Learning Decision Policies with Instrumental Variables through Double Machine Learning

Daqian Shao [1] **Ashkan Soleymani** [2] **Francesco Quinzan** [1] **Marta Kwiatkowska** [1]

Abstract

A common issue in learning decision-making policies in data-rich settings is spurious correlations in the offline dataset, which can be caused by *hidden confounders*. Instrumental variable (IV) regression, which utilises a key unconfounded variable known as the *instrument*, is a standard technique for learning causal relationships between confounded *action*, *outcome*, and *context* variables. Most recent IV regression algorithms use a two-stage approach, where a deep neural network (DNN) estimator learnt in the first stage is directly plugged into the second stage, in which another DNN is used to estimate the causal effect. Naively plugging the estimator can cause heavy bias in the second stage, especially when regularisation bias is present in the first stage estimator. We propose DML-IV, a non-linear IV regression method that reduces the bias in two-stage IV regressions and effectively learns high-performing policies. We derive a novel learning objective to reduce bias and design the DML-IV algorithm following the *double/debiased machine learning (DML)* framework. The learnt DML-IV estimator has strong convergence rate and $O(N^{-1/2})$ suboptimality guarantees that match those when the dataset is unconfounded. DML-IV outperforms state-of-the-art IV regression methods on IV regression benchmarks and learns high-performing policies in the presence of instruments.

1. Introduction

Recent advances in deep learning (DL) have greatly facilitated the learning of decision-making policies in data-rich settings, but they often lack optimality guarantees. A common issue for learning from offline observational data is the existence of spurious correlations, which are relationships between variables that appear to be causal, but in fact are not. For example, suppose we have aeroplane ticket sales and pricing data in a ticket demand scenario (Hartford et al., 2017), and we wish to learn a policy from this offline data that maximises revenue. During holiday season, observational data may contain evidence of a concurrent surge in both ticket sales and prices, which may result in the learning algorithm to learn an incorrect policy that higher ticket prices will drive higher sales.

Spurious correlations are often caused by *hidden confounders* (Pearl, 2000), which are unobserved variables that influence both the *actions* (or *interventions*) and the *outcome*. In the aeroplane ticket example, the occurrence of popular events and holidays serves as a hidden confounder that raises both ticket prices (actions) and sales (outcome). To properly account for these hidden confounders and understand the true causal effect of actions, we need to model the causal (or structural) relationship between the action and the outcome, which is expressed through a *causal function*. However, learning the causal function in the presence of hidden confounders is known to be challenging and sometimes infeasible (Shpitser & Pearl, 2008).

A popular approach to deal with hidden confounders is via *instrumental variables* (IVs) (Wright, 1928), which are heterogeneous random variables that only affect the action, but not the outcome. These IVs have been used extensively to identify the causal effect of actions in many applications, including econometrics (Reiersöl, 1945; Angrist & Pischke, 2009), drug testings (Angrist et al., 1996), and social sciences (Angrist, 1990). In the aeroplane ticket example, we can employ supply cost-shifters (e.g., fuel price) as instrumental variables, as their variations are independent of the demand for aeroplane tickets and affect sales solely via ticket prices (Blundell et al., 2012).

We focus on the problem of learning the causal function in the presence of hidden confounders using IVs (known as *IV regression*), in order to learn a decision policy that maximises the expected outcome in this setting (which we refer to as the *offline IV bandit* problem, described in Section 2.3)

[1]Department of Computer Science, University of Oxford, UK [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA. Correspondence to: Daqian Shao <daqian.shao@cs.ox.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

and comes with suboptimality guarantees. Two-stage least squares (2SLS) (Angrist et al., 1996) is a classical IV regression algorithm, which has been extended to non-linear settings that utilise machine learning (ML) techniques, including deep neural networks (DNNs), to learn the causal function. The use of DNNs allows for greater flexibility in IV regression, as it does not impose strong assumptions on the functional form and can learn directly from data. However, regularisation is often employed to trade-off overfitting with the induced regularisation bias, especially for high-dimensional inputs. Both regularisation bias and overfitting may cause heavy bias (Chernozhukov et al., 2018) in estimating the causal function when the first stage estimator is naively plugged in, which causes slow convergence of the causal function estimator.

Double/Debiased Machine Learning (Chernozhukov et al., 2018) (DML) is a statistical technique that provides an unbiased estimator with convergence rate guarantees for general two-stage regressions. DML relies on having a Neyman orthogonal (Neyman & Scott, 1965) score function to deal with regularisation bias, and uses cross-fitting, that is, an efficient form of (randomised) data splitting, to tackle overfitting bias. However, the use of DML for IV regression that utilises neural networks has not been explored.

In this work, we propose DML-IV, a novel IV regression algorithm that adopts the DML framework to provide an unbiased estimation of the causal function with fast convergence rate guarantees. We derive a novel Neyman orthogonal score for IV regression, and design a cross-fitting regime such that, under mild regularity conditions, our estimator is guaranteed to converge at the rate of $N^{-1/2}$, where N is the sample size. We then extend DML-IV to solve the offline IV bandit problem, where we derive a policy from the DML-IV estimator and provide a $O(N^{-1/2})$ suboptimality bound with high probability that matches the suboptimality bounds of *unconfounded* offline bandit algorithms (Jin et al., 2021; Nguyen-Tang et al., 2022). Finally, we evaluate DML-IV on multiple benchmarks for IV regression and offline IV bandits, where superior results are demonstrated compared to state-of-the-art (SOTA) methods.

Novel Contributions.

- We propose DML-IV, a novel IV regression algorithm that leverages the DML framework to provide unbiased estimation of the causal function.
- We derive a novel, Neyman orthogonal, score function for IV regression, and design a cross-fitting regime for the DML-IV estimator to mitigate the bias.
- We provide the first convergence rate guarantees for IV regression algorithms that use DL. Namely, we show that DML-IV converges at $N^{-1/2}$ rate leading to $O(N^{-1/2})$ suboptimality for the derived policy.
- On a range of IV regression and offline IV bandit benchmarks, including two real-world datasets, we experimentally demonstrate that DML-IV outperforms other SOTA methods.

1.1. Related Works

IV Regression. A number of approaches have been developed to extend the two-stage least squares (2SLS) algorithm (Angrist et al., 1996) to non-linear settings. A common approach is to use non-linear basis functions, such as Sieve IV (Newey & Powell, 2003; Blundell et al., 2007; Chen & Christensen, 2018), Kernel IV (Singh et al., 2019) and Dual IV (Muandet et al., 2020). These methods enjoy theoretical benefits, but their flexibility is limited by the set of basis functions. More recently, DFIV (Xu et al., 2020) proposed to use basis functions parameterised by DNNs, which remove the restrictions on the functional form. Another approach is to perform stage 1 regression through conditional density estimation (Darolles et al., 2011), where DeepIV (Hartford et al., 2017) adopts DNNs to perform these regressions. DeepGMM (Bennett et al., 2019) is a DNN-based method that is inspired by the Generalised Method of Moments (GMM) to find a causal function that ensures the regression residual and the instrument are independent. The learning procedure of DeepGMM does not offer stability comparable to 2SLS approaches, as it is based on solving a smooth zero-sum game, similar to training Generative Adversarial Networks (Goodfellow et al., 2014). Our approach allows DNNs in both stages and compares favourably to Deep IV, DeepGMM, Kernel IV and DFIV.

Double Machine Learning (DML). DML was originally proposed for semiparametric regression (Robinson, 1988); it relies on the derivation of a score function, which describes the regression problem that is Neyman orthogonal (Neyman & Scott, 1965). DML was later extended by adopting DNNs for generalised linear regressions (Chernozhukov et al., 2021). Its strength is that it provides unbiased estimations for causal effects when the causal effect is identifiable (Jung et al., 2021) or there are no hidden confounders (Chernozhukov et al., 2022b). DML offers strong ($N^{-1/2}$, where N is the size of the dataset) guarantees on the convergence rate, even in the presence of high-dimensional input.

There are previous works on combining DML with IV regression, but they are mainly focused on linear and partially linear models. Belloni et al. (2012) propose a method to use Lasso and then Post-Lasso methods for the first stage estimation of linear IV to estimate the optimal instruments. To avoid selection biases, Belloni et al. (2012) leverage techniques from weak identification robust inference. In addition, Chernozhukov et al. (2015) propose a Neyman-orthogonalised score for the linear IV problem with control

and instrument selection, to potentially be robust to regularisation and selection biases of Lasso as a model selection method. Neyman orthogonality for partially linear models with instruments was primarily discussed in the work of Chernozhukov et al. (2018). Furthermore, DML techniques for identifying the local average treatment effects (LATE) for nonlinear models with a binary instrument and treatment (action) have been explored before (Chernozhukov et al., 2024). For additional discussion, we refer to the book (Chernozhukov et al., 2024).

DML for semiparametric models (Chernozhukov et al., 2022a; Ichimura & Newey, 2022) has been previously applied to solve the nonparametric IV (NPIV) problem. However, their methods require that the average moment of the Neyman orthogonal score is affine (linear) in the nuisance parameters. Therefore, when applied to solve NPIV, functional assumptions regarding the IV set and the residual function were made. Such assumptions are not required in our work since we are considering a different problem setting and their Neyman orthogonal score is very different from ours. To the best of our knowledge, there is no work that adopts the DML framework for IV regression with DNNs.

Causality. Doubly robust estimation for causality problems predominantly revolved around the estimation of average treatment effects (ATE) (Robins et al., 1994; Funk et al., 2011; Benkeser et al., 2017; Bang & Robins, 2005; Słoczyński & Wooldridge, 2018). Recently, there has been a surge in doubly robust identification of causal structures beyond the ATE settings. Soleymani et al. (2022); Quinzan et al. (2023) focus on finding direct causes of the target variable by orthogonalised scores. Angelis et al. (2023) extend this line for testing Granger causality in the time-series domain. In this work, we focus on doubly robust estimation of the counterfactual prediction function, a central problem in the field of causal inference, which could be of independent interest beyond the IV settings.

Offline Bandit. Most bandit algorithms assume unconfoundedness (e.g., Nguyen-Tang et al. (2022); Subramanian & Ravindran (2022)). For bandit algorithms that consider hidden confounders, most of them work in the online setting, aiming to learn the best policy from scratch using the least amount of online interactions (Zhang & Bareinboim, 2020; Subramanian & Ravindran, 2022), or with the help of a pre-collected dataset (Lu et al., 2020). Few works are dedicated to the offline confounded bandit, where only the offline dataset is provided, as it is essentially a causal inference problem. However, offline reinforcement learning (RL) with hidden confounders has been studied. Pace et al. (2023) develop a pessimistic algorithm based on the Delphic uncertainty due to hidden confounders, while other methods adopt IV regression in combination with value iteration (Liao et al., 2021) and Actor-Critic methods (Li et al., 2021) to learn policies in offline RL. Offline policy evaluation (OPE) under hidden confounders has also been studied. Using IVs, doubly robust estimators for policy values are derived through efficient influence functions (Xu et al., 2023) and marginalised importance sampling (Fu et al., 2022). Bennett et al. (2021) solve OPE under an infinite-horizon ergodic MDP with hidden confounders using states and actions as proxies for the hidden confounders to identify policy values. Chen et al. (2021) consider the OPE problem in a standard unconfounded MDP, where they view the previous (action, state) pair as the instrument for the Bellman residual estimation problem of the current (action, state) pair and directly apply existing IV regression methods to estimate the Q value. We consider the setting of the offline confounded bandit with IVs, for which we leverage DML to obtain convergence and suboptimality guarantees.

2. Preliminaries

2.1. Notation

We use uppercase letters such as C to denote random variables. An observed realisation of C is denoted by a lowercase letter c. We abbreviate $\mathbb{E}[R|C = c]$, a realisation of the conditional expectation $\mathbb{E}[R|C]$, as $\mathbb{E}[R|c]$. $[N]$ denotes the set $\{1, ..., N\}$ for $N \in \mathbb{N}$. We write $\mathbb{E}[R|do(A = a)]$ for the expectation of R under *do* intervention (Pearl, 2000) of setting $A = a$. We use $\|\cdot\|_p$ to denote the functional norm, defined as $\|f\|_p \coloneqq \mathbb{E}[|f(C)|^p]^{1/p}$, where the measure is implicit from the context. For a function f, we use f_0 to denote the true function and $\hat{f}$ an estimator of the true function. We use O and o to denote big-O and little-o notations (Weisstein, 2023) respectively.

