Double Machine Learning for Conditional Moment Restrictions: IV regression, Proximal Causal Learning and Beyond

Daqian Shao DAQIAN.SHAO@CS.OX.AC.UK
Department of Computer Science
University of Oxford
Oxford, UK

Ashkan Soleymani ASHKANSO@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, USA

Francesco Quinzan* FRANCESCO.QUINZAN@CS.OX.AC.UK
Department of Computer Science
University of Oxford
Oxford, UK

Marta Kwiatkowska MARTA.KWIATKOWSKA@CS.OX.AC.UK
Department of Computer Science
University of Oxford
Oxford, UK

Abstract

Solving conditional moment restrictions (CMRs) is a key problem considered in statistics, causal inference, and econometrics, where the aim is to solve for a function of interest that satisfies some conditional moment equalities. Specifically, many techniques for causal inference, such as instrumental variable (IV) regression and proximal causal learning (PCL), are CMR problems. Most CMR estimators use a two-stage approach, where the first-stage estimation is directly plugged into the second stage to estimate the function of interest. However, naively plugging in the first-stage estimator can cause heavy bias in the second stage. This is particularly the case for recently proposed CMR estimators that use deep neural network (DNN) estimators for both stages, where regularisation and overfitting bias is present. We propose DML-CMR, a two-stage CMR estimator that provides an unbiased estimate with fast convergence rate guarantees. We derive a novel learning objective to reduce bias and develop the DML-CMR algorithm following the *double/debiased machine learning (DML)* framework. We show that our DML-CMR estimator can achieve the minimax optimal convergence rate of $O(N^{-1/2})$ under parameterisation and mild regularity conditions, where N is the sample size. We apply DML-CMR to a range of problems using DNN estimators, including IV regression and proximal causal learning on real-world datasets, demonstrating state-of-the-art performance against existing CMR estimators and algorithms tailored to those problems.

Keywords: conditional moment restrictions, double machine learning, instrumental variable regression, proximal causal learning, two-stage regression

∗. Current address and email: Department of Engineering Science, University of Oxford, UK. francesco.quinzan@eng.ox.ac.uk

1 Introduction

In this work, we study the problem of *conditional moment restrictions* (CMRs). Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$, $C \in \mathcal{C} \subseteq \mathbb{R}^p$, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ be random variables with their corresponding distributions. The CMR problem consists of estimating a function of interest $f_0 \in \mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$ in some hypothesis function space $\mathcal{F}$, such that

$$\mathbb{E}[Y - f_0(X)|C] = 0. \tag{P}$$

This problem appears in many fields of research, such as statistical learning (Vapnik, 1998), causal inference (Liao et al., 2020), integral equations (Honerkamp and Weese, 1990), deconvolution (Carrasco et al., 2007), and econometrics (Carrasco et al., 2007), and includes causal inference problems such as instrumental variable (IV) regression (Wright, 1928) and proximal causal learning (Miao et al., 2018).

Solving CMRs analytically is ill-posed (Nashed and Wahba, 1974; Kress, 1999) because it is an inverse problem that requires the derivation of the function f_0 inside the conditional expectation. Hence, various techniques have been proposed to estimate the solutions to CMR problems. The classic econometric approaches begin with the framework of generalised method of moments (GMM) (Hansen, 1982), which gives rise to the classic two-stage least squares (2SLS) (Angrist et al., 1996) algorithm and sieve methods (Newey and Powell, 2003; Ai and Chen, 2003). These techniques have strong theoretical foundations for convergence analysis, but they are restricted to the class of basis functions used in their theoretical analysis, and their empirical performance may suffer when estimating complex function classes and with high-dimensional input. More recent works (Hartford et al., 2017; Shao et al., 2024; Bennett et al., 2019; Singh et al., 2019), inspired by the development of deep learning, proposed to use deep neural networks (DNNs) to parameterise and estimate the solutions to CMR problems. These methods allow for greater flexibility since they do not impose strong assumptions on the functional form such as linear functions or polynomials, and can learn directly from data with strong approximation power.

Most of the existing CMR methods, including those that adopt DNNs, take a two-stage approach (Angrist et al., 1996; Newey and Powell, 2003; Chen and Christensen, 2018; Singh et al., 2019; Muandet et al., 2020). In the first stage, they estimate some nuisance parameters, which are parameters or infinite-dimensional functions of no direct interest, but are necessary for the second stage estimation. However, in these settings, regularisation is often employed to trade off overfitting with the induced regularisation bias, especially for high-dimensional inputs. This is problematic because both regularisation and overfitting can cause heavy bias (Chernozhukov et al., 2018) in two-stage estimations when the first stage estimator is naively plugged in for the second stage estimation, which results in slow convergence rate of the estimator.

In order to mitigate this problem, we take inspiration from *double/debiased machine learning* (Chernozhukov et al., 2018) (DML), which is a technique that provides an unbiased estimator with strong convergence rate guarantees for general two-stage regressions. DML relies on having a Neyman orthogonal (Neyman and Scott, 1965) score function to deal with regularisation bias, and uses cross-fitting, that is, an efficient form of (randomised) data splitting, to address overfitting bias. However, the use of DML for CMR estimation involving neural networks has not been explored to the best of our knowledge.

