\varepsilon + \mathbf{1}_{\{\delta \geq \tau\}} \sqrt{(A + 2s)} \int_0^\delta \sqrt{\log(1/\varepsilon)} d\varepsilon \right). \tag{30}$$

Hence, it suffices to bound $\int_0^\delta \sqrt{\log(1/\epsilon)}\, d\epsilon$. Using a change of variable $\log(1/\epsilon) = z^2$ the integral can be transformed into $2\int_\nu^\infty z^2 e^{-z^2} dz$ with $\nu = \sqrt{\log(1/\delta)}$. To study the last integral we will use $\int_\nu^\infty e^{-(az)^2} dz$. Note that for any $a \neq 0$ we have

$$\int_\nu^\infty e^{-(az)^2} dz = \frac{1}{a}\int_{a\nu\sqrt{2}}^\infty 2^{-\frac{y^2}{2}} dy = \frac{\sqrt{2\pi}}{a}\left(1 - \Phi(a\nu\sqrt{2})\right).$$

Differentiating the above display with respect to a we get (ϕ is the standard Gaussian density)

$$2a\int_\nu^\infty z^2 e^{-(az)^2} dz = \frac{\sqrt{2\pi}}{a^2}\left(1 - \Phi(a\nu\sqrt{2})\right) + \frac{\sqrt{2\pi}}{a}\phi(a\nu\sqrt{2})\nu\sqrt{2}.$$

Plugging in $a = 1$ and using Mill's ratio bound $1 - \Phi(x) \leq \frac{\phi(x)}{x}$ for $x > 0$ we get

$$2\int_\nu^\infty z^2 e^{-z^2} dz \leq \frac{\sqrt{2\pi}\phi(\nu\sqrt{2})}{\nu\sqrt{2}} + \sqrt{2\pi}\phi(\nu\sqrt{2})\nu\sqrt{2} \leq \frac{e^{-\nu^2}}{\nu\sqrt{2}} + \nu\sqrt{2}e^{-\nu^2}.$$

Finally substituting $\nu = \sqrt{\log(1/\delta)}$ we get

$$\int_0^\delta \sqrt{\log(1/\epsilon)}\, d\epsilon \leq \delta\left(\sqrt{2\log\frac{1}{\delta}} + \frac{1}{\sqrt{2\log\frac{1}{\delta}}}\right) \leq 2\delta\log(1/\delta).$$

Then, in view of (30), the result follows. This completes the proof. □

C Proof of Theorem 2

Proving the result related to π^* is similar to the proof of (Fan and Gu, 2024, Theorem 2) as the proofs depend on the entire sample space, so it is omitted here. We prove the result related to estimating μ_0^*, μ_1^* and the proof differs from the standard functional guarantees in (Fan and Gu, 2024, Theorem 2) as the above work analyses function estimation with fixed data and our estimators for μ_0^*, μ_1^* are constructed using the random subsamples given by the control group $\{(y_i, \boldsymbol{x}_i, T_i) : T_i = 0\}$ and the treatment group $\{(y_i, \boldsymbol{x}_i, T_i) : T_i = 1\}$ respectively. Our proof will rely on the following auxiliary result. Let P_j denote the conditional law of $\boldsymbol{x}$ given $T = j, j = 0, 1$ and define

$$\mathbb{E}_{P_j}[h] = \int h(\boldsymbol{x}) dP_j(\boldsymbol{x}), \quad \mathbb{E}_{n,j}[h] = \frac{1}{n_j}\sum_{i:T_i=j} h(\boldsymbol{x}_i), \quad n_j = \sum_{i=1}^n \mathbf{1}_{\{T_i=j\}}, \quad j = 0, 1.$$

Lemma 4. Suppose that the conditions in Theorem 2 hold. Then, with probability at least $1 - O(e^{-t} + e^{-n\alpha_*^2/2})$, the following holds, for n large enough and $j = 0, 1$

$$\mathbb{E}_{P_j}\left[\left(\widehat{\mu}_j^{\mathsf{FAST}} - \mu_j^*\right)^2\right] + \mathbb{E}_{n,j}\left[(\widehat{\mu}_j^{\mathsf{FAST}} - \mu_j^*)^2\right] \leq \frac{c}{\alpha_*}\left\{\delta_{\mathsf{opt}} + \delta_{j,a} + \delta_{j,s} + \delta_{j,f} + \frac{t}{n}\right\},$$

where c is a constant.

In view of the above result, the proof of Theorem 2 relies on bounding $\mathbb{E}_P$ using the conditional expectation $\mathbb{E}_{P_j}$, as Assumption 2 implies that $dP(\boldsymbol{x}) = \frac{\mathbb{P}[T=j]}{\mathbb{P}[T=j|\boldsymbol{x}]} dP_j(\boldsymbol{x}) \leq \frac{dP_j(\boldsymbol{x})}{\alpha_*}$, for each $j = 0, 1$. Hence, it remains to prove Lemma 4.