2.2. Contextual IV Setting

We begin with a description of the contextual IV setting (Hartford et al., 2017) that we use in this paper. We observe an *action* $A \in \mathcal{A} \subseteq \mathbb{R}^{d_A}$, a *context* $C \in \mathcal{C} \subseteq \mathbb{R}^{d_C}$, an *instrumental variable (IV)* $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$ and an *outcome* $R \in \mathbb{R}$, where there exist *unobserved confounders* that affect all of A, C and R through a hidden variable (or *noise*) ϵ. IV directly affects the action A, does not directly affect the outcome R and is not correlated with the hidden confounder ϵ. These causal relationships are illustrated in Figure 1 and are represented by the following structural causal model (Pearl, 2000):

$$R \coloneqq f_r(C, A) + \epsilon, \quad \mathbb{E}[\epsilon] = 0, \quad \mathbb{E}[\epsilon|A, C] \neq 0, \quad (1)$$

where f_r is an unknown, continuous, and potentially nonlinear causal function, and $\mathbb{E}[\epsilon|A, C]$ is not necessarily zero. Denote the set of observations (c_i, z_i, a_i, r_i), where $i \in [N]$, generated from this model as the *offline dataset* $\mathcal{D}$. The goal of this paper is to learn the *counterfactual prediction*

function (Hartford et al., 2017),

$$h_0(C, A) := f_r(C, A) + \mathbb{E}[\epsilon|C] = \mathbb{E}[R|do(A), C],$$

which is the expected outcome under $do(A)$ intervention conditional on C, from the offline dataset $\mathcal{D}$. This task is also known as *IV regression*, and we aim to estimate h_0 using a DNN. The term $\mathbb{E}[\epsilon|C]$ is typically nonzero[1], but learning h_0 still allows us to compare between different actions when given a context as $h_0(C, a_1) - h_0(C, a_2) = f_r(C, a_1) - f_r(C, a_2)$ for all $a_1, a_2 \in \mathcal{A}$, and in particular, $\arg\max_a h_0(C, a) = \arg\max_a f_r(C, a)$.

Generally, h_0 is allowed to be infinite-dimensional, as commonly seen in nonparametric IV literature (Newey & Powell, 2003). We also allow h_0 to be infinite-dimensional for the Neyman orthogonal score introduced in Section 3.1, but later, in Section 3.2, we restrict h_0 to be finite-dimensional and parameterised to obtain the theoretical results of the convergence rate and the suboptimality bound of $O(N^{-1/2})$.

The challenge of learning h_0 from $\mathcal{D}$ is that $\mathbb{E}[\epsilon|C, A] \neq 0$, which reflects the existence of hidden confounders that obscure the true causal effect. It has been shown (Bareinboim & Pearl, 2012) that we cannot learn the causal effect of actions in the presence of hidden confounders without structural assumptions. Fortunately, IVs enable the identification of h_0 if the following assumptions hold:

Assumption 2.1. (a) ϵ is additive to R and $\mathbb{E}[\epsilon] = 0$; (b) $\mathbb{E}[\epsilon|C, Z] = 0$; and (c) $\mathbb{P}(A|C, Z)$ is not constant in Z.

Intuitively, Assumption 2.1 (a) and (b), introduced by Newey & Powell (2003), is known as the exclusion restriction, and requires that the instrument Z is uncorrelated with the hidden confounder ϵ. Assumption 2.1 (c), known as the relevance condition, ensures that Z induces variation in action and should be satisfied by the data generation policy. These assumptions are standard for the IV setting (Newey & Powell, 2003; Xu et al., 2020; Singh et al., 2019), and allow for the minimal condition to identify the causal effect.

2.3. Offline IV Bandit

The learnt estimator of h_0 from the offline dataset $\mathcal{D}$ can be used to solve the *offline bandit problem in the contextual IV setting* (Zhang et al., 2022), that is, to identify a (deterministic) policy $\pi : \mathcal{C} \to \mathcal{A}$ that maximises the value $V(\pi) := \mathbb{E}_{c\sim\mathbb{P}_{\text{test}}}[R|do(A = \pi(c)), c] = \mathbb{E}_{c\sim\mathbb{P}_{\text{test}}}[h_0(c, \pi(c))]$, which is the expected outcome when performing actions following π. $\mathbb{P}_{\text{test}}$ is a test context distribution that can potentially differ from the distribution of $\mathcal{D}$. The optimal policy π^* should satisfy $V(\pi^*) = \max_\pi V(\pi)$, and suboptimality is defined as $\text{subopt}(\pi) := V(\pi^*) - V(\pi)$.

[1] In the setting where $\mathbb{E}[\epsilon|C] = 0$ is assumed (Bennett et al., 2019; Xu et al., 2020), $h_0 = f_r$ and all our results apply.

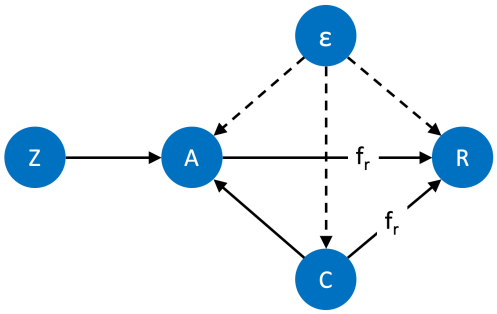


Figure 1: The causal graph of the contextual IV setting, where $R = f_r(C, A) + \epsilon$ and Z is an *instrumental variable* that affects R only through A.

We see that the optimal policy π^* can be retrieved from h_0 by selecting $\pi^*(c) = \arg\max_{a\in\mathcal{A}} h_0(c, a)$.

2.4. Two-Stage IV Regression

In order to identify h_0, a key observation (Newey & Powell, 2003) is that, by taking the expectation on both sides of Eq. (1) conditional on (C, Z), we have

$$\begin{aligned}\mathbb{E}[R|C, Z] &= \mathbb{E}\Big[f_r(C, A) + \mathbb{E}[\epsilon|C]\Big|C, Z\Big] \\ &= \mathbb{E}[h_0(C, A)|C, Z] \\ &= \int h_0(C, A)\mathbb{P}(A|C, Z)dA,\end{aligned} \tag{2}$$

where the expectation $\mathbb{E}[R|C, Z]$ and the distribution $\mathbb{P}(A|C, Z)$ are both observable. However, solving this equation analytically is ill-posed (Nashed & Wahba, 1974). This is an inverse problem for definite integrals that requires the derivation of a function inside the definite integral based on numerical integration values, which is thus not solvable analytically. Recent IV regression methods instead estimate $\hat{h}$ in some space of continuous functions $\mathcal{H}$ by solving the following optimisation problem with a two-stage approach:

$$\min_{h\in\mathcal{H}} \mathbb{E}[(R - \mathbb{E}[h(C, A)|C, Z])^2]. \tag{3}$$

In the first stage, the conditional expectation $\mathbb{E}[h(C, A)|c, z]$ is learnt as a function of (c, z) using observations, and in the second stage, the loss in Eq. (3) is minimised using the estimator obtained in stage 1. In both stages, linear regression or parametric ML methods, such as DNN, can be used to learn the true functions.

2.5. Double Machine Learning

DML is a parameter estimation method that can mitigate certain biases in the learning process (Chernozhukov et al., 2018; 2021; 2022b), which has been extended to work with ML methods, including DL. DML considers the problem

of estimating a function of interest h as a solution to an equation of the form

$$\mathbb{E}[\psi(\mathcal{D}; h, \eta)] = 0, \tag{4}$$

where ψ is referred to as a score function. Here, η is a nuisance parameter, which is of no direct interest, but must be estimated to obtain h. DML provides a set of tools to derive an unbiased estimator of h with convergence rate guarantees, even when the nuisance parameter η suffers from regularisation, overfitting and other type of biases present in the training of ML models, which typically causes slow convergence when learning h.

In order to estimate h, DML reduces biases by using score functions ψ that are Neyman orthogonal (Neyman & Scott, 1965) in η, which require the Gateaux derivative

$$\frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}[\psi(\mathcal{D}; h_0, \eta_0 + r\eta)] = 0, \tag{5}$$

for all η. Here, h_0 and η_0 are the true parameters that minimise the expected score, that is, $\mathbb{E}[\psi(\mathcal{D}; h_0, \eta_0)] = 0$. Intuitively, the condition in Eq. (5) is met if small changes of the nuisance parameter do not significantly affect the score function around the true parameter h_0. Neyman orthogonality is key in DML, as it allows fast convergence guarantees for h, even if the estimator for the nuisance parameter η is biased. For score functions that are Neyman orthogonal, we define DML with *K-fold cross-fitting* as follows.

Definition 2.2 (DML, Definition 3.2 (Chernozhukov et al., 2018))**.** Given a dataset $\mathcal{D}$ of N observations, consider a score function ψ as in Eq. (4), and suppose that ψ is Neyman orthogonal that satisfies Eq. (5). Take a *K-fold* random partition $\{I_k\}_{k=1}^K$ of observation indices $[N]$ each with size $n = N/K$, and let $\mathcal{D}_{I_k}$ be the set of observations $\{\mathcal{D}_i : i \in I_k\}$. Furthermore, define $I_k^c := [N] \setminus I_k$ for each fold k, and construct estimators $\hat{\eta}_k$ of the nuisance parameter using $\mathcal{D}_{I_k^c}$. Then, construct an estimator $\hat{h}$ as a solution to the equation

$$\frac{1}{K}\sum_{k=1}^{K} \hat{\mathbb{E}}_k[\psi(\mathcal{D}_{I_k}; \hat{h}, \hat{\eta}_k)] = 0, \tag{6}$$

where $\hat{\mathbb{E}}_k$ is the empirical expectation over $\mathcal{D}_{I_k}$.

In the definition above, $\hat{h}$ is defined as a solution to Eq. (6). In practice, however, finding an exact solution may not be feasible. To circumvent this problem, we can also define the estimator of interest $\hat{h}$ as an ϵ_N-approximate solution to Eq. (6), where $\epsilon_N = O(N^{-1/2})$, which allows for a small optimisation error.

3. DML-IV Algorithm

We now present the main contributions of this paper. The key to our results is the DML-IV algorithm, a novel two-stage IV regression algorithm utilising DNNs in both stages that provides guarantees on the convergence rate by leveraging the DML framework (see Section 2.5). The DML-IV estimator is then utilised to solve an offline IV bandit (see Section 2.3) by retrieving a deterministic policy with suboptimality guarantees that match those of the unconfounded bandit.

Firstly, we remark that, in order to estimate the counterfactual prediction function h_0 with convergence rate guarantees, we need a Neyman orthogonal score. We let $g_0(h, c, z) := \mathbb{E}[h(C, A)|c, z]$ and let $\mathcal{G}$ to be some function space that includes g_0 and its potential estimators $\hat{g}$. Unfortunately, the standard score (or loss) function for two-stage IV regression $\ell = (R - g(h, c, z))^2$ in Eq. (3) is not Neyman orthogonal (details in Appendix B), which means that small misspecifications or bias on g may lead to significant changes to this loss function, and there are no guarantees on the convergence rate if the first stage estimator $\hat{g}$ is naively plugged into the loss to estimate h_0. To address this, we first derive a novel Neyman orthogonal score function for the IV regression problem and then design a DML algorithm with K-fold cross-fitting adapted to the IV regression problem.

3.1. Neyman Orthogonal Score

We first derive a novel Neyman orthogonal score for learning h_0 in the contextual IV setting. The key to constructing a Neyman orthogonal score usually involves estimating additional nuisance parameters (Chernozhukov et al., 2018) and adding terms to the original score function to debias it, so we first select relevant quantities that should be estimated as nuisance parameters. Following two-stage IV regression approaches (Hartford et al., 2017), estimating g_0 is essential for identifying h_0, so we will estimate it as a nuisance parameter. We found that, by additionally estimating $s_0(c, z) := \mathbb{E}[R|c, z]$ inside some function space $\mathcal{S}$, we can construct a new score function

$$\psi(\mathcal{D}; h, (s, g)) = (s(c, z) - g(h, c, z))^2, \tag{7}$$

by replacing R in the standard score with $s(c, z)$. Here, the nuisance parameters are $\eta = (s, g)$. We see that ψ is a valid score function since $\mathbb{E}[\psi(\mathcal{D}; h_0, (s_0, g_0))] = 0$ with the true functions (s_0, g_0) by Eq. (2), and the next theorem shows that our score function is in fact Neyman orthogonal by checking its Gateaux derivative vanishes at $(h_0, (s_0, g_0))$, where the proof is deferred to Appendix C.1.

Theorem 3.1. *The score function* $\psi(\mathcal{D}; h, (s, g)) = (s(c, z) - g(h, c, z))^2$ *obeys the Neyman orthogonality conditions at* $(h_0, (s_0, g_0))$.

This Neyman orthogonal score function is abstract, in the sense that it allows for general estimation methods for g_0 and s_0, as long as they satisfy certain convergence conditions, which are introduced in the next section.

Algorithm 1 DML-IV with K-fold cross-fitting

Input: Dataset $\mathcal{D}$ of size N, number of folds K for cross-fitting, mini-batch size n_b
Output: The DML-IV estimator $h_{\hat{\theta}}$
Get a partition $(I_k)_{k=1}^K$ of dataset indices $[N]$
for $k = 1$ **to** K **do**
  $I_k^c := [N] \setminus I_k$
  Learn $\hat{s}_k$ and $\hat{g}_k$ using $\{(\mathcal{D}_i) : i \in I_k^c\}$
end for
Initialise $h_{\hat{\theta}}$
repeat
  for $k = 1$ **to** K **do**
    Sample n_b data (c_i^k, z_i^k) from $\{(\mathcal{D}_i) : i \in I_k\}$
    $\mathcal{L} = \hat{\mathbb{E}}_{(c_i^k, z_i^k)} \left[(\hat{s}_k(c,z) - \hat{g}_k(h_\theta, c, z))^2\right]$
    Update $\hat{\theta}$ to minimise loss $\mathcal{L}$
  end for
until convergence

3.2. Learning Causal Effects through DML

With the Neyman orthogonal score, we now introduce DML-IV. While the DML-IV algorithm does not require any assumptions on h, we assume that h is finite-dimensional and parameterised for the theoretical analysis of DML-IV. Let $h_0 = h_{\theta_0}$ and $\Theta \subseteq \mathbb{R}^{d_\theta}$ be a compact space of parameters of h, where the true parameter $\theta_0 \in \Theta$ is in the interior of Θ, and $\mathcal{H} := \{h_\theta : \theta \in \Theta\}$ is the function space of h. The procedure of the DML-IV algorithm for estimating h_0 is described in Algorithm 1. Given a dataset $\mathcal{D}$ of size N, we split the dataset using a random partition $\{I_k\}_{k=1}^K$ of dataset indices $[N]$ such that the size of each fold I_k is N/K.