In this work, we propose a novel CMR estimator, referred to as DML-CMR, with fast convergence rate guarantees based on the DML framework. We derive a novel Neyman orthogonal score for CMR problems and design a cross-fitting regime such that, under mild regularity conditions, it is guaranteed to converge at the rate of $N^{-1/2}$ with high probability, where N is the sample size. For empirical evaluation, we apply DML-CMR with DNNs to solve two common CMR problems, IV regression and proximal causal learning. We evaluate its performance on multiple benchmarks, where superior results are demonstrated compared to state-of-the-art (SOTA) methods. Our main contributions are summarised below.

- We propose DML-CMR, a novel CMR estimator that leverages the DML framework to provide unbiased estimates of the solution to CMR problems. To the best of our knowledge, this is the first work that uses DML for CMR estimation with neural networks.
- We derive a novel, Neyman orthogonal, score function for CMR problems in Section 4.1, which does not rely on influence functions or the classical techniques in Chernozhukov et al. (2018), and design a cross-fitting regime for the DML-CMR estimator to mitigate the bias in Section 4.2.
- In Section 4.3, we show an asymptotic convergence rate for the DML-CMR estimator at the rate of $O(N^{-1/2})$, which is minimax optimal under parameterisation and mild regularity conditions.
- On a range of IV regression and proximal causal learning problems, including two real-world datasets, we experimentally demonstrate that DML-CMR outperforms existing SOTA methods in Section 5.

This paper extends our earlier approach of Shao et al. (2024), which developed an IV regression algorithm, called DML-IV, that uses the DML framework to provide fast convergence rate guarantees. Compared to Shao et al. (2024), we consider the more general class of CMR estimators leveraging the DML (Chernozhukov et al., 2018) framework, derive new theoretical results and analysis, and conduct new experiments for the causal proximal learning problem.

2 Related Works

2.1 Conditional Moment Restriction

The classic framework of the generalised method of moments (GMM) (Hansen, 1982) was first proposed to translate conditional moment restrictions into unconditional moments, which can thus be estimated under the GMM framework. However, finding the set of unconditional moments that fully capture the conditional moments is challenging. Sometimes, for nonlinear CMRs and functions of interest f, an infinite set of unconditional moments may be required to represent the conditional moments, and a misspecification of the unconditional moments can bias the results (Domínguez and Lobato, 2004). To mitigate some of these issues, Angrist et al. (1996) considered linear functions of interest and proposed the classic two-stage least squares (2SLS) algorithm as a special case of GMMs. Following this, many efforts aim to

extend 2SLS to solve CMRs, where the function of interest is nonlinear or nonparametric. One common approach is the sieve method, which uses nonlinear basis functions. Sieve minimum distance (SMD) estimator (Newey and Powell, 2003; Ai and Chen, 2003) performs regression in two stages using an increasing set of nonlinear basis functions as the number of samples increases. Later, a penalised version of SMD (Chen and Pouzo, 2012) was proposed to generalise SMD by allowing non-smooth residuals and high-dimensional function spaces. These sieve-GMM methods and later works (Blundell et al., 2007; Chen and Christensen, 2018; Singh et al., 2019; Muandet et al., 2020) that consider different dictionaries of basis functions enjoy strong consistency and efficiency guarantees, but their flexibility is limited by the set of basis functions, and they can be sensitive to the choice of such functions and regularisation hyperparameters. Note that these works also perform the estimation in two stages.

More recently, various CMR estimation methods based on DNNs have been proposed, since machine learning methods can model highly nonlinear and high-dimensional relationships with greater flexibility. DeepIV (Hartford et al., 2017) extended the classic 2SLS algorithm to the nonlinear setting by adopting DNNs for both stages with conditional density estimation (Darolles et al., 2011) in the first stage. GMM methods are also extended to use DNNs (Bennett et al., 2019; Liao et al., 2020; Dikkala et al., 2020; Bennett and Kallus, 2020), where a minimax criterion is optimised adversarially. Minimax approaches aim to solve a two-player zero-sum game where the players play adversarially. Specifically, player one chooses a hypothesis function to minimise moment violation, and player two chooses a test function that maximises moment violation. These DNN-based GMM methods require minimax optimisation, which is similar to the training of Generative Adversarial Networks (Goodfellow et al., 2014) and could be experimentally unstable.

In this work, we propose a doubly robust CMR estimator for nonlinear functions of interest f that leverages the DML (Chernozhukov et al., 2018) framework and DNNs to provide fast convergence rate guarantees for the estimator, extending our previous IV regression algorithm, DML-IV, that also uses the DML framework (Shao et al., 2024).

2.1.1 IV regression

Instrumental variable (IV) regression is a typical example of a CMR problem in causal inference, which aims to estimate the causal effect in the presence of hidden confounders (see Appendix D.3 for details). While most of the previously introduced CMR estimators can be directly applied to solve IV regression, here we review algorithms specifically designed for IV regression. Following the sieve-based approach, DFIV (Xu et al., 2020) proposed to use basis functions parameterised by DNNs, which remove restrictions on the functional form. In addition, Kernel IV (Singh et al., 2019) and Dual IV (Muandet et al., 2020) used different dictionaries of basis functions in reproducing kernel Hilbert spaces (RKHS) to solve the IV regression problem. DeepGMM (Bennett et al., 2019) is a DNN-based method that was inspired by GMM to solve IV regression using a minimax approach. Kremer and Schölkopf (2024) improved GMM-based IV regression methods in settings where the data manifold is not uniform through data-derivative information. With the exception of DML-IV (Shao et al., 2024), the precursor of this work, none of these approaches utilise the DML framework.