Notations: We use the following notation for the proofs in this section. Given any matrix $\boldsymbol{B}$ with n rows and a subset $\mathcal{J}$ of the index set $\{1, \ldots, n\}$, let $[\boldsymbol{B}]_{\mathcal{J},:}$ denote the submatrix consisting of the rows corresponding to the $\mathcal{J}$ index set. In view of (2), define

$$\widetilde{\boldsymbol{f}} = \frac{1}{p}\boldsymbol{W}^\top \boldsymbol{x}, \quad \boldsymbol{H} = \frac{1}{p}\boldsymbol{W}^\top \boldsymbol{B}. \tag{31}$$

Given a matrix $\boldsymbol{H} \in \mathbb{R}^{\bar{r}\times r}, \bar{r} \geq r$ with full column rank define $\boldsymbol{H}^+$ to be it's left inverse $\boldsymbol{H}^+ = (\boldsymbol{H}^\top \boldsymbol{H})^{-1}\boldsymbol{H}^\top$. Assume that $\mu_1^*(\boldsymbol{x}) = \mu_1(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}_1})$, i.e., the coordinates corresponding to $\mathcal{J}_1$ are active in the output function μ_1^*. Also define

$$v_n = (N^2 L + N\bar{r})\frac{L\log(BNn)}{n}, \quad \varrho_n = \frac{\log(np(N+\bar{r})) + L\log(BN)}{n}. \tag{32}$$

Proof of Lemma 4. We only prove results related to μ_1^* as the result for μ_0^* is similar. We first outline the key steps of the proof of Lemma 4.

- **Step 1:** Show that $\widetilde{\mu_1}^*(\boldsymbol{x}) = \mu_1^*(\boldsymbol{H}^+\widetilde{\boldsymbol{f}}, \boldsymbol{x}_{\mathcal{J}_1} - [\boldsymbol{B}]_{\mathcal{J}_1,:}\boldsymbol{H}^+\widetilde{\boldsymbol{f}})$ is close to $\mu_1^*(\boldsymbol{f}, \boldsymbol{u}_{\mathcal{J}_1})$

  $$\mathbb{E}_P\left[(\widetilde{\mu_1}^* - \mu_1^*)^2\right] \lesssim \frac{|\mathcal{J}_1| r \cdot \bar{r}}{(\nu_{\min}(\boldsymbol{H}))^2 p} = \delta_{1,f}, \tag{33}$$

  where $\mathbb{E}_P$ is the expectation with respect to the unconditional distribution of $\boldsymbol{x}$. Then noting that $\pi^*(\boldsymbol{x}) = \mathbb{P}[T = 1|\boldsymbol{x}] \in (\alpha_*, 1-\alpha_*)$ for all $\boldsymbol{x}$, we get

  $$\mathbb{E}_{P_1}\left[(\widetilde{\mu_1}^* - \mu_1^*)^2\right] \leq \frac{\mathbb{E}_P\left[(\widetilde{\mu_1}^* - \mu_1^*)^2\right]}{\alpha_*} = \delta_{1,f}/\alpha_*. \tag{34}$$

- **Step 2:** Define the function class

  $$\mathcal{G}_m = \left\{m^{\mathsf{FAST}}(\boldsymbol{x}; \boldsymbol{W}, g, \boldsymbol{\Theta}) : g \in \mathcal{G}(L, \bar{r}+N, 1, N, M, B), \boldsymbol{\Theta} \in \mathbb{R}^{p\times N}, \|\boldsymbol{\Theta}\|_{\max} \leq B\right\} \tag{35}$$

  Then there exists $\widetilde{\mu_1} \in \mathcal{G}_m$ (i.e., with corresponding $\widetilde{\boldsymbol{\Theta}}_1$) such that $\|\widetilde{\boldsymbol{\Theta}}_1\|_0 \leq |\mathcal{J}_1|$ and

  $$\mathbb{E}_P\left[(\widetilde{\mu_1} - \widetilde{\mu_1}^*)^2\right] \lesssim \delta_{1,f} + \delta_{1,a}. \tag{36}$$

  Similar to (37) we get $\mathbb{E}_{P_1}\left[(\widetilde{\mu_1} - \widetilde{\mu_1}^*)^2\right] \lesssim (\delta_{1,f} + \delta_{1,a})/\alpha_*$, which implies

  $$\mathbb{E}_{P_1}\left[(\widetilde{\mu_1} - \mu_1^*)^2\right] \leq 2\left(\mathbb{E}_{P_1}\left[(\widetilde{\mu_1} - \widetilde{\mu_1}^*)^2\right] + \mathbb{E}_{P_1}\left[(\widetilde{\mu_1}^* - \mu_1^*)^2\right]\right) \lesssim (\delta_{1,f} + \delta_{1,a})/\alpha_*. \tag{37}$$

- **Step 3:** Derive the basic inequality

  $$\begin{aligned}
  &\mathbb{E}_{n,1}[(\widehat{\mu_1}^{\mathsf{FAST}} - \widetilde{\mu_1})^2] + 2\lambda_1 \sum_{i,j} \psi_{\tau_1}(\widehat{\Theta}_{1,i,j}) \\
  &\leq 4\mathbb{E}_{n,1}[(\widetilde{\mu_1} - \mu_1^*)^2] + \frac{4}{n_1}\sum_{i\in[n]:T_i=1} \varepsilon_i(1)(\widehat{\mu_1}^{\mathsf{FAST}}(\boldsymbol{x}_i) - \widetilde{\mu_1}(\boldsymbol{x}_i)) + 2\lambda_1|\mathcal{J}_1| + 2\delta_{\mathsf{opt}}.
  \end{aligned} \tag{38}$$

- **Step 4:** Show that the following event occurs with a probability at least $1 - e^{-t}$

$$\mathcal{B}_{t,1/2} = \Bigg\{ \forall \mu_1 = m^{\mathsf{FAST}}\ (\cdot; \boldsymbol{W}, g, \boldsymbol{\Theta}) \in \mathcal{G}_m,\ \frac{4}{n_1} \sum_{i \in [n]: T_i = 1} \varepsilon_i(1)(\mu_1(\boldsymbol{x}_i) - \widetilde{\mu}_1(\boldsymbol{x}_i)) \\ - \lambda_1 \sum_{i,j} \psi_\tau(\Theta_{1,i,j}) \le \frac{1}{2}\mathbb{E}_{n,1}[(\mu_1 - \widetilde{\mu}_1)^2] + 2\left(v_{n_1} + \varrho_{n_1} + \frac{t}{n_1}\right) \Bigg\}. \tag{39}$$