In the first stage of DML-IV, for each fold $k \in [K]$, we learn $\hat{s}_k$ and $\hat{g}_k$ using data $\mathcal{D}_{I_k^c}$ with indices $I_k^c := [N] \setminus I_k$. $\hat{s}_k \approx \mathbb{E}[R|C, Z]$ can be learnt through standard supervised learning using a neural network with inputs (C, Z) and label R. For $\hat{g}_k$, we follow (Hartford et al., 2017) to estimate $F_0(A|C, Z)$, the conditional distribution of A given (C, Z), with $\hat{F}$, and then estimate $\hat{g}$ via

$$\hat{g}(h, c, z) = \sum_{\dot{A} \sim \hat{F}(A|C,Z)} h(C, \dot{A}) \approx \int h(C, A)\hat{F}(A|C, Z)dA \approx \mathbb{E}[h(C, A)|c, z].$$

If the action space is discrete, $\hat{F}$ is a categorical model, e.g., a DNN with softmax output. For a continuous action space, a mixture of Gaussian models is adopted to estimate the distribution $F_0(A|C, Z)$, where a DNN is used to predict the means and standard deviations of the Gaussian distributions.

In the second stage of DML-IV, we estimate $\hat{\theta}$ using our Neyman orthogonal score function ψ in Eq. (7). The key here is to optimise $\hat{\theta}$ with data from the k-th fold using nuisance parameters $\hat{s}_k$, $\hat{g}_k$ that are trained with data $\mathcal{D}_{I_k^c}$, the complement of the data from the k-th fold. This is important to fully debias the estimator $\hat{\theta}$. We alternate between the K folds while sampling a mini-batch (c_i^k, z_i^k) of size n_b from each fold k of the dataset to update $\hat{\theta}$ by minimising the empirical loss on the mini-batch following our Neyman orthogonal score ψ,

$$\hat{\mathbb{E}}_{(c_i^k, z_i^k)} \left[(\hat{s}_k(c,z) - \hat{g}_k(h_\theta, c, z))^2\right] = \sum_{(c_i^k, z_i^k)} \frac{1}{n_b} \left((\hat{s}_k(c,z) - \hat{g}_k(h_\theta, c, z))^2\right).$$

When the second stage converges, we return the DML-IV estimator $h_{\hat{\theta}}$.

To obtain the DML convergence rate guarantees (Chernozhukov et al., 2018) for $h_{\hat{\theta}}$, i.e., for $\hat{\theta}$ to converge to the true parameters θ_0 at the rate of $O(N^{-1/2})$ with high probability, there are two key conditions: i) Neyman orthogonality of the score function, and ii) the nuisance parameters should converge to their true values at the crude rate of $o(N^{-1/4})$. The Neyman orthogonal score is given in Theorem 3.1, so it remains to prove the convergence rate of the nuisance parameters. Define $\mathcal{G}_N$ to be the *realisation set* such that $\hat{g}_N$, the estimator of g_0 using a dataset of size N, takes values in this set. Similarly, define $\mathcal{S}_N$ to be the *realisation set* of $\hat{s}_N$. These realisation sets are properly shrinking neighbourhoods of the true functions g_0 and s_0, and we later provide Lemma 3.3 that describes the rate of shrinkage of these realisation sets, for which we require boundedness of functions g, s, h and the outcome variable R as stated in Assumption 3.2.

Assumption 3.2. We assume that (a): $g_0, s_0, h_0 \in \mathcal{G}, \mathcal{S}, \mathcal{H}$ are all bounded i.e., $\|g_0\|_\infty, \|s_0\|_\infty, \|h_0\|_\infty \leq B$; and (b): the outcome $\|R\|_\infty \leq B$, where $B \in \mathbb{R}^+$.

To improve readability, we provide here an informal statement of the lemma, which expresses the relationship between the critical radius (Wainwright, 2019; Bartlett et al., 2005) of the realisation sets and the convergence rate of the nuisance parameters. We defer the formal statement and the proof to Appendix C.1.

Lemma 3.3 (Informal: nuisance parameters convergence[2]). *If Assumption 3.2 holds, let δ_N be an upper bound on the critical radius of the function spaces related to the realisation sets $\mathcal{S}_N$ and $\mathcal{G}_N$. Then, with probability $1 - \zeta$:*

$$\|\hat{s} - s_0\|_2^2 = O\left(\delta_N^2 + \frac{\ln(1/\zeta)}{N}\right);$$
$$\|\hat{g} - g_0\|_2^2 = O\left(\delta_N^2 + \frac{\ln(1/\zeta)}{N}\right).$$

[2]See Lemma C.2 for the formal statement.

The critical radius is a quantity that describes the complexity of estimation, and it is typically shown that $\delta_N = O(d_N N^{-1/2})$ (Chernozhukov et al., 2022b; 2021), where d_N is the effective dimension of the hypothesis space (see Appendix C.3 for the derivation and formal definitions). This, together with Lemma 3.3, implies that $\|\hat{s} - s_0\|_2 = O(d_N N^{-1/2})$. Therefore, for function classes with $d_N = o(N^{1/4})$, $\|\hat{s} - s_0\|_2 \leq o(N^{-1/4})$ (and similarly for $\hat{g}$). This is a broad class of functions that covers many machine learning methods such as deep ReLU networks and shallow regression trees (Chernozhukov et al., 2021). It has also been shown that conditional density and expectation estimation used for $\hat{g}$ satisfies $d_N = o(N^{1/4})$ under mild assumptions (Grünewälder, 2018; Bilodeau et al., 2021). We refer to Chernozhukov et al. (2021) for additional discussion and concrete convergence rates of nuisance estimators.

Lemma 3.3 shows that the nuisance parameters converge to their true values at the rate of $o(N^{-1/4})$ if $d_N = o(N^{1/4})$, thus satisfying the second key condition to get the DML convergence rate guarantees. This allows us, after checking some mild regularity and continuity conditions, to obtain the following theorem regarding the convergence of the DML-IV estimator by applying Theorem 3.3 of Chernozhukov et al. (2018), with proof deferred to Appendix C.1.

Theorem 3.4 (Convergence of the DML-IV estimator)**.** *If the effective dimension $d_N = o(N^{1/4})$ for $\hat{s}$, $\hat{g}$, and Assumption 2.1, & 3.2 hold, we have that the DML-IV estimator $\hat{\theta}$ is concentrated in a $1/\sqrt{N}$ neighbourhood of θ_0, and is approximately linear and centred Gaussian:*

$$\sqrt{N}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, \sigma^2) \text{ in distribution},$$

where the estimator variance is given by

$$\sigma^2 := J_0^{-1}\mathbb{E}[\psi(\mathcal{D}, \theta_0, \eta_0)\psi(\mathcal{D}, \theta_0, \eta_0)^T](J_0^{-1})^T,$$

which is constant w.r.t N and J_0 denotes the Jacobian matrix of $\mathbb{E}[\psi]$ w.r.t θ.

Theorem 3.4 states that, with adequately trained nuisance parameter estimators, the estimator error $\hat{\theta} - \theta_0$ is normally distributed and variance shrinks at the rate of $N^{-1/2}$. This implies that $\hat{\theta}$ converges to θ_0 at the rate $O(N^{-1/2})$ with high probability, which allows us to deduce suboptimaltiy bounds for the policy induced by $h_{\hat{\theta}}$ in the next section.

3.3. Suboptimality Bounds

From the DML-IV estimator $h_{\hat{\theta}}$, we retrieve (an estimate of) the induced optimal policy as $\hat{\pi}(c) := \arg\max_a h_{\hat{\theta}}(c, a)$. Recall that the suboptimality of a policy is subopt$(\hat{\pi}) := V(\pi^*) - V(\hat{\pi})$. Next, we show a suboptimality bound for the DML-IV policy in terms of the sample size N.

Theorem 3.5 (Suboptimality Bounds)**.** *Let the learnt policy from a dataset of size N be $\hat{\pi}(c) := \arg\max_a h_{\hat{\theta}}(c, a)$, where $\hat{\theta}$ is the DML-IV estimator. Let L be a constant such that $|h_\theta(C, A) - h_{\theta'}(C, A)| \leq L\|\theta - \theta'\|$ for all C in the support of $\mathbb{P}_{test}$, $A \in \mathcal{A}$, and $\theta, \theta' \in \Theta$. Then, for all $\zeta \in (0, 1]$, we have that the suboptimality of $\hat{\pi}$ satisfies*

$$\textit{subopt}(\hat{\pi}) = O\left(L\sqrt{\frac{\ln(1/\zeta)}{N}}\right),$$

with probability $1 - \zeta$.

The proof is deferred to Appendix C.2. To the best of our knowledge, this is the first time that the convergence rate and suboptimality bounds of $O(N^{-1/2})$ have been proved for IV regression methods that use DL, matching the suboptimality bounds of the unconfounded bandit. On the other hand, most other DL-based IV regression methods only demonstrate that their estimators converge in the limit.

4. Experimental Results

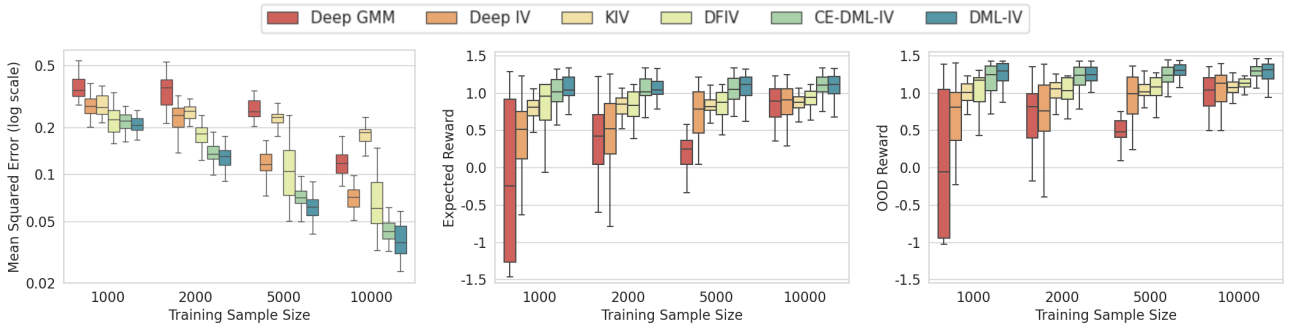
In this section, we empirically evaluate DML-IV for IV regression and offline IV bandit problems. In addition, we evaluate a computationally efficient version of DML-IV, referred to as CE-DML-IV, which does not apply K-fold cross-fitting. It trains $\hat{s}$ and $\hat{g}$ only once (instead of K times) using the entire dataset, and can also be considered as an ablation study on K-fold cross-fitting. Without K-fold cross-fitting, it lacks the theoretical convergence rate guarantees but it still enjoys the partial debiasing effect (Mackey et al., 2018) from the Neyman orthogonal score and trades off computational complexity with bias. We found that CE-DML-IV empirically performs as well as standard DML-IV on low-dimensional datasets. We provide details and discussion regarding CE-DML-IV in Appendix A.

Our evaluation considers both low- and high-dimensional contexts, as well as semi-synthetic real-world datasets. We compare our methods with leading modern IV regression methods Deep IV (Hartford et al., 2017), DeepGMM (Bennett et al., 2019), KIV (Singh et al., 2019) and DFIV (Xu et al., 2020). In this section we use DNN estimators for both stages with network architecture and hyper-parameters provided in Appendix F. Additional results of DML-IV using tree-based estimators such as Random Forests and Gradient Boosting are provided in Appendix G.2, where SOTA performance is also demonstrated. The algorithms are implemented using PyTorch (Paszke et al., 2019), and the code is available on GitHub[3].

4.1. Aeroplane Ticket Demand Dataset

We first conduct experiments for IV regression on the aeroplane ticket demand dataset, which is a synthetic dataset introduced by Hartford et al. (2017) that is now a standard

[3] https://github.com/shaodaqian/DML-IV

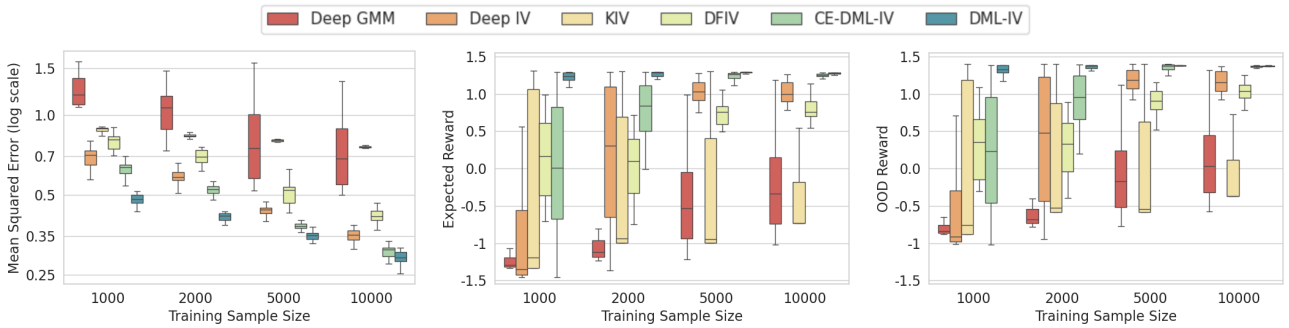


(a) The mean squared error of $\hat{h}$.