2.1.2 Proximal Causal Learning

Another example of a CMR problem in causal inference is *proximal causal learning* (PCL, see Appendix D.4 for details). PCL was first proposed by Miao et al. (2018) to leverage two *proxy variables* for causal identification in estimating the causal function. This was extended by Shi et al. (2020) to a general semiparametric framework, where Tchetgen et al. (2024) introduced a two-stage procedure for linear causal models based on ordinary least squares regression. Mastouri et al. (2021) resolved how to handle nonlinear causal models by replacing linear regression with kernel ridge regression. To extend the kernel-based PCL methods, Xu et al. (2021) used DNNs as feature maps instead of fixed kernels. This improves the flexibility of the method, especially for highly nonlinear models. Kompa et al. (2022) proposed a single-stage PCL method based on maximum moment restrictions, where they train neural networks to minimise a loss function derived to satisfy the maximum moment restrictions. Cui et al. (2023) introduced a treatment bridge function and incorporated it into the Proximal Inverse Probability Weighting (PIPW) estimator. They considered only binary treatments and derived the Proximal Doubly Robust (PDR) estimator via influence functions. A similar approach by Wu et al. (2024) derived a doubly robust estimator for PCL with continuous treatment through influence functions, but none of these works adopted the DML framework, to the best of our knowledge.

The algorithms proposed specifically for IV regression and PCL often require additional problem-specific assumptions about variables and the functional form. We instead provide a general method for CMRs that can be directly applied to a range of problems, including IV regression and PCL.

2.2 Double Machine Learning (DML)

DML was originally proposed for semiparametric regression (Robinson, 1988). It relies on the derivation of a Neyman orthogonal (Neyman and Scott, 1965) score function that serves as the learning objective. DML was then extended by adopting DNNs for generalised linear regressions (Chernozhukov et al., 2021). Its strength is that it provides unbiased estimates for two-stage estimations (Jung et al., 2021; Chernozhukov et al., 2022b) under certain identifiability conditions and offers $N^{-1/2}$ convergence rate guarantees.

There are previous works on combining DML with CMR estimation, specifically for the IV regression problem, but they are mainly focused on linear and partially linear functions of interest. Belloni et al. (2012) proposed a method to use Lasso and Post-Lasso methods for the first-stage estimation of linear IV to estimate the optimal instruments. To avoid selection biases, Belloni et al. (2012) leveraged techniques from weak identification robust inference. In addition, Chernozhukov et al. (2015) proposed 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 IVs was mainly discussed in the work of Chernozhukov et al. (2018). For an additional discussion, we refer to the book (Chernozhukov et al., 2024).

DML for semiparametric models (Chernozhukov et al., 2022a; Ichimura and Newey, 2022) has been previously applied to solve the nonparametric IV (NPIV) problem. However, their methods require additional assumptions on the IVs and residual functions such that the average moment of the Neyman orthogonal score is linear in the nuisance parameters. Such

assumptions are not required in our work since we are considering a different problem setting and we formulate a novel Neyman orthogonal score. To the best of our knowledge, there is no work that adopts the DML framework for nonlinear IV regression and general CMR problems with DNNs.

3 Preliminaries

Notations

We use uppercase letters such as X to denote random variables and use the corresponding calligraphic letter such as $\mathcal{X}$ to denote the set from which the random variable takes its value. For example, $X \in \mathcal{X} \subseteq \mathbb{R}^d$ is a d-dimensional real-valued random variable in $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, where $\mathcal{B}_\mathcal{X}$ is the Borel algebra on $\mathcal{X}$. An observed realisation of X is denoted by a lowercase letter x. We abbreviate $\mathbb{E}[Y|X = x]$, a realisation of the conditional expectation $\mathbb{E}[Y|X]$, as $\mathbb{E}[Y|x]$. $[N]$ denotes the set $\{1, ..., N\}$ for $N \in \mathbb{N}$. We use $\|\cdot\|_p$ to denote the functional norm, defined as $\|f\|_p \coloneqq \mathbb{E}[|f(X)|^p]^{1/p}$, where the measure is implicit from the context. For a function f, we use f_0 to denote the true function and $\widehat{f}$ an estimator of the true function. We use O and o to denote big-O and little-o notations, respectively (Weisstein, 2023).

3.1 Conditional Moment Restrictions

Recall that the CMR problem as in Equation (P) consists of providing an estimate for a function f_0 such that $\mathbb{E}[Y - f_0(X)|C = c] = 0$ for all $c \in \mathcal{C}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^d$, $C \in \mathcal{C} \subseteq \mathbb{R}^p$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ are random variables with their corresponding distributions. As discussed in Section 2.1, many CMR estimators (e.g., Angrist et al. 1996; Newey and Powell 2003; Hartford et al. 2017; Singh et al. 2019) estimate $\widehat{f}$ in some space of functions $\mathcal{F}$ by solving the following objective function with a two-stage approach:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}[(Y - \mathbb{E}[f(X)|C])^2]. \tag{1}$$

Specifically, Newey and Powell (2003) take a minimax approach that indirectly optimises this objective by solving a minimax unconditional moment problem in two stages, as elaborated on in Section 2.1. Angrist et al. (1996); Singh et al. (2019); Hartford et al. (2017), on the other hand, directly optimise the above objective in two stages. The first stage involves learning the conditional expectation $\mathbb{E}[f(X)|c]$ using either density estimation or kernel methods from observations. In the second stage, the objective in Equation (1) is minimised using the estimations in the first stage. For both stages, linear regression, sieve methods, and DNNs are used for estimation, respectively, for each work.