- **Step 5:** Show that the following event occurs with a probability $1 - e^{-t} - e^{-n\alpha_*^2/2}$

$$\mathcal{C}_t = \Bigg\{ \forall \mu_1 = m^{\mathsf{FAST}}\ (\cdot; \boldsymbol{W}, g, \boldsymbol{\Theta}) \in \mathcal{G}_m,\ \frac{1}{2}\mathbb{E}_{P_1}\left[(\mu_1 - \widetilde{\mu}_1)^2\right] \\ \le \frac{1}{2}\mathbb{E}_{n,1}[(\mu_1 - \widetilde{\mu}_1)^2] + 2\lambda_1 \sum_{i,j} \psi_{\tau_1}(\Theta_{i,j}) + C_5\left(v_{n_1} + \varrho_{n_1} + \frac{t}{n_1}\right) \Bigg\}. \tag{40}$$

- **Step 6:** We bound the separation between μ_1^* and $\widetilde{\mu}_1$ from Step 2. For every $0 < t \le n$, there is an event $\mathcal{A}_t$ with $\mathbb{P}[\mathcal{A}_t] \ge 1 - e^{-t} - e^{-n\alpha_*^2/2}$ on which

$$\mathbb{E}_{n,1}[(\widetilde{\mu}_1 - \mu_1^*)^2] \lesssim \frac{1}{\alpha_*}\left(\delta_{1,f} + \delta_{1,a} + \frac{t}{n}\right). \tag{41}$$

- **Step 7:** We bound $\mathbb{E}_{n,1}[(\widehat{\mu}_1^{\mathsf{FAST}} - \mu_1^*)^2]$. Using (38), (39) and (41) we get on the event $\mathcal{B}_{t,1/2}$

$$\mathbb{E}_{n,1}[(\widehat{\mu}_1^{\mathsf{FAST}} - \widetilde{\mu}_1)^2] + 2\lambda_1 \sum_{i,j} \psi_{\tau_1}(\widehat{\Theta}_{1,i,j}) \\ \le 4\lambda_1 |\mathcal{J}_1| + 4\delta_{\mathsf{opt}} + C_4\left(v_{n_1} + \varrho_{n_1} + \frac{t}{n_1}\right) + \frac{\widetilde{C}_4}{\alpha_*}\left(\delta_{1,f} + \delta_{1,a} + \frac{t}{n}\right). \tag{42}$$

  Combining the last display and (41) with the following facts

  - $(\widehat{\mu}_1^{\mathsf{FAST}} - \mu_1^*)^2 \le 2[(\widetilde{\mu}_1 - \mu_1^*)^2 + (\widehat{\mu}_1^{\mathsf{FAST}} - \widetilde{\mu}_1)^2]$,
  - for $n_1 \in (n\alpha_*/2, n)$

$$v_{n_1} \lesssim \frac{v_n}{\alpha_*}, \quad \varrho_{n_1} \lesssim \frac{\varrho_n}{\alpha_*} \tag{43}$$

  we get on $\mathcal{E}_1 = \mathcal{A}_t \cap \mathcal{B}_{t,1/2} \cap \mathcal{D}$ ($\mathcal{D}$ is as in Lemma 5) with $\mathbb{P}[\mathcal{E}_1] \ge 1 - e^{-t} - e^{-n\alpha_*^2/2}$

$$\mathbb{E}_{n,1}[(\widehat{\mu}_1^{\mathsf{FAST}} - \mu_1^*)^2] \lesssim \lambda_1 |\mathcal{J}| + \delta_{\mathsf{opt}} + \frac{1}{\alpha_*}\left(v_n + \varrho_n + \delta_{1,f} + \delta_{1,a} + \frac{t}{n}\right). \tag{44}$$

- **Step 8:** We bound $\mathbb{E}_{P_1}[(\widehat{\mu}_1^{\mathsf{FAST}} - \mu_1^*)^2]$. On the event $\mathcal{C}_t \cap \mathcal{D}$ we use (42) to get

$$\mathbb{E}_{P_1}\left[\left(\widehat{\mu}_1^{\mathsf{FAST}} - \widetilde{\mu}_1\right)^2\right] \leq \mathbb{E}_{n,1}[(\widehat{\mu}_1^{\mathsf{FAST}} - \widetilde{\mu}_1)^2] + 4\lambda_1 \sum_{i,j} \psi_{\tau_1}(\widehat{\Theta}_{1,i,j}) + C_5\left(v_{n_1} + \varrho_{n_1} + \frac{t}{n_1}\right)$$
$$\lesssim \frac{1}{\alpha_*}\left(\lambda_1|\mathcal{J}_1| + \delta_{1,f} + \delta_{\mathsf{opt}} + v_n + \varrho_n + \frac{t}{n}\right). \tag{45}$$

where the last inequality followed using (43). We continue the last display using $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \leq 2(\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2)$ and (37) to get on the event $\mathcal{E}_2 = \mathcal{B}_{t,1/2} \cap \mathcal{C}_t \cap \mathcal{D}$ with $\mathbb{P}[\mathcal{E}_1] \geq 1 - e^{-t} - e^{-n\alpha_*^2/2}$

$$\mathbb{E}_{P_1}\left[\left(\widehat{\mu}_1^{\mathsf{FAST}} - \mu_1^*\right)^2\right] \leq 2\left(\mathbb{E}_{P_1}\left[\left(\widehat{\mu}_1^{\mathsf{FAST}} - \widetilde{\mu}_1\right)^2\right] + \mathbb{E}_{P_1}\left[(\widetilde{\mu}_1 - \mu_1^*)^2\right]\right)$$
$$\lesssim \frac{1}{\alpha_*}\left(\lambda_1|\mathcal{J}_1| + \delta_{\mathsf{opt}} + v_n + \varrho_n + \delta_{1,f} + \delta_{\mathsf{opt}} + \frac{t}{n}\right). \tag{46}$$