(b) The average reward following the policy $\hat{\pi}$ derived from $\hat{h}$.

(c) The average reward following $\hat{\pi}$ with out of training distribution context.

Figure 2: Results on the aeroplane ticket demand dataset with low-dimensional context.



(a) The mean squared error of $\hat{h}$.

(b) The average reward following the policy $\hat{\pi}$ derived from $\hat{h}$.

(c) The average reward following $\hat{\pi}$ with out of training distribution context.

Figure 3: Results on the aeroplane ticket demand dataset with high-dimensional context.

benchmark for nonlinear IV methods. In this dataset, we aim to understand how ticket prices p affect ticket sales r. We observe two context variables, which are the time of year $t \in [0, 10]$ and customer type $s \in [7]$ variables, the latter categorised by the level of price sensitivity. Price and context affect sales through $h_0((t, s), p) = 100 + (10 + p) \cdot s \cdot \psi(t) - 2p$, where $\psi(t)$ is a complex nonlinear function. However, the noise of r and p is correlated, which indicates the existence of unobserved confounders. The fuel price z is introduced as an instrumental variable. Details of this dataset are included in Appendix D.1.

The results for learning h_0 with this dataset of various sizes are provided in Figure 2a. We ran each method 20 times and report the mean squared errors (MSE) between the estimators $\hat{h}$ and h_0, where the median, 25th and 75th percentiles are shown. It can be seen that DML-IV performs better than other IV regression methods for all dataset sizes. CE-DML-IV, which requires significantly less computation, matches the performance of DML-IV in this case.

High-Dimensional Feature Space

In real applications, we typically do not observe variables such as the customer type as explicit categories. Therefore, we follow Hartford et al. (2017) and consider the case where the customer type $s \in [7]$ is replaced by images of the corresponding handwritten digits from the MNIST dataset (LeCun & Cortes, 2010) to evaluate our methods with high-dimensional (28^2=784 dimensions) inputs. The task remains to learn h_0, but the algorithms are no longer explicitly given the 7 customer types, and instead have to infer the relationship between the image data and the outcome. Results for IV regression are plotted in Figure 3a, where DML-IV and CE-DML-IV outperforms all other methods. In these high-dimensional settings, regularisation is heavily used to avoid overfitting. DML-IV demonstrates the benefits of using DML to reduce both the regularisation and overfitting bias caused by learning the nuisance parameters.

To demonstrate the robustness of DML-IV, we first provide a sensitivity analysis against hyperparameter changes in Appendix G.3. We evaluate DML-IV and CE-DML-IV on the aeroplane ticket demand datasets under a range of hyperparameters, where stable performance is observed. In addition, we consider the case when the IV is weakly correlated with the action in Appendix G.1, where we empirically demon-

strate that DML-IV and CE-DML-IV perform significantly better than SOTA methods under weak instruments.

4.2. Offline IV Bandit

We also evaluate DML-IV's ability to learn good decision policies in the offline IV bandit problem. We reuse the aeroplane ticket demand dataset and aim to find the best pricing policy that maximises sales. From the learnt $\hat{h}$, for each context sampled from the test distribution, we retrieve the best action by uniformly sampling actions from the action space $\mathcal{A}$ and selecting the action for which $\hat{h}$ returns the highest value. Using this induced policy $\hat{\pi}$, we compare the expected reward following $\hat{\pi}$ over the test distribution.

For the low-dimensional ticket demand dataset, we first set the test distribution to be the same as the training distribution and plot the average rewards in Figure 2b. In Figure 2c, we shift the test distribution out of the training distribution by incrementing the distribution of t by 1. For the high-dimensional setting, Figure 3b and Figure 3c demonstrate the expected rewards for test distributions in and out of the training distribution, respectively. There is a clear trend that a better fitted (low MSE) $\hat{h}$ leads to an induced policy with higher expected reward. In all cases, DML-IV outperforms all other methods, especially in the high-dimensional setting, where DML-IV consistently learns the near-optimal policy with only 2000 samples. CE-DML-IV, on the other hand, only matches the performance of DML-IV for the low-dimensional setting, but still outperforms the other methods in the high-dimensional setting.

We only compare with other IV regression methods because there are no offline bandit methods that consider the IV setting, and standard offline bandit algorithms (e.g., (Valko et al., 2013; Jin et al., 2021; Nguyen-Tang et al., 2022)) fail to learn meaningful policies when the dataset is confounded, as demonstrated in Appendix E.

4.3. Real-World Decision Problem

Lastly, we test the performance of DML-IV on real-world datasets. The true counterfactual prediction function is rarely available for real-world data. Therefore, in line with previous approaches (Shalit et al., 2017; Wu et al., 2023; Schwab et al., 2019; Bica et al., 2020), we instead consider two semi-synthetic real-world datasets IHDP[4] (Hill, 2011) and PM-CMR[5] (Wyatt et al., 2020). We directly use the continuous variables from IHDP and PM-CMR as context variables, and generate the outcome variable with a nonlinear synthetic function following Wu et al. (2023). There are 470 and 1350 training samples in IHDP and PM-CMR, respectively (for details see Appendix D.2). We also run

[4]IHDP: https://www.fredjo.com/.

[5]PM-CMR:https://doi.org/10.23719/1506014.

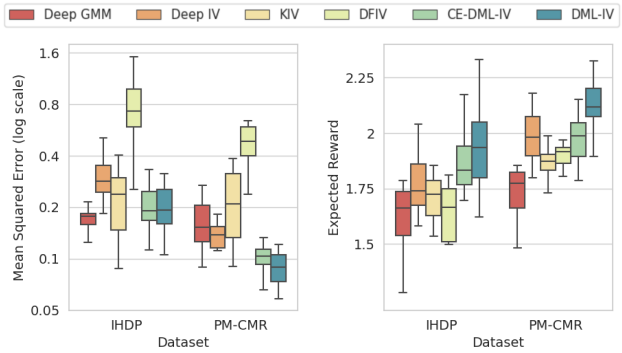


Figure 4: The mean squared error of $\hat{h}$ and average reward following $\hat{\pi}$ for the real-world datasets.

each method 20 times, where the MSE of $\hat{h}$ and the expected reward of the induced policy $\hat{\pi}$ on the test dataset are plotted in Figure 4. DML-IV and CE-DML-IV demonstrate comparable, if not lower, MSE of fitting $\hat{h}$ than the other methods, while outperforming all other methods in average reward. This shows that our algorithm can reliably learn the counterfactual prediction function and policies with the highest average reward from real-world data.

5. Conclusion

We have proposed a novel method for instrumental variable regression, DML-IV. By leveraging IVs and DML on offline data, DML-IV can learn counterfactual predictions and effective decision policies with fast convergence rate and suboptimality guarantees by mitigating the regularisation and overfitting biases of DL. We evaluated DML-IV on IV regression benchmarks and IV bandit problems, including semi-synthetic real-world data, experimentally showing it is superior compared to SOTA IV regression methods.

Future work includes considering other estimation methods for the nuisance parameters following our Neyman-orthogonal score, and extending the method to sequential decision problems and reinforcement learning in the presence of hidden confounders (Namkoong et al., 2020).

Acknowledgments

This work was supported by the EPSRC Prosperity Partnership FAIR (grant number EP/V056883/1). DS acknowledges funding from the Turing Institute and Accenture collaboration. AS was partially supported by AI Singapore, grant AISG2-RP-2020-018. FQ acknowledges funding from ELSA: European Lighthouse on Secure and Safe AI project (grant agreement No. 101070617 under UK guarantee). MK receives funding from the ERC under the European Union's Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115).

Impact Statement

The goal of the paper is to develop a methodology to learn high-performing decision policies from offline data. There are many applications of our work in automated decision making, for example, in planning, healthcare, and finance. The theoretical guarantees that we provide ensure the reliability and suboptimality guarantees of the learnt policies. We do not foresee negative implications of our methodology, but would caution against deploying it without human input and recommend additional validation in any new setting to reduce the risk of misapplication.

References

Andrews, I., Stock, J. H., and Sun, L. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11: 727–753, 8 2019. ISSN 19411391. doi: 10.1146/ANNUREV-ECONOMICS-080218-025643/1.

Angelis, E., Quinzan, F., Soleymani, A., Jaillet, P., and Bauer, S. Doubly robust structure identification from temporal data. *arXiv preprint arXiv:2311.06012*, 2023.

Angrist, J. D. Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, 80:1284–1286, 1990. ISSN 00028282.

Angrist, J. D. and Pischke, J.-S. *Mostly Harmless Econometrics*. Princeton University Press, 2 2009. doi: 10.2307/J.CTVCM4J72.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 6 1996. ISSN 1537274X. doi: 10.1080/01621459.1996.10476902.

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

Bareinboim, E. and Pearl, J. Causal inference by surrogate experiments: z-identifiability. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33:1497-1537, 2005.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6): 2369–2429, 2012.

Benkeser, D., Carone, M., Laan, M. V. D., and Gilbert, P. B. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.

Bennett, A., Kallus, N., Li, L., and Mousavi, A. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1999–2007, 3 2021. ISSN 2640-3498.

Bica, I., Jordon, J., and van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 2020-December, 2 2020. ISSN 10495258.

Bilodeau, B., Foster, D. J., and Roy, D. M. Minimax rates for conditional density estimation via empirical entropy. *Annals of Statistics*, 51:762–790, 9 2021. doi: 10.1214/23-AOS2270. URL http://arxiv.org/abs/2109.10461http://dx.doi.org/10.1214/23-AOS2270.

Blair, J. M., Edwards, C. A., and Johnson, J. H. Rational chebyshev approximations for the inverse of the error function. *Mathematics of Computation*, 30(136):827, 10 1976. ISSN 00255718. doi: 10.2307/2005402.

Blundell, R., Chen, X., and Kristensen, D. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75:1613–1669, 11 2007. ISSN 1468-0262. doi: 10.1111/J.1468-0262.2007.00808.X.

Blundell, R., Horowitz, J. L., and Parey, M. Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics*, 3:29–51, 3 2012. ISSN 1759-7331. doi: 10.3982/QE91.

Bound, J., Jaeger, D. A., and Baker, R. M. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90:443, 6 1995. ISSN 01621459. doi: 10.2307/2291055.

Chen, X. and Christensen, T. M. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9:39–84, 3 2018. ISSN 17597331. doi: 10.3982/qe722.

Chen, Y., Xu, L., Gulcehre, C., Paine, T. L., Gretton, A., de Freitas, N., and Doucet, A. On instrumental variable regression for deep offline policy evaluation. *Journal of Machine Learning Research*, 23, 5 2021. ISSN 15337928.

Chernozhukov, V., Hansen, C., and Spindler, M. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–490, 2015.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. ISSN 1368-4221. doi: 10.1111/ECTJ.12097.

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V., and Syrgkanis, V. Automatic debiased machine learning via neural nets for generalized linear regression. 4 2021. URL https://arxiv.org/abs/2104.14737v1.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 7 2022a. ISSN 0012-9682. doi: 10.3982/ecta16294.

Chernozhukov, V., Newey, W., Quintas-Martínez, V., and Syrgkanis, V. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. *Proceedings of Machine Learning Research*, 162: 3901–3914, 10 2022b. ISSN 26403498.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. Applied causal inference powered by ml and ai. *rem*, 12(1):338, 2024.

Darolles, S., Fan, Y., Florens, J. P., and Renault, E. Nonparametric instrumental regression. *Econometrica*, 79:1541–1565, 9 2011. ISSN 1468-0262. doi: 10.3982/ECTA6539.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. Offline reinforcement learning with instrumental variables in confounded markov decision processes. 2022.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63:139–144, 6 2014. ISSN 15577317. doi: 10.1145/3422622.

Grünewälder, S. Plug-in estimators for conditional expectations and probabilities. *Proceedings of the 21 International Conference on Artificial Intelligence and Statistics*, pp. 1513–1521, 3 2018. ISSN 2640-3498.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*, 2017. doi: 10.5555/3305381.3305527.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240, 3 2011. ISSN 10618600. doi: 10.1198/JCGS.2010.08162.

Ichimura, H. and Newey, W. K. The influence function of semiparametric estimators. *Quantitative Economics*, 13: 29–61, 1 2022. ISSN 1759-7331. doi: 10.3982/QE826.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? *International Conference on Machine Learning*, 2021.

Jung, Y., Tian, J., and Bareinboim, E. Estimating identifiable causal effects through double machine learning. *AAAI Conference on Artificial Intelligence*, 2021.

LeCun, Y. and Cortes, C. Mnist handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist/.

Li, J., Luo, Y., Zhang, X., Ai, C., Chernozhukov, V., Dai, J., Fernandez-Val, I., Forneron, J.-J., Jiang, W., and Kaido, H. Causal reinforcement learning: An instrumental variable approach. *SSRN Electronic Journal*, 3 2021. doi: 10.2139/ssrn.3792824.

Liao, L., Fu, Z., Yang, Z., Wang, Y., Kolar, M., and Wang, Z. Instrumental variable value iteration for causal offline reinforcement learning. 2021. doi: CoRRabs/2102.09907.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*, 11 2017.