3.2 Double Machine Learning

DML considers the problem of estimating a function of interest f as a solution to an equation of the form

$$\mathbb{E}[\psi(\mathcal{D}; f_0, \eta)] = 0, \tag{2}$$

where ψ is referred to as a score function and f_0 is the true function. Here, η is a nuisance parameter, which can be of parametric form or an infinite-dimensional function. It is of

no direct interest, but must be estimated to obtain an estimate of f_0. For example, in a two-stage CMR estimator, nuisance parameters such as conditional density are estimated in the first stage, and in the second stage, they are used to estimate f_0. DML provides a set of tools to derive an unbiased estimator of f_0 with convergence rate guarantees, even when the nuisance parameter η suffers from regularisation, overfitting and other types of biases present in the training of ML models, which typically cause slow convergence when learning f_0.

In order to estimate f_0, DML reduces biases by using score functions ψ that are Neyman orthogonal (Neyman and Scott, 1965) in η. This requires the Gateaux derivative, which defines the directional derivative for functionals, of the score function ψ w.r.t. the nuisance parameters at f_0, η_0 to be zero:

$$\frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}[\psi(\mathcal{D}; f_0, \eta_0 + r\eta)] = 0, \tag{3}$$

for all η. Here, f_0 and η_0 are the true parameters that minimise the expected score, that is, $\mathbb{E}[\psi(\mathcal{D}; f_0, \eta_0)] = 0$. Intuitively, the condition in Equation (3) is met if small changes in the nuisance parameter do not significantly affect the score function around the true function f_0. Neyman orthogonality is key in DML, as it allows fast convergence for estimating f_0, 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 1 (DML, Definition 3.2 (Chernozhukov et al., 2018)) *Given a dataset $\mathcal{D}$ of N observations, consider a score function ψ as in Equation (2), and suppose that ψ is Neyman orthogonal that satisfies Equation (3). 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 $\widehat{\eta}_k$ of the nuisance parameter using $\mathcal{D}_{I_k^c}$. Then, construct an estimator $\widehat{f}$ as a solution to the equation*

$$\frac{1}{K}\sum_{k=1}^{K} \widehat{\mathbb{E}}_k[\psi(\mathcal{D}_{I_k}; \widehat{f}, \widehat{\eta}_k)] = 0, \tag{4}$$

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

In the above definition, $\widehat{f}$ is defined as an exact solution to the empirical expectation equation in Equation (4). In practice, we can also define the estimator $\widehat{f}$ as an approximate solution to Equation (4)[1].

4 Solving CMRs with DML

We now present the main contributions of this paper — an estimator for solving CMRs under the DML framework. We propose a novel Neyman orthogonal score for our estimator

1. This approximation error is different to the estimation error. The estimation error measures the difference between $\widehat{f}$ and f_0, whereas the approximation error concerns the error of minimising the empirical risk. In fact, the approximation error contributes to the estimation error, which is analysed in Section 4.3.

and a novel two-stage algorithm for solving CMRs that can use DNN estimators in both stages and provides guarantees on the convergence rate by leveraging the DML framework.

To solve the CMR problem defined in Equation (P) using the DML framework, we first need a Neyman orthogonal score. As introduced in Section 3.1, the optimisation objective for many two-stage CMR estimators (Angrist et al., 1996; Hartford et al., 2017; Singh et al., 2019) is Equation (1). Let $g_0(f, c) \coloneqq \mathbb{E}[f(X)|c]$ and $\mathcal{G}$ be some function space that includes g_0 and its potential estimators $\widehat{g}$. Then, these two-stage CMR estimators estimate g_0 with $\widehat{g}$ in the first stage, and then use $\widehat{g}$ to optimise the following loss function, $\ell = (Y - \widehat{g}(f, c))^2$, in the second stage. Unfortunately, as we show in Proposition 2 with proof deferred to Appendix B, this objective, or the score function, is not Neyman orthogonal.

Proposition 2 *The score (or objective) function for standard two-stage CMR estimators, $\ell = (Y - \widehat{g}(f, c))^2$, is not Neyman orthogonal at (f_0, g_0).*

This means that small misspecifications or estimation biases of $\widehat{g}$ can lead to significant changes to the score function, and there are no guarantees on the convergence rate if the first stage estimator $\widehat{g}$ is naively plugged into the second stage to estimate f_0. To address this, we first derive a novel Neyman orthogonal score function for the CMR problem and then design a CMR algorithm with K-fold cross-fitting that uses the DML framework.