To complete the proof of Lemma 4 we only prove steps 1–6, as the other steps follow from them via simple algebra. The proof of Steps 1-2 follows from the proof of (Fan and Gu, 2024, Theorem 2) and uses properties of the functions $\widetilde{\mu}_1, \widetilde{\mu}_1{}^*, \mu_1^*$. Next, we outline the proof of Step 3. Note that from the definition of $\widehat{g}_1, \widehat{\boldsymbol{\Theta}}_1$ in (16) it follows

$$\frac{1}{n_1}\sum_{i\in[n],T_i=1}\left\{y_i - \widehat{\mu}_1^{\mathsf{FAST}}(\boldsymbol{x}_i)\right\}^2 + \lambda_1\sum_{i,j}\psi_{\tau_1}(\widehat{\Theta}_{1,i,j})$$
$$\leq \frac{1}{n_1}\sum_{i\in[n],T_i=1}\{y_i - \widetilde{\mu}_1(\boldsymbol{x}_i)\}^2 + \lambda_1\sum_{i,j}\psi_{\tau_1}(\widetilde{\Theta}_{1,i,j}) + \delta_{\mathsf{opt}}$$

Substituting $y_i = \mu_1^*(\boldsymbol{x}_i) + \varepsilon_i(1), T_i = 1$ in the above expression, we get

$$\mathbb{E}_{n,1}[(\mu_1^* - \widehat{\mu}_1^{\mathsf{FAST}})^2] + \lambda_1\sum_{i,j}\psi_{\tau_1}(\widehat{\Theta}_{1,i,j})$$
$$\leq \mathbb{E}_{n,1}[(\mu_1^* - \widetilde{\mu}_1)^2] + \frac{2}{n_1}\sum_{i\in[n]:T_i=1}\varepsilon_i(1)(\mu_1^{\mathsf{FAST}}(\boldsymbol{x}_i) - \widetilde{\mu}_1(\boldsymbol{x}_i)) + \lambda_1\sum_{i,j}\psi_{\tau_1}(\widetilde{\Theta}_{1,i,j}) + \delta_{\mathsf{opt}}$$

In view of the construction of $\widetilde{\boldsymbol{\Theta}}_1$ in Step 2 we have

$$\sum_{i,j}\psi_{\tau_1}(\widetilde{\Theta}_{1,i,j}) \leq \|\widetilde{\boldsymbol{\Theta}}_1\|_0 \leq |\mathcal{J}_1|.$$

Using the last display and $\frac{1}{2}(\widehat{\mu}_1^{\mathsf{FAST}} - \widetilde{\mu}_1)^2 \leq (\widehat{\mu}_1^{\mathsf{FAST}} - \mu_1^*)^2 + (\mu_1^* - \widetilde{\mu}_1)^2$ we can rearrange the expressions to derive (38). The proof of Step 4 follows from (Fan and Gu, 2024, Lemma 10) along with a union bound argument. In view of the proof of (Fan and Gu, 2024, Lemma

10) we note that by defining

$$\begin{aligned}
&\mathcal{B}_{t,1/2}(\{\boldsymbol{x}_i : T_i = 1, i \in [n]\}) \\
&= \Bigg\{ \forall \mu_1 = m^{\mathsf{FAST}}\ (\cdot; \boldsymbol{W}, g, \boldsymbol{\Theta}) \in \mathcal{G}_m,\ \frac{4}{n_1} \sum_{i \in [n]: T_i = 1} \varepsilon_i(1)(\mu_1(\boldsymbol{x}_i) - \widetilde{\mu}_1(\boldsymbol{x}_i)) \\
&- \lambda_1 \sum_{i,j} \psi_\tau(\Theta_{1,i,j})\ \leq \frac{1}{2n_1} \sum_{i \in [n]: T_i = 1} (\mu_1(\boldsymbol{x}_i) - \widetilde{\mu}_1(\boldsymbol{x}_i))^2 + 2\left(v_{n_1} + \varrho_{n_1} + \frac{t}{n_1}\right) \Bigg\},
\end{aligned}$$

we get $\mathbb{P}[\mathcal{B}_{t,1/2}(\{\boldsymbol{x}_i : T_i = 1\})] \geq 1 - e^{-t}$ holds for every fixed realization of $\{\boldsymbol{x}_i : T_i = 1\}$. Hence, we can then apply the Law of Total Probability to conclude the statement.