Lu, Y., Meisami, A., Tewari, A., and Yan, Z. Regret analysis of bandit problems with causal background knowledge. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 10 2020.

Mackey, L., Syrgkanis, V., and Zadik, D. Orthogonal machine learning: Power and limitations. *35th International Conference on Machine Learning, ICML 2018*, 13:9112–9124, 11 2018.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 2020-December, 10 2020. ISSN 10495258.

Namkoong, H., Keramati, R., Yadlowsky, S., and Brunskill, E. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

Nashed, M. Z. and Wahba, G. Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. *SIAM Journal on Mathematical Analysis*, 5, 1974.

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71: 1565–1578, 9 2003. ISSN 1468-0262. doi: 10.1111/1468-0262.00459.

Neyman, J. and Scott, E. L. Asymptotically optimal tests of composite hypotheses for randomized experiments with noncontrolled predictor variables. *Journal of the American Statistical Association*, 60:699–721, 1965. ISSN 1537274X. doi: 10.1080/01621459.1965.10480822.

Nguyen-Tang, T., Gupta, S., Nguyen, A. T., and Venkatesh, S. Offline neural contextual bandits: Pessimism, optimization and generalization. *Proceeding of the International Conference on Learning Representations*, 2022.

Pace, A., Eche, H. Y., Schölkopf, B., Rätsch, G., and Tennenholtz, G. Delphic offline reinforcement learning under nonidentifiable hidden confounding. *Workshop on New Frontiers in Learning, Control, and Dynamical Systems at ICML*, 6 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 12 2019. ISSN 10495258.

Pearl, J. Causality: models, reasoning, and inference. *Econometric Theory*, 2000.

Quinzan, F., Soleymani, A., Jaillet, P., Rojas, C. R., and Bauer, S. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*, pp. 28468–28491, 2023.

Reiersöl, O. Confluence analysis by means of instrumental sets of variables. *astronomi och fysik*, 1945.

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

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56:931, 7 1988. ISSN 00129682. doi: 10.2307/1912705.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., and Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 5612–5619, 2 2019. doi: 10.1609/aaai.v34i04.6014.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. *34th International Conference on Machine Learning, ICML 2017*, 6:4709–4718, 6 2017.

Shpitser, I. and Pearl, J. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008. ISSN 1533-7928.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 6 2019. ISSN 10495258.

Słoczyński, T. and Wooldridge, J. M. A general double robustness result for estimating average treatment effects. *Econometric Theory*, 34(1):112–133, 2018.

Soleymani, A., Raj, A., Bauer, S., Schölkopf, B., and Besserve, M. Causal feature selection via orthogonal search. *Transactions on Machine Learning Research*, 2022.

Subramanian, C. and Ravindran, B. Causal contextual bandits with targeted interventions. In *International Conference on Learning Representations*, 1 2022.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, pp. 654–663, 9 2013.

Van Handel, R. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

Wainwright, M. J. High-dimensional statistics: A non-asymptotic viewpoint. *Cambridge University Press*, pp. 1–552, 1 2019. doi: 10.1017/9781108627771.

Weisstein, E. W. Asymptotic notation, 2023. URL https://mathworld.wolfram.com/AsymptoticNotation.html.

Wright, P. G. The tariff on animal and vegetable oils. *https://doi.org/10.1086/254144*, 38:619–620, 10 1928. ISSN 0022-3808. doi: 10.1086/254144.

Wu, A., Kuang, K., Xiong, R., Zhu, M., Liu, Y., Li, B., Liu, F., Wang, Z., and Wu, F. Learning instrumental variable from data fusion for treatment effect estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 8 2023.

Wyatt, L. H., Peterson, G. C. L., Wade, T. J., Neas, L. M., and Rappold, A. G. Annual pm2.5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 us counties. *Data in brief*, 30, 6 2020. ISSN 2352-3409. doi: 10.1016/J.DIB.2020.105318.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. Learning deep features in instrumental variable regression. *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020.

Xu, Y., Zhu, J., Shi, C., Luo, S., and Song, R. An instrumental variable approach to confounded off-policy evaluation. *Proceedings of the 40th International Conference on Machine Learning*, 2023.

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

Zhang, J. and Bareinboim, E. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. *Proceedings of the 37th International Conference on Machine Learning*, pp. 119, 2020.

Zhang, J., Chen, Y., Allen, P. G., and Singh, A. Causal bandits: Online decision-making in endogenous settings. *A causal view on dynamical systems workshop at NeurIPS 2022*, 2022.

A. Computationally Efficient CE-DML-IV

Algorithm 2 Computationally Efficient CE-DML-IV

Input: Dataset $\mathcal{D}$ with size N, mini-batch size n_b
Output: The CE-DML-IV estimator $h_{\hat{\theta}}$
Learn $\hat{s}$ and $\hat{g}$ using $\mathcal{D}$
Initialise $h_{\hat{\theta}}$
repeat
  Sample n_b data (c_i, z_i) from $\mathcal{D}$
  $\mathcal{L} = \hat{\mathbb{E}}_{(c_i,z_i)}\left[(\hat{s}(c,z) - \hat{g}(h_\theta, c, z))^2\right]$
  Update $\hat{\theta}$ to minimise loss $\mathcal{L}$
until convergence

The standard DML-IV with K-fold cross-fitting trains $\hat{s}$ and $\hat{g}$ K times on different subsets of the dataset to tackle overfitting bias, but it is computationally expensive. Therefore, as mentioned in Section 4, we also evaluate CE-DML-IV, a computationally efficient version of DML-IV that does not apply K-fold cross-fitting and trains $\hat{s}$ and $\hat{g}$ only once using the entire dataset. It uses the same Neyman orthogonal score as the standard DML-IV, so it still enjoys the partial debiasing effect (Mackey et al., 2018) from the Neyman orthogonal score. However, without K-fold cross-fitting, it lacks the theoretical convergence rate guarantees provided by Theorem 3.4 and Theorem 3.5. CE-DML-IV can be viewed as a trade-off between computational complexity and theoretical guarantees, and we found that CE-DML-IV empirically performs as well as standard DML-IV on low-dimensional datasets, where overfitting bias is not prevalent.

B. Standard Loss Function for IV Regression

The standard score (or loss) function for two-stage IV regression is $\ell = (R - g(h, c, z))^2$, as described in Eq. (3). This score is not Neyman orthogonal because, first of all, $\mathbb{E}[(R - g_0(h_0, c, z))^2] = \mathbb{E}[(R - \mathbb{E}[R|C, Z])^2] \neq 0$ since $\mathbb{E}[h_0|C, Z] = \mathbb{E}[R|C, Z]$ and $R - \mathbb{E}[R|C, Z] \neq 0$ due to the noise on R.

Secondly, the derivative against small changes in g for score $\mathbb{E}[(R - g_0(h_0, c, z))^2]$ is

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}\Big[(R - g_0(h_0, C, Z) - r \cdot g(h_0, C, Z))^2\Big] \\
&\quad = \frac{\partial}{\partial r}\mathbb{E}\Big[(R - g_0(h_0, C, Z))^2 - 2r \cdot (R - g_0(h_0, C, Z))g(h_0, C, Z) + r^2 \cdot g(h_0, C, Z)^2\Big] \\
&\quad = \mathbb{E}\Big[2(R - g_0(h_0, C, Z))g(h_0, C, Z) + 2r \cdot g(h_0, C, Z)^2\Big],
\end{aligned}$$

and, when $r = 0$, this derivative evaluates to

$$\mathbb{E}[2(R - g_0(h_0, c, z))g(h_0, c, z)] = \mathbb{E}[2(R - \mathbb{E}[R|C, Z])g(h_0, c, z)]$$

which does not equal to 0 for general $g \in \mathcal{G}$ since generally $g(h_0, c, z)$ and the residual $(R - \mathbb{E}[R|C, Z])$ are correlated. Therefore, the standard score function for two-stage IV regression can not be used to create a DML estimator.

C. Omitted Proofs

In this section, we state all the conditions required to prove the $N^{-1/2}$ convergence rate guarantees for the DML-IV estimator, and provide the omitted proofs in the main paper for Theorem 3.1, Lemma 3.3, Theorem 3.4 and Theorem 3.5.

C.1. DML-IV $N^{-1/2}$ Convergence Rate Guarantees

To obtain $N^{-1/2}$ convergence rate guarantees of the DML-IV estimator, the following conditions must be satisfied.

Condition C.1 (Conditions for $N^{-1/2}$ convergence of DML, Assumption 3.3 and 3.4 in Chernozhukov et al. (2018))**.** For $N \geq 3$, all the following conditions hold. (a): The true parameter θ_0 obeys $\mathbb{E}[\psi(\mathcal{D}; h_0, (s_0, g_0))] = 0$ and Θ contains a ball of radius $c_1 N^{-1/2} \log N$ centered at θ_0. (b): The map $(\theta, (s, g)) \mapsto \mathbb{E}[\psi(\mathcal{D}; h_\theta, (s, g))]$ is twice continuously

Gateaux-differentiable. (c): For all $\theta \in \Theta$, the identification relationship

$$\|\mathbb{E}[\psi(\mathcal{D}; h_\theta, (s_0, g_0))]\| \gtrsim \|J_0(\theta - \theta_0)\| \tag{8}$$

is satisfied, where $J_0 := \partial_{\theta'}\{\mathbb{E}[\psi(\mathcal{D}; h_{\theta'}, (s_0, g_0))]\}|_{\theta'=\theta_0}$ is the Jacobian matrix, with singular values strictly positive (bounded away from zero). (d): The score ψ obeys the Neyman orthogonality. (e): Let K be a fixed integer. Given a random partition $\{I_k\}_{k=1}^K$ of indices $[N]$ each of size $n = N/K$, we have that the nuisance parameter estimator $\hat{\eta}$ learnt using data with indices I_k^c belongs to a shrinking realisation set $\mathcal{T}_N$, and the nuisance parameters should be estimated at the $o(N^{-1/4})$ rate, i.e., $\|\hat{\eta} - \eta_0\|_2 = o(N^{-1/4})$. (f): All eigenvalues of the matrix $\mathbb{E}[\psi(\mathcal{D}; h_{\theta_0}, (s_0, g_0))\psi(\mathcal{D}; h_{\theta_0}, (s_0, g_0))^T]$ are strictly positive (bounded away from zero).

We will check all these conditions in Theorem 3.1, Lemma C.2 and Theorem 3.4.

*Proof of **Theorem 3.1**:* Firstly, by Equation 2, we have $s_0(C, Z) = g_0(h_0, C, Z)$, thus

$$\psi(\mathcal{D}; h_0, (s_0, g_0)) = \mathbb{E}\Big[(s_0(C, Z) - g_0(h_0, C, Z))^2\Big] = 0$$

Then we compute the derivative w.r.t. small changes in the nuisance parameters. For all $s, g \in \mathcal{S}, \mathcal{G}$,

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}\Big[(s_0(C, Z) + r \cdot s(C, Z) - g_0(h_0, C, Z) - r \cdot g(h_0, C, Z))^2\Big] \\
=&\frac{\partial}{\partial r}\mathbb{E}\Big[2r(s_0(C, Z) - g_0(h_0, C, Z))(s(C, Z) - g(h_0, C, Z)) + r^2(s(C, Z) - g(h_0, C, Z))^2\Big] \\
=&\mathbb{E}\Big[2(s_0(C, Z) - g_0(h_0, C, Z))(s(C, Z) - g(h_0, C, Z)) + 2r(s(C, Z) - g(h_0, C, Z))^2\Big],
\end{aligned}$$

and, when at $r = 0$, the derivative evaluates to

$$\begin{aligned}
&\mathbb{E}\Big[2(s_0(C, Z) - g_0(h_0, C, Z))(s(C, Z) - g(h_0, C, Z))\Big] \\
&\quad = \mathbb{E}\Big[0 \times (s(C, Z) - g(h_0, C, Z))\Big] \\
&\quad = 0 \quad \forall s, g \in \mathcal{S}, \mathcal{G},
\end{aligned}$$

since $s_0(C, Z) = \mathbb{E}[R|C, Z] = \mathbb{E}[h_0|C, Z] = g_0(h_0, C, Z)$. Therefore, our moment function ψ is Neyman orthogonal at $(h_0, (s_0, g_0))$. □

Lemma C.2 (Formal version of Lemma 3.3: Nuisances parameters convergence)**.** *If Assumption 3.2 holds, let δ_N be an upper bound on the critical radius of the two following function spaces:*

$$\{(C, Z) \mapsto \gamma(s(C, Z) - s_0(C, Z)) : s \in \mathcal{S}_N, \gamma \in [0, 1]\}; \tag{9}$$

$$\{(C, Z) \mapsto \gamma(g(C, Z, h_0) - g_0(C, Z, h_0)) : g \in \mathcal{G}_N, \gamma \in [0, 1]\}, \tag{10}$$

and suppose that all functions f in the two spaces above satisfy $\|f\|_\infty \leq B$ for some $B \in \mathbb{R}^+$. Then, for some universal constants c_1 and c_2, we have that with probability $1 - \zeta$:

$$\|\hat{s} - s_0\|_2^2 \leq c_1 \left(\delta_N^2 + \frac{B^2 \log(1/\zeta)}{N} + \inf_{s_* \in \mathcal{S}_N} \|s_* - s_0\|_2^2\right);$$

$$\|\hat{g} - g_0\|_2^2 \leq c_2 \left(\delta_N^2 + \frac{B^2 \log(1/\zeta)}{N} + \inf_{g_* \in \mathcal{G}_N} \|g_* - g_0\|_2^2\right).$$

*Proof of **Lemma C.2**:* We will mainly use the result from Theorem 1 of Chernozhukov et al. (2021), which states the following. For a function α that is the minimizer of a loss function that can be represented as $\mathbb{E}[-2m(\mathcal{D}, \alpha) + \alpha(x)^2]$, where $\mathcal{D}$ is the offline dataset and m is some moment function that satisfies

$$\mathbb{E}[(m(W, \alpha) - m(W, \alpha'))^2] \leq M\|\alpha - \alpha'\|_2^2 \quad \forall \alpha, \alpha' \in \mathcal{A}_N.$$

Let δ_N be an upper bound on the critical radius of the two function spaces:

$$\{W \mapsto \gamma(\alpha(W) - \alpha_0(W)) : \alpha \in \mathcal{A}_N, \gamma \in [0,1]\};$$
$$\{W \mapsto \gamma(m(W,\alpha) - m(W,\alpha_0)) : \alpha \in \mathcal{A}_N, \gamma \in [0,1]\}.$$

Then, if $\|\alpha\|_\infty \leq B$ for some $B \in \mathbb{R}^+$, there exists a universal constant c such that with probability $1-\zeta$,

$$\|\hat{\alpha} - \alpha_0\|_2^2 \leq c\left(\delta_N^2 + \frac{M\log(1/\zeta)}{N} + \inf_{\alpha_* \in \mathcal{A}_N} \|\alpha_* - \alpha_0\|_2^2\right).$$

In our case, we show that the loss function for both s and g satisfies the above conditions, and thus Theorem 1 of Chernozhukov et al. (2021) is applicable to provide an upper bound on the convergence rate of our nuisance parameters.