4.1 Neyman Orthogonal Score

Typically, to construct a Neyman orthogonal score from a non-orthogonal score, additional nuisance parameters need to be estimated (Chernozhukov et al., 2018). These additional nuisance terms adjust the score in a way that makes it orthogonal, where the error in estimating f_0 due to errors in the nuisance parameters becomes second order in the Taylor expansion (Foster and Syrgkanis, 2019). In our case, to construct a Neyman orthogonal score for CMR problems from the standard objective in Equation (1), we first select relevant functions that should be estimated as nuisance parameters. Following two-stage IV regression approaches (Hartford et al., 2017), estimating g_0 is essential for identifying f_0, so we will estimate it as a nuisance parameter. We find that, by additionally estimating $s_0(c) \coloneqq \mathbb{E}[Y|c]$ inside some function space $\mathcal{S}$, we can construct the score function:

$$\psi(\mathcal{D}; f, (s, g)) = (s(c) - g(f, c))^2. \tag{5}$$

Here, the nuisance parameters are $\eta = (s, g)$. For this to be a valid Neyman orthogonal score function, we check that $\mathbb{E}[\psi(\mathcal{D}; f_0, (s_0, g_0))] = 0$ with the true functions (s_0, g_0), and its Gateaux derivative vanishes at $(f_0, (s_0, g_0))$ with the following theorem, where the proof is deferred to Appendix C.1.

Theorem 3 (Neyman orthogonality) *The score function $\psi(\mathcal{D}; f, (s, g)) = (s(c) - g(f, c))^2$ obeys the Neyman orthogonality conditions at $(f_0, (s_0, g_0))$.*

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. In addition, having a Neyman orthogonal score is useful in general to debias two-stage estimators (Foster and Syrgkanis, 2019), beyond the specific DML regimes we are considering in this paper.

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

1: **Input:** Dataset $\mathcal{D}$ of size N, number of folds K for cross-fitting, mini-batch size n_b
2: Get a partition $(I_k)_{k=1}^K$ of dataset indices $[N]$
3: **for** $k = 1$ **to** K **do**
4: $\quad I_k^c \coloneqq [N] \setminus I_k$
5: $\quad$ Learn $\widehat{s}_k$ and $\widehat{g}_k$ using $\{(\mathcal{D}_i) : i \in I_k^c\}$
6: **end for**
7: Initialise $f_{\widehat{\theta}}$
8: **repeat**
9: $\quad$ **for** $k = 1$ **to** K **do**
10: $\quad\quad$ Sample n_b data c_i^k from $\{(\mathcal{D}_i) : i \in I_k\}$
11: $\quad\quad$ $\mathcal{L} = \widehat{\mathbb{E}}_{c_i^k}\left[(\widehat{s}_k(c) - \widehat{g}_k(f_{\widehat{\theta}}, c))^2\right]$
12: $\quad\quad$ Update $\widehat{\theta}$ to minimise loss $\mathcal{L}$
13: $\quad$ **end for**
14: **until** convergence
15: **Output:** The DML-CMR estimator $f_{\widehat{\theta}}$

4.2 A DML Estimator for Solving CMRs

With the Neyman orthogonal score, we now propose a novel DML estimator, DML-CMR, that solves Equation (P). Note that, in general, f_0 is allowed to be infinite-dimensional, as commonly seen in the nonparametric IV literature (Newey and Powell, 2003). We also allow f_0 to be infinite-dimensional for the Neyman orthogonal score introduced in Section 4.1. For the theoretical analysis of DML-CMR, while it is possible to provide a general analysis following Foster and Syrgkanis (2019) for nonparametric f_0 with the Neyman orthogonal score, the analysis would require more assumptions and the convergence rate will depend on the complexity of the function classes involved in the estimation. For our analysis, since we propose a concrete estimator, we would like to provide a concrete analysis following the DML framework (Chernozhukov et al., 2018), which is designed for semiparametric estimation, to show an optimal parametric rate for DML-CMR. Therefore, we assume that f_0 is finite-dimensional and parameterised for the theoretical analysis of DML-CMR.

Assumption 1 (Parameterisation) *Let $f_0 = f_{\theta_0}$ and $\Theta \subseteq \mathbb{R}^{d_\theta}$ be a compact space of parameters of f, where the true parameter $\theta_0 \in \Theta$ is in the interior of Θ.*

From this assumption, we can define $\mathcal{F} \coloneqq \{f_\theta : \theta \in \Theta\}$ as the function space of f.

The DML-CMR Estimator. The procedure of our DML-CMR estimator with k-fold cross-fitting is outlined in Algorithm 1. Given a dataset $\mathcal{D} = (y_i, x_i, c_i)_{i\in[N]}$ of size N, we first 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, and let $\mathcal{D}_{I_k}$ denote the set of observations $\{\mathcal{D}_i : i \in I_k\}$.

As introduced in Section 3.2, our DML estimator will be a two-stage procedure. In the first stage (lines 4-7 in Algorithm 1), for each fold $k \in [K]$, we learn $\widehat{s}_k$ and $\widehat{g}_k$ using data $\mathcal{D}_{I_k^c}$ with indices $I_k^c \coloneqq [N] \setminus I_k$. Then $\widehat{s}_k \approx \mathbb{E}[Y|C]$ can be learnt through standard supervised learning using a neural network with inputs C and labels Y. For $\widehat{g}_k$, we follow Hartford

et al. (2017) to estimate $F_0(X|C)$, the conditional distribution of X given C, with $\widehat{F}$, and then estimate $\widehat{g}$ via

$$\widehat{g}(f_\theta, c) = \sum_{\dot{X} \sim \widehat{F}(X|C)} f_\theta(\dot{X}) \approx \int f_\theta(X) \widehat{F}(X|C=c) dX \approx \mathbb{E}[f_\theta(X)|c].$$

For example, if the action space is discrete, $\widehat{F}$ can be a categorical model, e.g., a DNN with softmax output. For a continuous action space, a mixture of Gaussian models can be adopted to estimate the distribution $F_0(X|C)$, where a DNN is used to predict the mean and standard deviation of the Gaussian distributions.