To prove Step 5, define $\mathcal{I} = \{i \in [n] : T_i = 1\}$ and let $|\mathcal{I}|$ denote the size of $\mathcal{I}$. Note that conditioned on a fixed realization of $\mathcal{I}$, the points $\{\boldsymbol{x}_i\}_{i \in \mathcal{I}}$ are independently distributed with the distribution P_1. Next, we restrict ourselves to the event $\mathcal{D} = \{\sum_{i \in [n]} T_i \geq n\alpha_*/2\}$, which occurs with a probability at least $1 - e^{-n\alpha_*^2/2}$ in view of Lemma 5. In view of the above, we can first show that for each fixed realization from the event $\mathcal{D}$, the following event holds with a probability $1 - e^{-t}$, for any fixed $\widetilde{\mu}_1 \in \mathcal{G}_m$

$$\begin{aligned}
\mathcal{C}_{t,\mathcal{I}} = \Bigg\{ &\forall \mu_1 = m^{\mathsf{FAST}}\ (\cdot; \boldsymbol{W}, g, \boldsymbol{\Theta}) \in \mathcal{G}_m,\ \frac{1}{2}\mathbb{E}_{P_1}\left[(\mu_1 - \widetilde{\mu}_1)^2\right] \\
&\leq \frac{1}{2|\mathcal{I}|} \sum_{i \in \mathcal{I}} (\mu_1(\boldsymbol{x}_i) - \widetilde{\mu}_1(\boldsymbol{x}_i))^2 + 2\lambda_1 \sum_{i,j} \psi_{\tau_1}(\Theta_{i,j}) + C_5\left(v_{|\mathcal{I}|} + \varrho_{|\mathcal{I}|} + \frac{t}{|\mathcal{I}|}\right) \Bigg\}.
\end{aligned}$$

In addition, by further conditioning on the event in Lemma 5 we get that $|\mathcal{I}|$ is of constant order compared to n. Then we can follow the proof of (Fan and Gu, 2024, Lemma 9) to show that

$$\mathbb{P}\left[\mathcal{C}_{t,\mathcal{I}} \cap \mathcal{D}\right] \geq 1 - e^{-t} - e^{-n\alpha_*^2/2}.$$

Hence, using the law of total probability we get

$$\mathbb{P}[\mathcal{C}_{t,\mathcal{I}}] \geq \mathbb{P}[\mathcal{C}_{t,\mathcal{I}}|\mathcal{D}] \cdot \mathbb{P}[\mathcal{D}] \geq (1 - e^{-t})(1 - e^{-n\alpha_*^2/2}) \geq 1 - e^{-t} - e^{-n\alpha_*^2/2}.$$

We present the proof of Step 6 below. We will first apply (Fan and Gu, 2024, Lemma 9, Lemma 10) based on every fixed realization from the event $\mathcal{D}$ as in Lemma 5. Note that the random variables $\{\boldsymbol{x}_i : i \in \mathcal{I}\}$ are independently and identically distributed. This implies that for μ_1^* and $\widetilde{\mu}_1 \in \mathcal{G}_m$ as in Step 2, the following collection of random variables

$$z_i = (\widetilde{\mu}_1(\boldsymbol{x}_i) - \mu_1^*(\boldsymbol{x}_i))^2, \quad i \in [n], T_i = 1$$

are independent and satisfies (as $\widetilde{\mu}_1 \in [-M, M], \mu_1^* \in [-M^*, M^*]$ using Assumption 2)

$$z_i \leq (M + M^*)^2,\ \mathbb{E}_{P_1}[z_i^2] \leq (M + M^*)^2 \mathbb{E}_{P_1}[(\widetilde{\mu}_1 - \mu_1^*)^2] \lesssim \frac{\delta_{1,f} + \delta_{1,a}}{\alpha_*},$$

Hence, conditioned on n_1 we can apply the Bernstein Inequality (Boucheron et al., 2003) to conclude that with a probability $1 - e^{-t}$

$$\mathbb{E}_{n,1}[(\widetilde{\mu}_1 - \mu_1)^2] \leq \mathbb{E}_{P_1}[(\widetilde{\mu}_1 - \mu_1^*)^2] + C\left(\frac{\delta_{1,f} + \delta_{1,a}}{\alpha_*}\sqrt{\frac{t}{n_1}} + \frac{t}{n_1}\right), \tag{47}$$

for a constant $C > 0$. Next we show that $n_1 \geq \frac{n\alpha_*}{2}$ with a probability at least $1 - e^{-\frac{n\alpha_*^2}{2}}$. In view of (47) we use the last display and use a union bound to conclude that there exists a constant $C > 0$ such that for all $0 < t \leq n$

$$\mathbb{P}\left[\mathbb{E}_{n,1}[(\widetilde{\mu}_1 - \mu_1^*)^2] \leq \frac{C}{\alpha_*}\left(\delta_{1,f} + \delta_{1,a} + \frac{t}{n}\right)\right] \geq 1 - e^{-t} - e^{-n\alpha_*^2/2}.$$

□

Lemma 5. Define the event $\mathcal{D} = \{\sum_{i\in[n]} T_i \geq n\alpha_*/2\}$. Then $\mathbb{P}[\mathcal{D}] \geq 1 - e^{-n\alpha_*^2/2}$.