The loss function for $s \in \mathcal{S}_N$ is

$$\begin{aligned} s_0 &= \arg\min_{s \in \mathcal{S}} \mathbb{E}\left[(R - s(C,Z))^2\right] \\ &= \arg\min_{s \in \mathcal{S}} \mathbb{E}\left[R^2 - 2Rs(C,Z) + s(C,Z)^2\right] \\ &= \arg\min_{s \in \mathcal{S}} \mathbb{E}\left[-2Rs(C,Z) + s(C,Z)^2\right], \end{aligned}$$

where we can set $m(W,s) = Rs(C,Z)$ and check that

$$\begin{aligned} \mathbb{E}[(Rs(C,Z) - Rs_0(C,Z))^2] &\leq \mathbb{E}[R^2(s(C,Z) - s_0(C,Z))^2] \\ &\leq \|R^2\|_\infty \mathbb{E}[(s(C,Z) - s_0(C,Z))^2] \\ &= B^2\|s(C,Z) - s_0(C,Z)\|_2^2, \end{aligned}$$

by Hölder's inequality and the assumption that $\|R\|_\infty \leq B$. Therefore, by Theorem 1 of Chernozhukov et al. (2021), there exists a universal constant c_1 such that with probability $1-\zeta$,

$$\|\hat{s} - s_0\|_2^2 \leq c_1\left(\delta_N^2 + \frac{B^2\log(1/\zeta)}{N} + \inf_{s_* \in \mathcal{S}_N} \|s_* - s_0\|_2^2\right),$$

where recall δ_N is an upper bound on the critical radius of the function spaces defined in Eq. (9).

For the second part of the proof, recall that

$$g(h,c,z) = \int h(C,A)F(A \mid C,Z)dA,$$

where $F(A \mid C,Z)$ is some distribution over A and $F_0(A \mid C,Z) = \mathbb{P}(A \mid C,Z)$ is the distribution of A conditional on (C,Z). Therefore, g_0 should minimise the following loss:

$$\begin{aligned} g_0 &= \arg\min_{g \in \mathcal{G}} \mathbb{E}\left[\left(\int h_0(C,A)\mathbb{P}(A \mid C,Z)dA - g(C,Z,h_0)\right)^2\right] \\ &= \arg\min_{g \in \mathcal{G}} \mathbb{E}\left[-2\int h_0(C,A)\mathbb{P}(A \mid C,Z)dA \cdot g(C,Z,h_0) + g(C,Z,h_0)^2\right], \end{aligned}$$

where we can set $m(\mathcal{D}, g) = \int h_0(C, A)\mathbb{P}(A \mid C, Z)dA \cdot g(C, Z, h_0)$ and check that

$$\begin{aligned}
&\mathbb{E}\left[\left(\int h_0(C,A)\mathbb{P}(A \mid C,Z)dA \cdot g(C,Z,h_0) - \int h_0(C,A)\mathbb{P}(A \mid C,Z)dA \cdot g_0(C,Z,h_0)\right)^2\right] \\
=&\mathbb{E}\left[\left(\int h_0(C,A)\mathbb{P}(A \mid C,Z)dA\right)^2 \cdot (g(C,Z,h_0) - g_0(C,Z,h_0))^2\right] \\
=&\mathbb{E}\left[g_0(C,Z,h_0)^2 \cdot (g(C,Z,h_0) - g_0(C,Z,h_0))^2\right] \\
\leq&\|g_0^2\|_\infty \|g(C,Z,h_0) - g_0(C,Z,h_0)\|_2^2 \\
\leq&B^2\|g(C,Z,h_0) - g_0(C,Z,h_0)\|_2^2,
\end{aligned}$$

by Hölder's inequality, where M is a constant since g is bounded. Therefore, by Theorem 1 of Chernozhukov et al. (2021), there exists a universal constant c_2 such that with probability $1 - \zeta$,

$$\|g - g_0\|_2^2 \leq c_2 \left(\delta_N^2 + \frac{B^2 \log(1/\zeta)}{N} + \inf_{g_* \in \mathcal{G}_N} \|g_* - g_0\|_2^2\right),$$

where again δ_N is an upper bound on the critical radius of the function spaces defined in Equation 10, which completes the proof. □

Now, we are ready to prove Theorem 3.4, which is our main theorem that states the $N^{-1/2}$ convergence rate guarantees for the DML-IV estimator.

*Proof of **Theorem 3.4**:* We mainly use Theorem 3.3 from Chernozhukov et al. (2018), where properties of the DML estimator for non-linear scores are demonstrated. It states that, if Condition C.1 holds, the DML estimator $\hat{\theta}$ is concentrated in a $1/\sqrt{N}$ neighbourhood of θ_0:

$$\frac{\sqrt{N}}{\sigma}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum \bar{\psi}(\mathcal{D}_i) + O(\rho_N) \to \mathcal{N}(0, 1) \text{ in distribution},$$

where $\bar{\psi}(\cdot) := -\sigma^{-1} J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$ is the influence function, J_0 is the Jacobian of ψ, the approximate variance is $\sigma^2 := J_0^{-1}\mathbb{E}[\psi(\mathcal{D}, \theta_0, \eta_0)\psi(\mathcal{D}, \theta_0, \eta_0)^T](J_0^{-1})^T$, and the size of the remainder ρ_N converges to 0. Therefore, we only need to check whether, under Assumption 2.1 and 3.2, all of Condition C.1 for DML $N^{-1/2}$ convergence rate is satisfied. Conditions (a) and (d) are satisfied by Theorem 3.1. Condition (b) is satisfied since $(s - g)^2$ is twice continuously differentiable with respect to s and g.

Condition (c) is a sufficient identifiability condition, which states the closeness of the loss function at point θ to zero and implies the closeness of θ to θ_0. This assumption is standard in condition moment problems. To check condition (c), we first point out that under analytical assumptions for s, g, and h, we can write down first order Taylor series for the score function $\mathbb{E}[\psi(\mathcal{D}; h_\theta, (s_0, g_0))]$ around the point θ_0,

$$\mathbb{E}[\psi(\mathcal{D}; h_\theta, (s_0, g_0))] = \mathbb{E}[\psi(\mathcal{D}; h_{\theta_0}, (s_0, g_0))] + J_0(\theta - \theta_0) + O(\|\theta - \theta_0\|^2).$$

Plugging in validity of the score function $\psi(\mathcal{D}; h_\theta, (s_0, g_0))$, i.e., $\mathbb{E}[\psi(\mathcal{D}; h_{\theta_0}, (s_0, g_0))] = 0$, we infer that

$$\|\mathbb{E}[\psi(\mathcal{D}; h_\theta, (s_0, g_0))]\| \gtrsim \|J_0(\theta - \theta_0)\|.$$

Now for identifiability, we only need to assume that $J_0 J_0^T$ is non-singular, which is a common technical assumption.

Condition (e) is satisfied since we have that the effective dimension $d_N = o(N^{1/4})$, and together with Lemma C.2 and the fact that the upper bound of the critical radius $\delta_N = O(d_N N^{-1/2})$ (see Appendix C.3), the nuisance parameters converge sufficiently quickly to ensure $\|\hat{s} - s_0\|_2 \leq O(\delta_N + N^{-1/2}) = O(d_N N^{-1/2}) = o(N^{-1/4})$ and $\|\hat{g} - g_0\|_2 \leq O(\delta_N + N^{-1/2}) = O(d_N N^{-1/2}) = o(N^{-1/4})$.

Condition (f) is the non-degeneracy assumption for covariance of the score function $\psi(\mathcal{D}; h_\theta, (s_0, g_0))$. By definition,

$$\mathbb{E}[\psi(\mathcal{D}; h_\theta, (s_0, g_0))\psi(\mathcal{D}; h_\theta, (s_0, g_0))^T] = \int \psi(\mathcal{D}; h_\theta, (s_0, g_0))\psi(\mathcal{D}; h_\theta, (s_0, g_0))^T d\mathbb{P}(\mathcal{D}).$$

By trace trick, for each datapoint $\mathcal{D}$, the only eigenvalue of $\psi(\mathcal{D}; h_\theta, (s_0, g_0))\psi(\mathcal{D}; h_\theta, (s_0, g_0))^T$ is $\|\psi(\mathcal{D}; h_\theta, (s_0, g_0))\|^2 \geq 0$, with $\psi(\mathcal{D}; h_\theta, (s_0, g_0))$ as the corresponding eigenvector. Therefore, $\mathbb{E}[\psi(\mathcal{D}; h_\theta, (s_0, g_0))\psi(\mathcal{D}; h_\theta, (s_0, g_0))^T]$ is positive-definite if for each member d of the support of $\mathbb{P}$, which is the distribution of $\mathcal{D}$, there are at least as many eigenvectors of d as the number of dimension of $\psi(\mathcal{D}; h_\theta, (s_0, g_0))$, which is true in our setting as the co-domain of $\psi(\mathcal{D}; h_\theta, (s_0, g_0))$ is $\mathbb{R}$.

Therefore, all conditions for Theorem 3.3 (Chernozhukov et al., 2018) to hold are satisfied, which concludes the proof. □

C.2. Suboptimaltiy

*Proof of **Theorem 3.5**:* From theorem 3.4, we have that the parameters $\hat{\theta}$ for $h_{\hat{\theta}}$ learned from a dataset of size N using DML-IV satisfy $(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2/N)$, where σ^2 is the is the DML-IV estimator variance. This means that, for all $\epsilon > 0$ and $\zeta > 0$, there exists an integer $K > 0$ such that for all $N \geq K$,

$$\mathbb{P}(\|\hat{\theta} - \theta_0\| > \epsilon) \leq 1 - \Phi\left(\epsilon \cdot \sqrt{N}/\sigma\right) + \zeta/2,$$

where Φ is the CDF of a standard Gaussian distribution. If we assume L to be a constant such that $|h_\theta(C, A) - h_{\theta'}(C, A)| \leq L\|\theta - \theta'\|$ for all $C, A \in \text{supp}^M(C, A)$ and $\theta \in \Theta$, we have that for all $\epsilon > 0$ and $\zeta > 0$, there exists an integer $K > 0$ such that for all $N \geq K$,

$$\mathbb{P}(|h_{\hat{\theta}}(C, A) - h_{\theta_0}(C, A)| > L \cdot \epsilon) \leq 1 - \Phi(\epsilon \cdot \sqrt{N}/\sigma) + \zeta/2 \quad \forall C, A \in \text{supp}^M(C, A). \tag{11}$$

Next, we can show that the suboptimality of $\hat{\pi}$ satisfies

$$\begin{aligned}
\text{subopt}(\hat{\pi}) &= V(\pi^*) - V(\hat{\pi}) \\
&= \mathbb{E}_{C\sim\mathbb{P}_{\text{test}}}[R \mid C, do(A = \pi^*(c))] - \mathbb{E}_{C\sim\mathbb{P}_{\text{test}}}[R \mid C, do(A = \hat{\pi}(c))] \\
&= \mathbb{E}_{C\sim\mathbb{P}_{\text{test}}}[f_r(C, \pi^*(C)) - f_r(C, \hat{\pi}(C))] \\
&= \mathbb{E}_{C\sim\mathbb{P}_{\text{test}}}[h(C, \pi^*(C)) - h(C, \hat{\pi}(C))] \\
&\leq \max_{c\in\text{supp}(\mathbb{P}_{\text{test}})} (h(c, \pi^*(c)) - h(c, \hat{\pi}(c))) \\
&\leq \max_{c\in\text{supp}(\mathbb{P}_{\text{test}})} |h(c, \pi^*(c)) - h_{\hat{\theta}}(c, \pi^*(c))| + (h_{\hat{\theta}}(c, \pi^*(c)) - h_{\hat{\theta}}(c, \hat{\pi}(c))) + |h_{\hat{\theta}}(c, \hat{\pi}(c)) - h(c, \hat{\pi}(c))| \\
&\leq 2L \cdot \epsilon \quad \text{with probability } \left(\Phi(\epsilon \cdot \sqrt{N}/\sigma) - \zeta/2\right)
\end{aligned} \tag{12}$$

where $\text{supp}(\mathbb{P}_{\text{test}})$ is the support of $\mathbb{P}_{\text{test}}$, by Equation 11 and the fact that $h_{\hat{\theta}}(C, \pi^*(C)) - h_{\hat{\theta}}(C, \hat{\pi}(C)) \leq 0$. Setting $\Phi(\epsilon \cdot \sqrt{N}/\sigma) = 1 - \zeta/2$ in Equation 12 and substituting ϵ yields

$$\text{subopt}(\hat{\pi}) \leq 2L\Phi^{-1}(1 - \zeta/2)\sigma/\sqrt{N} \quad \text{with probability } 1 - \zeta.$$