In the second stage (lines 8-15 in Algorithm 1), we estimate $\widehat{\theta}$ using our Neyman orthogonal score function ψ in Equation (5). The key is to optimise $\widehat{\theta}$ with data from the k-th fold $\mathcal{D}_{I_k}$ using nuisance parameters $\widehat{s}_k$, $\widehat{g}_k$ that are trained with $\mathcal{D}_{I_k^c}$, the complement of $\mathcal{D}_{I_k}$. This is important to fully debias the estimator $\widehat{\theta}$. The DML estimator $f_{\widehat{\theta}}$ is then defined as

$$f_{\widehat{\theta}} \coloneqq \min_{f_\theta \in \mathcal{F}} \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathbb{E}}_k[(\widehat{s}_k(c) - \widehat{g}_k(f_{\widehat{\theta}}, c)^2], \tag{6}$$

where $\widehat{\mathbb{E}}_k$ is the empirical expectation over $\mathcal{D}_{I_k}$. In practice, we can alternate between the K folds while sampling a mini-batch c_i^k of size n_b from each fold $\mathcal{D}_{I_k}$ to update $\widehat{\theta}$ by minimising the empirical loss on the mini-batch following our Neyman orthogonal score ψ,

$$\mathcal{L} = \widehat{\mathbb{E}}_{c_i^k}\left[(\widehat{s}_k(c) - \widehat{g}_k(f_{\widehat{\theta}}, c))^2\right] = \sum_{c_i^k} \frac{1}{n_b} \left((\widehat{s}_k(c) - \widehat{g}_k(f_{\widehat{\theta}}, c))^2\right).$$

When the second stage converges, we return the DML-CMR estimator $f_{\widehat{\theta}}$.

4.3 Theoretical Analysis

In this section, we provide a theoretical analysis of the convergence of DML-CMR. The key benefit of DML is its debiasing effect for two-stage regressions, and crucially, it is possible to leverage the DML framework (Chernozhukov et al., 2018) to show a fast asymptotic convergence rate of $O(N^{-1/2})$, i.e., the DML estimator $\widehat{\theta}$ converges to the true parameters θ_0 at the rate of $O(N^{-1/2})$ with high probability. To provide a road map of this section, we first list all the technical conditions required for a general DML estimator to converge at the fast rate of $O(N^{-1/2})$ following Theorem 3.3 from (Chernozhukov et al., 2018). Later, in Theorem 6, we will prove that all these conditions hold for our DML-CMR estimator.

Condition 4 (Technical conditions of DML $N^{-1/2}$ rate proved later in Theorem 6) *For sample size $N \geq 3$:*

(a) The map $(\theta, (s, g)) \mapsto \mathbb{E}[\psi(\mathcal{D}; f_\theta, (s, g))]$ is twice continuously Gateaux-differentiable.

(b) The score ψ obeys the Neyman orthogonality conditions in Equation (3).

(c) The true parameter θ_0 obeys $\mathbb{E}[\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))] = 0$ and Θ contains a ball of radius $c_1 N^{-1/2} \log N$ centered at θ_0.

(d) For all $\theta \in \Theta$, the identification relationship

$$2\|\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]\| \gtrsim \|J_0(\theta - \theta_0)\|$$

is satisfied, where $J_0 := \partial_{\theta'}\{\mathbb{E}[\psi(\mathcal{D}; f_{\theta'}, (s_0, g_0))]\}|_{\theta'=\theta_0}$ is the Jacobian matrix, with singular values bounded between $c_0 > 0$ and $c_1 > 0$.

(e) All eigenvalues of the matrix $\mathbb{E}[\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))^T]$ are strictly positive (bounded away from zero).

(f) Let K be a fixed integer. Given a random partition $\{I_k\}_{k=1}^K$ of indices $[N]$, each of size $n = N/K$, the nuisance parameter estimator $\widehat{s}_k$ and $\widehat{g}_k$ learnt using data with indices I_k^c belongs to shrinking realisation sets $\mathcal{S}_N$ and $\mathcal{G}_N$, respectively, and the nuisance parameters should be estimated at the $o(N^{-1/4})$ rate, e.g., $\|\widehat{s} - s_0\|_2 = o(N^{-1/4})$.

Among these conditions, (a), (b), and (c) are conditions regarding the Neyman orthogonal score ψ. The Neyman orthogonality in (b) is shown in Theorem 3 and the other conditions (b) and (c) are mild regularity conditions, standard for moment problems. (d) is an identification condition that ensures sufficient identifiability of θ_0. This condition also implies bounded ill-posedness, which we will discuss in detail in Section 4.4. (e) is a non-degeneracy assumption for the covariance of the score function. Finally, (f) is a key condition that states the nuisance parameters should converge to their true values at the crude rate of $o(N^{-1/4})$, where a shrinking realisation set $\mathcal{S}_N$ is a decreasing set of possible estimators $\widehat{s}$ as the sample size N increases.

In Lemma 5, following recent works (Chernozhukov et al., 2018, 2021, 2022b), we show that the convergence condition for the nuisance parameters in Condition 4 (f) can be transformed to a condition on the critical radius (Bartlett et al., 2005) of the realisation sets, which is a property widely studied for various function classes. For this analysis, we start by assuming the realisability of the true functions g_0, s_0 and f_0 in their corresponding function classes, and further assuming that they are bounded. We formalise these assumptions in Assumption 2.