Proof. Note that as $\inf_{\boldsymbol{x}} \pi^*(\boldsymbol{x}) \geq \alpha_*$ we get $\sum_{i\in[n]} T_i$ are stochastically larger than $Z \sim \text{Binom}(n, \alpha_*)$. Hence, we get for $c > 1$ to be chosen later

$$\mathbb{P}[n_1 \leq \frac{n\alpha_*}{c}] \leq \mathbb{P}[Z \leq \frac{n\alpha_*}{c}] \leq \mathbb{P}[n - Z \geq n\left(1 - \frac{\alpha_*}{c}\right)] \tag{48}$$

Here $n - Z \sim \text{Binom}(1-\alpha_*)$. We will use Chernoff's inequality for Binomial random variables

Lemma 6. (Boucheron et al., 2003, Section 2.2) For a random variable $Z \sim \text{Binom}(m, q)$, we have

$$\mathbb{P}\left[Z \geq ma\right] \leq \exp\left(-mh_q(a)\right); \quad q < a < 1, \; h_q(a) = a\log\frac{a}{q} + (1-a)\log\frac{1-a}{1-q}.$$

Using $q = 1 - \alpha_*, a = 1 - \alpha_*/c$ in the definition of $h_q(a)$ in the above result and using Pinsker's inequality $h_q(a) \geq 2(a-q)^2$ we get $h_q(a) \geq 2\alpha_*^2(c-1)^2/c^2$. Hence we continue (48) using Lemma 6 to get

$$\mathbb{P}\left[n_1 \leq \frac{n\alpha_*}{c}\right] \leq \exp\left(-2n\alpha_*^2(c-1)^2/c^2\right).$$

Plugging in $c = 2$ in the above inequality, we get the desired result. □

Proof of Theorem 2 for $r = 0$. The proof here mainly deviates from the proof of the case $r \geq 1$ in establishing (34) and (37). We modify the steps as follows. Note that we have $\delta_{1,f} = 0$ and $r = 0$. We also get from $r = 0$ that $\mu_1^*(\boldsymbol{x}) = \mu_1^*(\boldsymbol{x}_{\mathcal{J}_1})$. Then, from the definition of $\delta_{1,a}$ as in Theorem 2 note that there exists $\widetilde{\mu}_1 \in \mathcal{G}_m$ (i.e., with corresponding $\widetilde{\boldsymbol{\Theta}}_1$, and $\mathcal{G}_m$ is as defined in (35)) such that $\|\widetilde{\boldsymbol{\Theta}}_1\|_0 \leq |\mathcal{J}_1|$ and $\mathbb{E}_P\left[(\widetilde{\mu}_1 - \mu_1^*)^2\right] \lesssim \delta_{1,a}$. Similar to (37), we get $\mathbb{E}_{P_1}\left[(\widetilde{\mu}_1 - \widetilde{\mu}_1^{\,*})^2\right] \lesssim \delta_{1,a}/\alpha_*$. Then the other parts of the proof, from Step 3 onwards, can be carried out as before. This completes our argument. □

D Empirical Implementation Details

D.1 Parameter choices for candidate methods

The scripts to submit each simulation as a job on the cluster are named identically to the file for the corresponding Python codes with an extension of '*.sh*'. All the directories in the Python code are saved relatively, so people can execute the code under any directory without changing the paths inside the Python scripts. GPUs are recommended to simulate Double Deep Learning, Vanilla Neural Networks with L_2 regularization.

- Factor Informed Double Deep Learning Estimator (FIDDLE): We implement a factor-augmented sparse throughput deep (FAST) ReLU neural network to estimate the average treatment effect (ATE). Set the number of epochs in training to be 100, the batch size to be 64, the learning rate $lr = 0.001$, the depth of the neural network $L = 4$, and the width of the neural network $N = 400$ and the column number of the diversified projection matrix $\overline{r} = 10$. The hyperparameters for the penalty in the FAST architecture are set to $\tau = 0.005$ and $\lambda = 1.3\log(p)/n$. We randomly sample $m = 50$ unlabeled observations of covariates to pre-train the diversified projection matrix $\boldsymbol{W}$ and leave the rest of the dataset to estimate the propensity and outcomes models. The column number of the diversified projection matrix $\overline{r} = 3$ only for the experiments on the MBSAQIP dataset in Section 5.4, due to a smaller number of covariates.

- Vanilla Neural Networks (Vanilla-NN): We adopt a fully connected ReLU neural network with the same number of epochs in training to be 100, the batch size 64, and the learning rate $lr = 0.001$, the depth of the neural network $L = 4$ and the width of the neural network $N = 400$. We penalize the loss function by an L_2 norm term with weight $\lambda = 1$.

- Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE): We adapt the official package published by (van der Schaar Lab, 2025) to implement this method. The hyperparameters for the simulation implementation are set as default: the hidden dimensions $h\_dim = 100$, number of training iterations $num\_iterations = 5000$, the batch size 256, hyperparameters to adjust the loss importance $\alpha = 0.1$ and $\beta = 0$.

- Double Robust Forest Model (DR): We implement DR by the function *econml.dr.DRLearner* from the EconML package (Research, 2025). The propensity score is modeled by the function *sklearn.ensemble.RandomForestClassifier* in the package (Pedregosa et al., 2011) with number of trees $n\_estimators = 100$ and the maximum depth of the tree $max\_depth = 2$. Both the outcome and the final model are implemented by the function *sklearn.ensemble.RandomForestRegressor* with both models have number of trees $n\_estimators = 100$ and the maximum depth of the tree $max\_depth = 2$ separately. We truncate the $min\_propensity = 0.1$, which is the minimum propensity at which to clip propensity estimates to avoid dividing by zero.

- Double Machine Learning Forest Model (DML): We implement DML by the function *econml.dml.CausalForestDML* from the EconML package (Research, 2025). The

propensity score is modeled by the function *sklearn.ensemble.RandomForestClassifier* in the package (Pedregosa et al., 2011) with number of trees $n\_estimators = 100$ and the maximum depth of the tree $max\_depth = 2$. The outcome is modeled by the function *sklearn.ensemble.RandomForestRegressor* with number of trees $n\_estimators = 100$ and the maximum depth of the tree $max\_depth = 2$ separately.