From Blair *et al.*'s approximation for the inverse of the error function (erf) (Blair et al., 1976), we have that for all $y \in (0, 1]$, $\Phi^{-1}(1 - y) \leq \sqrt{-2\ln(y)}$. Thus, we conclude that there exists $K > 0$ such that for all $N > K$

$$\text{subopt}(\hat{\pi}_N) \leq 2\sqrt{2}L\sigma\sqrt{\frac{\ln(2/\zeta)}{N}} \quad \text{with probability } 1 - \zeta,$$

which completes the proof. □

C.3. Critical Radius and Effective Dimension

Definition C.3 (Wainwright (2019))**.** The critical radius denoted by δ_N is defined as the minimum δ that satisfies the following upper bound on the local Gaussian complexity of a star-shaped function class $\mathcal{F}^*$[6], $\mathcal{G}(\mathcal{F}^*, \delta) \leq \delta^2/2$, where local Gaussian complexity is defined as

$$\mathcal{G}(\mathcal{F}^*, \delta) = \mathbb{E}_\epsilon[\sup_{g\in\mathcal{F}^*:\|g\|_N\leq\delta} \langle\epsilon, g\rangle],$$

with ϵ being a random i.i.d. zero-mean Gaussian vector.

[6]A function class $\mathcal{F}$ is star-shaped if for every $f \in \mathcal{F}$ and $\alpha \in [0, 1]$, we have $\alpha f \in \mathcal{F}$.

The critical radius is a standard notion to bound the estimation error in the regression problem. Since local Gaussian complexity can be viewed as an expected value of a supremum of a stochastic process indexed by g, we can apply empirical process theory tools, namely the Dudley's entropy integral (Wainwright, 2019; Van Handel, 2014), to provide a bound on the critical radius,

$$\mathcal{G}(\mathcal{F}^*, \delta) \leq \inf_{\alpha \geq 0} \left\{ \alpha + \frac{1}{\sqrt{N}} \int_{\alpha/4}^{\delta} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_N), \epsilon)}\, d\epsilon \right\},$$

where $\mathcal{N}(\mathcal{F}^*, L^2(P_N), \epsilon)$ is the ϵ-covering number of function class $\mathcal{F}^*$ in $L^2(P_N)$ norm. Now, by placing $\alpha = 0$, when the integral is a single scale value of $\sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \epsilon)}$, we infer that

$$\mathcal{G}(\mathcal{F}^*, \delta) \leq \frac{\delta}{\sqrt{N}} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_N), \epsilon)}.$$

Thus, the critical radius will be upper bounded by

$$\delta_N \lesssim \frac{\sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_N), \epsilon)}}{\sqrt{N}} = O(d_N N^{-1/2}).$$

Chernozhukov et al. (2022b; 2021) referred to $d_N = \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_N), \epsilon)}$ as the effective dimension of the hypothesis space. Note that this matches the minimax lower bound of fixed design estimation for this setting (Yang & Barron, 1999).

D. Datasets Details

In this section, we provide details of the datasets considered in this paper.

D.1. Aeroplane Ticket Demand Dataset

Here, we describe the aeroplane ticket demand dataset, first introduced by Hartford et al. (2017). The observable variables are generated by the following model:

$$\begin{aligned} r &= h_0((t,s),p) + \epsilon, \quad \mathbb{E}[\epsilon | t, s, p] = 0; \\ p &= 25 + (z+3)\psi(t) + \omega, \end{aligned}$$

where r is the ticket sales (as the outcome variable) and p is the ticket price (as the action variable). (t, s) are observed context variables, where t is the time of year and s is the customer type. The fuel price z is introduced as an instrumental variable, which only affects the ticket price p. The noises ϵ and ω are correlated with correlation $\rho \in [0, 1]$, where in our experiments we set $\rho = 0.9$. h_0 is the true counterfactual prediction function, defined as

$$\begin{aligned} h_0((t,s),p) &= 100 + (10+p) \cdot s \cdot \psi(t) - 2p, \\ \psi(t) &= 2\left(\frac{(t-5)^4}{600} + \exp(-4(t-5)^2) + \frac{t}{10} - 2 \right), \end{aligned}$$

where $\psi(t)$ is a complex non-linear function of t plotted in Figure 5. The offline dataset is sampled with the following distributions:

$$\begin{aligned} s &\sim \text{Unif}\{1, ..., 7\} \\ t &\sim \text{Unif}(0, 10) \\ z &\sim \mathcal{N}(0, 1) \\ \omega &\sim \mathcal{N}(0, 1) \\ \epsilon &\sim \mathcal{N}(\rho\omega, 1 - \rho^2). \end{aligned}$$

From the observations (r, p, t, s, z), we estimate $\hat{h}$ using IV regression methods, and the mean squared error between $\hat{h}$ and the true causal function h_0 are computed on 10000 random samples from the above model. For the out of distribution test samples, we sample $t \sim \text{Unif}(1, 11)$ instead.

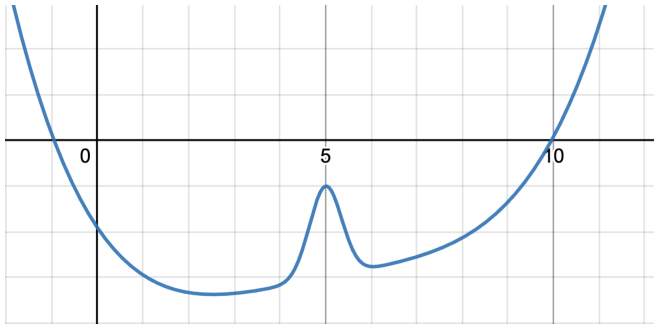


Figure 5: A graph of the nonlinear function $\psi(t)$ in the aeroplane ticket demand dataset.

We standardise the action and outcome variables p and r to centre the data around a mean of zero and a standard deviation of one following Hartford et al. (2017). This is standard practice for DNN training, which improves training stability and optimization efficiency.

High-Dimensional Setting

For the high-dimensional setting, we again follow Hartford et al. (2017) to replace the customer type $s \in [7]$ in the low-dimensional setting with images of the corresponding handwritten digits from the MNIST dataset (LeCun & Cortes, 2010). For each digit $d \in [7]$, we select a random MNIST image from the digit class d as the new customer type variable s. The images are $28 \times 28 = 784$ dimensional.

D.2. Real-World Datasets

Following previously studied causal inference methods (Shalit et al., 2017; Wu et al., 2023; Schwab et al., 2019; Bica et al., 2020), we consider two semi-synthetic real-world datasets IHDP[7] (Hill, 2011) and PM-CMR[8] (Wyatt et al., 2020) for experiments, since the true counterfactual prediction function is rarely available for real-world datasets.

IHDP, the Infant Health and Development Program (IHDP), comprises 747 units with 6 pre-treatment continuous variables, one action variable and 19 discrete variables related to the children and their mothers, aiming at evaluating the effect of specialist home visits on the future cognitive test scores of premature infants. From the original data, We select all 6 continuous covariance variables as our context variable C.

PM-CMR studies the impact of PM2.5 particle level on the cardiovascular mortality rate (CMR) in 2132 counties in the United States using data provided by the National Studies on Air Pollution and Health (Wyatt et al., 2020). We use 6 continuous variables about CMR in each city as our context variable C.

Following Wu et al. (2023), from the context variables C obtained from real-world datasets, we generate the instrument Z, the action A and the outcome R using the following model:

$$
\begin{aligned}
&Z \sim \mathbb{P}(Z = z) = 1/K, \quad z \in [1..K]; \\
&A = \sum_{z=1}^{K} 1_{Z=z} \sum_{i=1}^{d_C} w_{iz}(C_i + 0.2\epsilon + f_z(z)) + \delta_A, \quad w_{iz} \sim \text{Unif}(-1, 1); \\
&R = 9A^2 - 1.5A + \sum_{i=1}^{d_C} \frac{C_i}{d_C} + |C_1 C_2| - \sin(10 + C_2 C_3) + 2\epsilon + \delta_R,
\end{aligned}
$$

where C_i denotes the i-th variable in C, f_z is a function that returns different constants depending on the input z, $\delta_R, \delta_A \sim \mathcal{N}(0, 1)$ and $\epsilon \sim \mathcal{N}(0, 0.1)$ is the unobserved confounder. The fully generated semi-synthetic datasets IHDP and PM-CMR have 747 and 2132 samples respectively, and we randomly split them into training (63%), validation (27%), and testing (10%) following Wu et al. (2023).

[7]IHDP: https://www.fredjo.com/.
[8]PM-CMR:https://doi.org/10.23719/1506014.

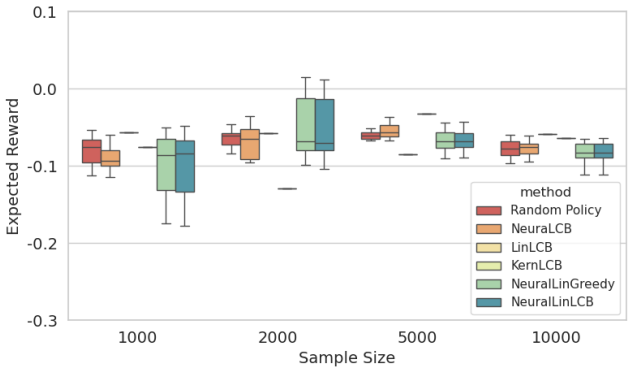


Figure 6: Comparing the average reward obtained by policies learned using offline bandit algorithms that do not take IVs into account with a random policy on the aeroplane ticket demand dataset with low-dimensional context.

E. Failure of Standard Offline Bandit Algorithms

It has been demonstrated that standard supervised learning that does not take IVs into account fails to learn the causal function or the counterfactual prediction function from a confounded offline dataset (Hartford et al., 2017). Similarly, we demonstrate here that standard offline bandit algorithms also fail to learn meaningful policies from confounded offline datasets. We evaluate PEVI, also called LinLCB (Jin et al., 2021), NeuraLCB (Nguyen-Tang et al., 2022), KernLCB (Valko et al., 2013), NeuralLinLCB (Nguyen-Tang et al., 2022) and NeuralLinGreedy (Nguyen-Tang et al., 2022) algorithms, for which we combine the context C and instrument Z variables together as the new context input for these offline bandit algorithms. For algorithms that only support discrete actions, we discretise the action space $\mathcal{A}$ into 20 discrete actions.

For all methods, we follow the network architecture and hyper parameters from the original papers, and we adopt the implementation[9] of Nguyen-Tang et al. (2022). We evaluate these methods on the aeroplane ticket demand dataset described in Appendix D.1 and compare the average reward obtained by the learned policies with a random policy in Figure 6. It can be seen that all the offline bandit algorithms do not outperform a random policy while DML-IV achieves an average reward higher then 1 as shown in Figure 2b. This is unsurprising because these bandit methods do not exploit IVs explicitly and are unable to learn the true causal effect of actions.

F. Network Structures and Hyper-Parameters

Here, we describe the network architecture and hyper-parameters of all experiments. Unless otherwise specified, all neural network algorithms are optimised using AdamW (Loshchilov & Hutter, 2017) with learning rate $= 0.001$, $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. In addition, we set $K = 10$ for K-fold cross-fitting in DML-IV.

F.1. Aeroplane Ticket Demand Dataset

For DML-IV and CE-DML-IV, we use the network architecture described in Table 1. We use a learning rate of 0.0002 with a weight decay of 0.001 (L2 regularisation) and a dropout rate of $\frac{1000}{5000+N}$ that depends on the data size N. For DeepGMM, we use the same structure as the outcome network of DML-IV with dropout $= 0.1$ and the same learning rate as DML-IV. For DFIV, we follow the original structure proposed in Xu et al. (2020) with regularisers $\lambda 1$, $\lambda 2$ both set to 0.1 and weight decay of 0.001. For DeepIV, we use the same network architectures as action network and stage 2 network for DML-IV, with the dropout rate in Hartford et al. (2017) and weight decay of 0.001. For KIV, we use the Gaussian kernel, where the bandwidth is determined by the median trick as originally described by Singh et al. (2019), and we use the random Fourier feature trick with 100 dimensions.

[9] https://github.com/thanhnguyentang/offline_neural_bandits

Table 1: Network architecture for DML-IV and CE-DML-IV for the aeroplane ticket demand low-dimensional dataset. For the input layer, we provide the input variables. For mixture of Gaussians output, we report the number of components. The dropout rate is given in the main text.