Assumption 2 (Realisable and bounded) *We assume that g_0, s_0, f_0 are realisable in the function classes $\mathcal{G}, \mathcal{S}, \mathcal{F}$, that is, $g_0, s_0, f_0 \in \mathcal{G}, \mathcal{S}, \mathcal{F}$, respectively, and furthermore, $\|f\|_\infty, \|s\|_\infty \leq B$ for all $f, s \in \mathcal{F}, \mathcal{S}$, where B is a positive constant. Moreover, we assume that the random variable $|Y| \leq B$ almost surely.*

As discussed in Section 4.1, DML-CMR has two nuisance parameters that are required to be estimated: $\widehat{s} \in \mathcal{S}$ and $\widehat{g} \in \mathcal{G}$. As we saw in Section 4.2, the estimation of $\widehat{s}$ is made through standard supervised learning algorithms that we can directly analyse. However, the estimation of $\widehat{g}$ has two steps: (i) we estimate the conditional distribution $\widehat{F}(X|C) \in \mathcal{P}$, where the density sieve $\mathcal{P}$ is defined as

$$\mathcal{P} \subset \left\{F : \int F(x|C = c)dx = 1 \quad \forall c \in \mathcal{C}\right\};$$

and (ii) we plug in the functional f_θ into the conditional expectation estimator,

$$\widehat{g}_{\widehat{F}}(f_\theta, c) := \int f_\theta(x)\widehat{F}(x|C = c)dx,$$

for all candidate test functions $f_\theta \in \mathcal{F} = \{f_\theta \ : \ \theta \in \Theta\}$. From the realisability assumption of $g_0 \in \mathcal{G}$ in Assumption 2, it follows that $F_0(X|C) \in \mathcal{P}$, and the hypothesis space for the estimand $\widehat{g}$ and the true parameter g_0 is defined as $\mathcal{G} \coloneqq \{g_F \ : \ F \in \mathcal{P}\}$.

Lemma 5 (Convergence of nuisance parameters) *Under Assumption 2, let $\mathcal{S}_N^*$ be the star-hull of the realisation set $\mathcal{S}_N$ of function class $\mathcal{S}$,*

$$\mathcal{S}_N^* = \{C \mapsto \gamma(s(C) - s_0(C)) \ : \ s \in \mathcal{S}_N, \gamma \in [0,1]\},$$

$\mathcal{P}_N^*$ be the star-hull of the realisation set $\mathcal{P}_N$ of the function class $\mathcal{P}$,

$$\mathcal{P}_N^* = \{C \mapsto \gamma(F(\cdot|C) - F_0(\cdot|C)) \ : \ F \in \mathcal{P}_N, \gamma \in [0,1]\},$$

and $\mathcal{G}_N^*$ be the star-hull of the realisation set $\mathcal{G}_N$ of the function class $\mathcal{G}$,

$$\mathcal{G}_N^* = \{C, f \mapsto \gamma(g(C,f) - g_0(C,f)) \ : \ g \in \mathcal{G}_N, \gamma \in [0,1]\},$$

where $\mathcal{S}_N$, $\mathcal{P}_N$ and $\mathcal{G}_N$ are properly shrinking neighbourhoods of the true functions s_0, F_0 and g_0. Then, there exist universal constants c_1 and c_2, for which we have that, with probability at least $1-\xi$, the estimation errors are bounded as

$$\|\widehat{s} - s_0\|_2^2 \leq c_1 \left(\delta_N(\mathcal{S}_N^*)^2 + \sqrt{\frac{\log(1/\zeta)}{N}} \right);$$
$$\|\widehat{g} - g_0\|_2^2 \leq c_2 \left(\delta_N(\mathcal{P}_N^*)^2 + \sqrt{\frac{\log(1/\zeta)}{N}} \right).$$

This lemma shows that, if we can upper bound the critical radii of $\delta_N(\mathcal{S}^*)$ and $\delta_N(\mathcal{P}^*)$ by $o(N^{-1/4})$, then $\|\widehat{s} - s_0\|_2 = o(N^{-1/4})$, and $\|\widehat{g} - g_0\|_2 = o(N^{-1/4})$, meaning that nuisance parameters converge to their true values at the rate of $o(N^{-1/4})$ as required by Condition 4 (f). Next, we provide an analysis and concrete examples of estimators that satisfy this requirement.

The critical radius is a quantity that describes the complexity of estimation, and it is typically shown that $\delta_N = O(d_N^{1/2} N^{-1/2})$ (Chernozhukov et al., 2022b, 2021) (see Appendix C.3), where d_N is the effective dimension of the hypothesis space (for a formal definition of the critical radius and the effective dimension, the relationship of these metrics to Dudley's entropy integral, and bounds on the excess risk of estimators, refer to Appendix C.3.). This, together with Lemma 5, implies that $\|\widehat{s} - s_0\|_2 = O(d_N(\mathcal{S}^*)^{1/2} N^{-1/2})$. Therefore, we can also see that, if the effective dimension satisfies $d_N(\mathcal{S}^*) = o(N^{1/4})$, then $\|\widehat{s} - s_0\|_2 = o(N^{-1/4})$ as required by Condition 4 (f) (and similarly for $\widehat{g}$ and $d_N(\mathcal{P}^*)$).