- Causal Forest (CF) on Covariates or Latent Factors: We implement CF by the function *econml.grf.CausalForest* from the EconML package (Research, 2025). The parameters to implement the function are set as default: number of trees $n\_estimators = 100$, the maximum depth of the tree $max\_depth = 50$.

D.2 Additional tables and plots

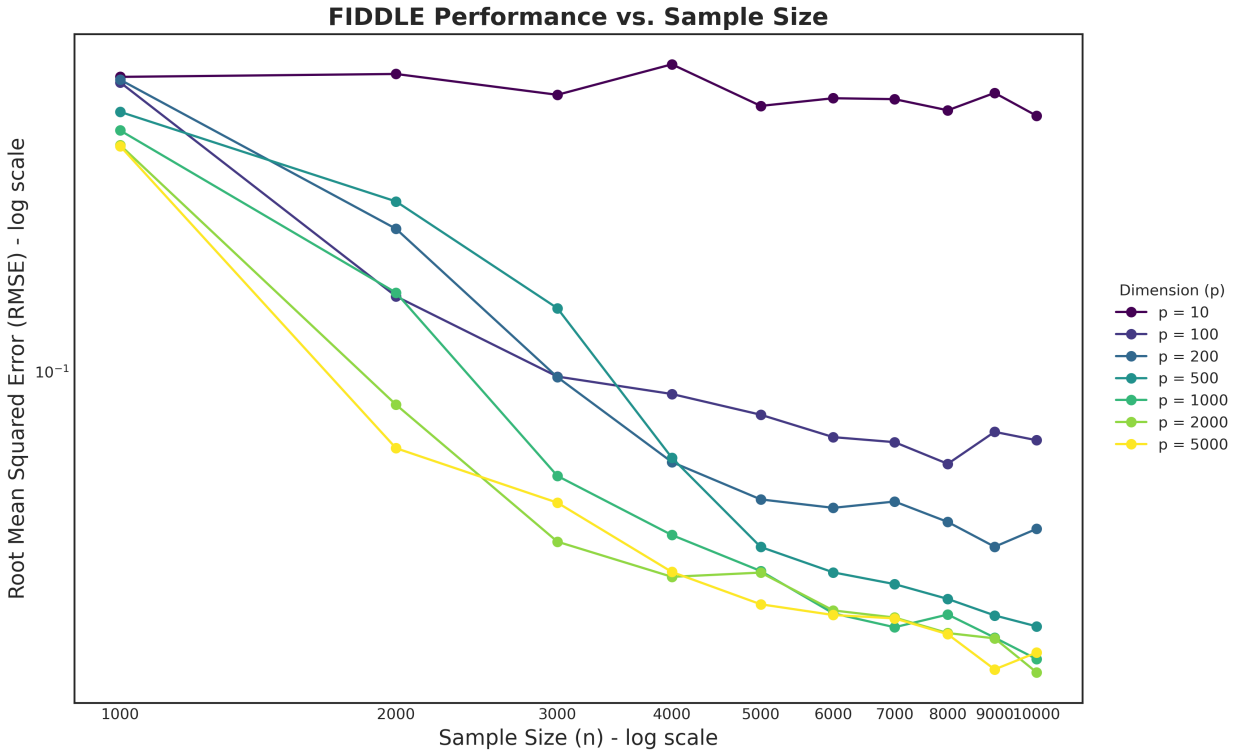

Figure 3: Plot of root mean squared error (RMSE) by FIDDLE performance across different sample sizes (n) and covariate dimensions (p) among 100 replications (Both x and y axis plot on a log scale).

	$p = 10$	$p = 100$	$p = 200$	$p = 500$	$p = 1000$	$p = 2000$	$p = 5000$
$n = 1000$	0.5247	0.5087	0.5164	0.4310	0.3883	0.3577	0.3557
	(0.0342)	(0.0167)	(0.0166)	(0.0139)	(0.0130)	(0.0145)	(0.0134)
$n = 2000$	0.5333	0.1529	0.2234	0.2606	0.1561	0.0832	0.0651
	(0.0299)	(0.0089)	(0.0079)	(0.0074)	(0.0086)	(0.0120)	(0.0150)
$n = 3000$	0.4744	0.0974	0.0971	0.1430	0.0557	0.0385	0.0480
	(0.0251)	(0.0057)	(0.0040)	(0.0058)	(0.0037)	(0.0026)	(0.0106)
$n = 4000$	0.5631	0.0883	0.0603	0.0617	0.0400	0.0316	0.0325
	(0.0282)	(0.0058)	(0.0031)	(0.0041)	(0.0025)	(0.0018)	(0.0025)
$n = 5000$	0.4457	0.0786	0.0489	0.0374	0.0327	0.0324	0.0271
	(0.0239)	(0.0044)	(0.0030)	(0.0029)	(0.0022)	(0.0019)	(0.0018)
$n = 6000$	0.4654	0.0693	0.0466	0.0324	0.0258	0.0262	0.0255
	(0.0231)	(0.0035)	(0.0026)	(0.0023)	(0.0018)	(0.0022)	(0.0018)
$n = 7000$	0.4629	0.0674	0.0483	0.0304	0.0238	0.0252	0.0250
	(0.0240)	(0.0044)	(0.0037)	(0.0023)	(0.0017)	(0.0016)	(0.0018)
$n = 8000$	0.4349	0.0597	0.0430	0.0279	0.0256	0.0231	0.0229
	(0.0213)	(0.0033)	(0.0026)	(0.0024)	(0.0019)	(0.0019)	(0.0017)
$n = 9000$	0.4795	0.0714	0.0374	0.0255	0.0225	0.0224	0.0188
	(0.0242)	(0.0054)	(0.0026)	(0.0020)	(0.0014)	(0.0014)	(0.0014)
$n = 10000$	0.4217	0.0681	0.0414	0.0239	0.0199	0.0185	0.0207
	(0.0200)	(0.0038)	(0.0021)	(0.0015)	(0.0013)	(0.0016)	(0.0013)