(a) Action Network for $\hat{g}$

Layer Type	Configuration
Input	C, Z
FC + ReLU	in:3 out:128
Dropout	-
FC + ReLU	in:128 out:64
Dropout	-
FC + ReLU	in:64 out:32
Dropout	-
MixtureGaussian	10

(b) Outcome Network for $\hat{s}$

Layer Type	Configuration
Input	C, Z
FC + ReLU	in:3 out:128
Dropout	-
FC + ReLU	in:128 out:64
Dropout	-
FC + ReLU	in:64 out:32
Dropout	-
FC	in:32 out:1

(c) Stage 2 Network for $\hat{h}$

Layer Type	Configuration
Input	C, A
FC + ReLU	in:3 out:128
Dropout	-
FC + ReLU	in:128 out:64
Dropout	-
FC + ReLU	in:64 out:32
Dropout	-
FC	in:32 out:1

F.2. Aeroplane Ticket Demand with MNIST

For DML-IV and CE-DML-IV, we use a convolutional neural network (CNN) feature extractor, which we denote as *ImageFeature*, described in Table 2, for all networks. The full network architecture is described in Table 3; we use weight decay of 0.05. For DeepGMM, we use the same structure as the outcome network of DML-IV, with a dropout rate of 0.1 and weight decay of 0.05. For DFIV, we follow the original structure proposed in Xu et al. (2020) with regularisers $\lambda 1$, $\lambda 2$ both set to 0.1 and weight decay of 0.05. For DeepIV, we use the same network architecture as the action network and stage 2 network for DML-IV, with the dropout rate in Hartford et al. (2017) and weight decay of 0.05. For KIV, we use the Gaussian kernel, where the bandwidth is determined by the median trick as originally described by Singh et al. (2019), and we use the random Fourier feature trick with 100 dimensions.

F.3. IHDP and PM-CMR

For the two real-world datasets, we use the same network architectures described in Table 1 as in the low-dimensional ticket demand setting, where the input dimension is increased to 7 for all networks. We use a dropout rate of 0.1 and weight decay of 0.001. For DeepGMM, we use the same structure as the outcome network of DML-IV with dropout $= 0.1$. For DFIV, we also use the same network architectures as in the low dimensional ticket demand setting with regularisers $\lambda 1$, $\lambda 2$ both set to 0.1 and weight decay of 0.001. For DeepIV, we use the same network architectures as the action network and stage 2 network of DML-IV, with a dropout rate of 0.1 and weight decay of 0.001. For KIV, we use the Gaussian kernel where the bandwidth is determined by the median trick as originally described by Singh et al. (2019), and we use the random Fourier feature trick with 100 dimensions.

F.4. Valiadation and Hyper-Parameter Tuning

Validation procedures are crucial for tuning DNN hyper-parameters and optimizer parameters. All the DML-IV and CE-DML-IV training stages can be validated by simply evaluating the respective losses on held-out data, as discussed

Table 2: Network architecture of the feature extractor used for the aeroplane ticket demand dataset with MNIST. For each convolution layer, we list the kernel size, input dimension and output dimension, where s stands for stride and p stands for padding. For max-pooling, we provide the size of the kernel. The dropout rate here is set to 0.3. We denote this feature extractor as *ImageFeature*.

Layer Type	**Configuration**
Input	28×28
Conv + ReLU	$3 \times 3 \times 32$, s:1, p:0
Max Pooling	2×2, s:2
Dropout	-
Conv + ReLU	$3 \times 3 \times 64$, s:1, p:0
Max Pooling	2×2, s:2
Dropout	-
Conv + ReLU	$3 \times 3 \times 64$, s:1, p:0
Dropout	-
FC + ReLU	in: 576, out:64

Table 3: Network architecture for DML-IV and CE-DML-IV for the aeroplane ticket demand dataset with MNIST. For the input layer, we provide the input variables. For a mixture of Gaussians output, we report the number of components. The dropout rate is given in the main text.

(a) Action Network for $\hat{g}$

Layer Type	**Configuration**
Input	ImageFeature$(C), Z$
FC + ReLU	in:66 out:32
Dropout	-
MixtureGaussian	10

(b) Outcome Network for $\hat{s}$

Layer Type	**Configuration**
Input	ImageFeature$(C), Z$
FC + ReLU	in:66 out:32
Dropout	-
FC	in:32 out:1

(c) Stage 2 Network for $\hat{h}$

Layer Type	**Configuration**
Input	ImageFeature$(C), A$
FC + ReLU	in:66 out:32
Dropout	-
FC	in:32 out:1

in Hartford et al. (2017). This allows independent validation and hyperparameter tuning of the two first stage networks (the action and the outcome networks), and perform second stage validation using the best network selected in the first stage. This validation procedure guards against the 'weak instruments' bias (Bound et al., 1995) that can occur when the instruments are only weakly correlated with the actions variable (see detailed discussion in Hartford et al. (2017)).

G. Additional Experimental Results

In this section, we provide additional experimental results including the effects of weak IVs, performance with tree-based estimators, and a hyperparameter sensitivity analysis.

G.1. Effects of Weak Instruments

When the correlation between instruments and the endogenous variable (the action in our case) is weak, IV regression methods generally become unreliable (Andrews et al., 2019) because the weak correlation induces variance and bias in the first stage estimator thus induces bias in the second stage estimator, especially for non-linear IV regressions. In theory, DML-IV should be more resistant to biases in the first stage thanks to the DML framework, as long as the causal effect

IV Strength	1.0	0.8	0.6	0.4	0.2	0.01
DML-IV	**0.0676(0.0116)**	**0.0984(0.0161)**	**0.1295(0.0168)**	**0.1859(0.0376)**	**0.2899(0.0494)**	**0.4872(0.1295)**
CE-DML-IV	**0.0765(0.0119)**	**0.1064(0.0120)**	**0.1514(0.0203)**	**0.2070(0.0329)**	**0.3194(0.0572)**	**0.5302(0.1625)**
DeepIV	0.1213(0.0209)	0.2039(0.0269)	0.3051(0.0415)	0.4476(0.0656)	0.6891(0.1210)	0.9293(0.2382)
DFIV	0.1124(0.0481)	0.1586(0.0320)	0.3080(0.1907)	0.8117(0.2779)	0.9622(0.3892)	1.6503(0.6845)
DeepGMM	0.2699(0.0522)	0.3330(0.1171)	0.4762(0.1056)	0.8666(0.2248)	1.0056(0.4334)	2.0218(0.6555)
KIV	0.2312(0.0272)	0.3149(0.0218)	0.4275(0.0368)	0.6646(0.0538)	0.8099(0.0657)	1.226(0.1014)

Table 4: Results for the low-dimensional ticket demand dataset when the IV is weakly correlated with the action.

is identifiable under the weak instrument. Under this identifiability condition, Lemma 3.3, Theorem 3.4 and 3.5 all hold, and the convergence rate guarantees still apply. However, while causal identifiability with weak instruments are studied theoretically in the linear setting (Andrews et al., 2019), such a theoretical study for non-linear IV models, to the best of our knowledge, does not exist due to the difficulty of analyzing non-linear models and estimators.

Experimentally, for the airplane ticket demand dataset, we alter the instrument strength by changing how much the instrument z affects the price p. Recall from Appendix D.1 that $p = 25 + (z + 3)\psi(t) + \omega$, where ψ is a nonlinear function and ω is the noise. We add an IV strength parameter ϱ such that $p = 25 + (\varrho \cdot z + 3)\psi(t) + \omega$. In Table 4, we present the mean and standard deviation of the MSE of $\hat{h}$ for various IV strengths ϱ from 0.01 to 1 and sample size $N = 5000$. It is very interesting to see that DML-IV indeed performs significantly better than SOTA nonlinear IV regression methods under weak instruments.

G.2. Performance of DML-IV with tree-based estimators

The DML-IV framework allows for general estimators following the Neyman orthogonal score function. While deep learning is flexible and widely used in SOTA non-linear IV regression methods, Gradient Boosting and Random Forests regression are all good candidate estimators for DML-IV. In addition, as discussed in Lemma 3.3, the convergence rate and suboptimality guarantees in Theorem 3.4 and 3.5 both hold for these tree-based regressions.

Empirically, we replace the DNN estimators in DML-IV, CE-DML-IV and DeepIV with Random Forests and Gradient Boosting regressors (using scikit-learn implementation). DeepIV is a good baseline for comparison, since it optimizes directly using a non-Neyman-orthogonal score and allows for direct replacement of all DNN estimators with tree-based estimators. We use 500 trees for both regressors, with minimum samples required at each leaf node of 100 for the nuisance parameters and 10 for $\hat{h}$.

In Table 5, we present the mean and standard deviation of the MSE of $\hat{h}$ with Random Forests and Gradient Boosting estimators on the aeroplane ticket demand dataset with various dataset sample sizes. The results demonstrate the benefits of our Neyman orthogonal score function, and interestingly the performance of Gradient Boosting is comparable to DNN estimators.

G.3. Sensitivity analysis for different Hyperparameters

The tunable hyperparameters in DML-IV are the learning rate, network width, weight decay and dropout rate (see Appendix F). As a sensitivity analysis, we provide results for the mean and standard deviation of the MSE of the DML-IV estimator $\hat{h}$ with different hyperparameter values for both the low-dimensional and high-dimensional datasets with sample size N=5000 in Table 6 and Table 7. Overall, we see that DML-IV is not very sensitive to small changes of the hyperparameters.

	Dataset Size	DNN (results in the paper)	Random Forests	Gradient Boosting
DML-IV	2000	**0.1308(0.0206)**	0.1689(0.0172)	**0.1301(0.0112)**
CE-DML-IV	2000	**0.1410(0.0246)**	0.1733(0.0198)	**0.1329(0.0125)**
DeepIV	2000	0.2388(0.0438)	0.2642(0.0261)	0.2052(0.0232)
DML-IV	5000	**0.0676(0.0129)**	0.1067(0.0131)	**0.0632(0.0107)**
CE-DML-IV	5000	**0.0765(0.0119)**	0.1154(0.0138)	**0.0699(0.0069)**
DeepIV	5000	0.1213(0.0209)	0.1626(0.0128)	0.1020(0.0091)
DML-IV	10000	**0.0378(0.0094)**	0.0657(0.0062)	**0.0482(0.0079)**
CE-DML-IV	10000	**0.0442(0.0070)**	0.0721(0.0039)	**0.0523(0.0059)**
DeepIV	10000	0.0714(0.0140)	0.1106(0.0080)	0.1017(0.0075)

Table 5: Results for the low-dimensional ticket demand dataset using tree-based estimators compared to DNN estimators.

Learning Rate	Weight Decay	Dropout	DNN Width	DML-IV	CE-DML-IV
0.0002	0.001	0.1	128	**0.0676(0.0129)**	**0.0765(0.0119)**
0.0005				0.0752(0.0122)	0.0897(0.0196)
0.0001				**0.0703(0.0195)**	**0.0794(0.0201)**
	0.0005			0.0794(0.0185)	0.0823(0.0149)
	0.005			0.0765(0.0135)	0.0809(0.0159)
	0.01			0.0820(0.0162)	0.0865(0.0174)
		0.05		**0.0715(0.0074)**	**0.0813(0.0089)**
		0.2		0.0836(0.0100)	0.0919(0.0157)
			64	0.0830(0.0162)	0.0924(0.0121)
			256	0.0943(0.0179)	0.0981(0.0126)
	0.0005	0.2		0.0805(0.0133)	0.0910(0.0106)
	0.005	0.05		**0.0672(0.0116)**	**0.0742(0.0102)**
	0.01	0.05		0.0825(0.0152)	0.0914(0.0125)
		0.2	256	0.0810(0.0129)	0.0852(0.0121)
		0.05	64	0.0907(0.0149)	0.0963(0.0161)
	0.005		256	0.0939(0.0146)	0.0991(0.0093)

Table 6: Results for the low-dimensional ticket demand dataset for a range of hyperparameter values. The default hyperparameters in this case are: learning rate=0.0002, weight decay=0.001, dropout=0.1 and DNN width 128.

Learning Rate	Weight Decay	Dropout	CNN Channels	DML-IV	CE-DML-IV
0.001	0.05	0.2	64	**0.3513(0.0125)**	**0.3808(0.0150)**
0.0005				0.4063(0.0129)	0.5008(0.0369)
0.002				0.3659(0.0219)	0.4133(0.0267)
0.005				**0.3377(0.0218)**	**0.3555(0.0202)**
	0.01			0.3935(0.0176)	0.4461(0.0478)
	0.02			**0.3595(0.03013)**	**0.3851(0.0293)**
	0.1			0.4066(0.0172)	0.5160(0.0329)
		0.1		0.4136(0.0211)	0.5386(0.0398)
		0.3		0.3857(0.0171)	0.4002(0.0249)
			128	0.4176(0.01941)	0.5129(0.0630)
			256	0.4942(0.0226)	0.6180(0.0396)
	0.1	0.1		0.4163(0.0214)	0.5952(0.0343)
	0.01	0.3		0.3636(0.0186)	0.3995(0.0250)
		0.3	128	0.4006(0.0187)	0.4764(0.0216)
		0.3	256	**0.3429(0.0215)**	**0.3971(0.0264)**
	0.1		256	0.4170(0.0283)	0.5335(0.0371)

Table 7: Results for the high-dimensional ticket demand dataset for a range of hyperparameter values. The default hyperparameters in this case are: learning rate 0.001, weight decay=0.05, dropout=0.2 and 64 CNN channels.