Table 4: Root mean squared error and standard error (in parentheses) of our proposed FIDDLE across different sample sizes (n) and covariate dimensions (p) over 100 replications. For each n, the first row shows RMSE and the second row shows SE in parentheses.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

American College of Surgeons (2025). Metabolic and bariatric surgery accreditation and quality improvement program (mbsaqip). Website: https://www.facs.org/quality-programs/accreditation-and-verification/metabolic-and-bariatric-surgery-accreditation-and-quality-improvement-program/. Accessed: 2025-06-16.

Aronow, P. M. and Carnegie, A. (2013). Beyond late: Estimation of the average treatment effect with an instrumental variable. *Political Analysis*, 21(4):492–506.

Arterburn, D., Wellman, R., Emiliano, A., Smith, S. R., Odegaard, A. O., Murali, S., Williams, N., Coleman, K. J., Courcoulas, A., Coley, R. Y., et al. (2018). Comparative effectiveness and safety of bariatric procedures for weight loss: a PCORnet cohort study. *Annals of Internal Medicine*, 169(11):741–750.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression.

Bhattacharya, S., Fan, J., and Mukherjee, D. (2024). Deep neural networks for nonparametric interaction models with diverging dimension. *The Annals of Statistics*, 52(6):2738–2766.

Boucheron, S., Lugosi, G., and Bousquet, O. (2003). Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer.

Brand, J. E., Zhou, X., and Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, 49:81–110.

Candes, E. and Romberg, J. (2007). Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969.

Chamberlain, G. and Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets.

Chen, X., Jana, S., Metzler, C. A., Maleki, A., and Jalali, S. (2025). Multilook coherent imaging: Theoretical guarantees and algorithms. *arXiv preprint arXiv:2505.23594*.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

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

Dax, A. (2013). From eigenvalues to singular values: a review. *Advances in Pure Mathematics*, 3(9):8–24.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236.

Du, X., Fan, Y., Lv, J., Sun, T., and Vossler, P. (2021). Dimension-free average treatment effect inference with deep neural networks. *arXiv preprint arXiv:2112.01574*.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J. and Gu, Y. (2024). Factor augmented sparse throughput deep relu neural networks for high dimensional regression. *Journal of the American Statistical Association*, 119(548):2680–2694.

Fan, J., Gu, Y., and Zhou, W.-X. (2024). How do noise tails impact on deep relu networks? *The Annals of Statistics*, 52(4):1845–1871.

Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y., and Yang, X. (2022). Optimal covariate balancing conditions in propensity score estimation. *Journal of Business & Economic Statistics*, 41(1):97–110.

Fan, J., Ke, Y., and Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of econometrics*, 216(1):71–85.

Fan, J. and Liao, Y. (2022). Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association*, 117(538):909–924.

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021). Robust high dimensional factor models with applications to statistical machine learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2):303.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

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

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

Hady, H. R., Dadan, J., Gołaszewski, P., and Safiejko, K. (2012). Impact of laparoscopic sleeve gastrectomy on body mass index, ghrelin, insulin and lipid levels in 100 obese patients. *Videosurgery and other Miniinvasive Techniques*, 7(4):251–259.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hoffmann, N. I. (2024). Double robust, flexible adjustment methods for causal inference: An overview and an evaluation.

Hutter, M. M., Schirmer, B. D., Jones, D. B., Ko, C.-Y., Cohen, M. E., Merkow, R. P., and Nguyen, N. T. (2013). Outcome analysis of early laparoscopic sleeve gastrectomy experience. *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques*, 23(6):515–519.

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

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Mousavi, A., Patel, A. B., and Baraniuk, R. G. (2015). A deep learning approach to structured signal recovery. In *2015 53rd annual allerton conference on communication, control, and computing (Allerton)*, pages 1336–1343. IEEE.

Oh, D. H. and Patton, A. J. (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics*, 35(1):139–154.

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review*, 96(1):152–175.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pereira, A. M., Moura, D., Pereira, S. S., Andrade, S., Almeida, R. F. d., Nora, M., Monteiro, M. P., and Guimarães, M. (2024). Beyond restrictive: sleeve gastrectomy to single anastomosis duodeno-ileal bypass with sleeve gastrectomy as a spectrum of one single procedure. *Obesity Facts*, 17(4):364–371.

Research, M. (2025). Econml: A python package for estimating heterogeneous treatment effects. https://github.com/py-why/EconML. Accessed: 2025-01-30.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function.

Tang, S., Jana, S., and Fan, J. (2024). Factor adjusted spectral clustering for mixture models. *arXiv preprint arXiv:2408.12564*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Vaart, A. v. d. and Wellner, J. A. (2023). Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer.

van der Laan, L., Luedtke, A., and Carone, M. (2024). Automatic doubly robust inference for linear functionals via calibrated debiased machine learning. *arXiv preprint arXiv:2411.02771*.

van der Schaar Lab (2025). Ganite: Counterfactual inference using generative adversarial networks. https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/ganite. Accessed: 2025-01-30.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural networks*, 94:103–114.

Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*.

Zhou, G., Han, Y., and Yu, X. (2025). Factor augmented tensor-on-tensor neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22928–22936.