Therefore, we can refer to results in the literature that analyse the effective dimension and critical radius of various estimators to provide examples of estimators that satisfy Condition 4 (f). For the estimation of $\widehat{s}$, we have a regression problem and Condition 4 (f) is satisfied by many supervised learning estimators such as parametric generalised linear models (Van Der Vaart et al., 1996), Lasso (Bickel et al., 2009), random forests (Syrgkanis and Zampetakis, 2020), boosting (Luo et al., 2016), Sobolev kernel regression with α-smooth RKHS ($\alpha > d/2$, where d is the dimension of X) (Caponnetto and De Vito, 2007; Christmann

and Steinwart, 2008) and neural networks (Chen and White, 1999; Yarotsky, 2018; Schmidt-Hieber, 2020; Farrell et al., 2021). For the conditional density estimator $\widehat{g}$, the above estimators also satisfy Condition 4 (f) if the conditional distribution can be parameterised accordingly; otherwise, the condition is also satisfied by Gaussian mixtures (mixture density networks) (Ho et al., 2022), polynomial sieve with Hölder smoothness $\alpha > \frac{d+1}{2}$ (Ai and Chen, 2003), and categorical-logistic models (Zhao et al., 2022), among others.

Lemma 5 allows us to obtain the following theorem regarding the convergence of the DML-CMR estimator by applying Theorem 3.3 of Chernozhukov et al. (2018). We prove satisfaction of all technical conditions for DML convergence rate guarantee mentioned in Condition 4 in Appendix C.1.

Theorem 6 (Convergence of the DML estimator for CMRs) *Let $f_{\theta_0} \in \mathcal{F}$ be a solution that satisfies the CMRs in Equation (P), let ψ be the Neyman orthogonal score defined in Equation (5) and let $J_0 := \partial_{\theta'}\{\mathbb{E}[\psi(\mathcal{D}; f_{\theta'}, (s_0, g_0))]\}|_{\theta'=\theta_0}$ be the Jacobian matrix of $\mathbb{E}[\psi]$ w.r.t. θ. Suppose that the upper bound of the critical radius $\delta_N = o(N^{-1/4})$, for $\widehat{s}$ and $\widehat{g}$, and J_0 have bounded singular values. Then, if Assumption 1 and 2 hold, our DML-CMR estimator $f_{\widehat{\theta}}$ satisfies that $\widehat{\theta}$ is concentrated in a $N^{-1/2}$ neighbourhood of θ_0, and is approximately linear and centered Gaussian:*

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

where the estimator variance is given by

$$\sigma^2 := J_0^{-1}\mathbb{E}[\psi(\mathcal{D}, \theta_0, (s_0, g_0))\psi(\mathcal{D}, \theta_0, (s_0, g_0))^T](J_0^{-1})^T,$$

which is constant w.r.t. N.

Theorem 6 states that, with adequately trained nuisance parameter estimators in terms of their critical radius and identifiability conditions in terms of the non-singularity of the Jacobian matrix J_0, the estimator error $\widehat{\theta} - \theta_0$ is normally distributed, where its variance shrinks at the rate of $N^{-1/2}$. This implies that $\widehat{\theta}$ converges to θ_0 at the rate $O(N^{-1/2})$ with high probability, which allows us to bound the estimation error $\|f_{\widehat{\theta}} - f_{\theta_0}\|_2$ of the DML-CMR estimator with high probability, under a Lipschitz condition of f_θ.

Corollary 7 *Let $f_{\widehat{\theta}}$ be our DML-CMR estimator. If all assumptions for Theorem 6 hold and there exists a constant $L > 0$ such that $\|f_\theta(x) - f_{\theta_0}(x)\|_2 \leq L\|\theta - \theta_0\|_2$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$, then for all $\zeta \in (0, 1]$, we have that*

$$\|f_{\widehat{\theta}} - f_{\theta_0}\|_2 = O\left(L\sqrt{\frac{\ln(1/\zeta)}{N}}\right),$$

with probability $1 - \zeta$.

Here, we assume a local Lipschitz condition of f_θ w.r.t. θ around θ_0: $\|f_\theta(x) - f_{\theta_0}(x)\|_2 \leq L\|\theta - \theta_0\|_2$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$ for some $L > 0$. Since Θ is compact, we see that it is enough for this Lipschitz condition to hold locally in a neighbourhood of θ_0.

4.4 DML Identifiability Condition and Ill-posedness

In this section, we provide a discussion regarding the relationship between our identifiability condition, Condition 4 (d), and a more common notion of identifiability for CMR problems, the *ill-posedness* (Chen and Pouzo, 2012). To begin with, we define the ill-posedness of CMR problems.

Definition 8 (Ill-posedness (Chen and Pouzo, 2012; Dikkala et al., 2020)) *Given a CMR problem as in Equation (P), the* ill-posedness ν *of the function space* $\mathcal{F}$ *is given by*

$$\nu = \sup_{f \in \mathcal{F}} \frac{\|f_0 - f\|_2}{\|\mathbb{E}[f_0(X) - f(X)|C]\|_2}.$$

Intuitively, ill-posedness describes how well a small CMR error (the projected error under conditional expectation) implies a small L_2 error (root mean squared error) to f_0. For identifiability, it is usually assumed that the ill-posedness ν is bounded. Otherwise, even solving the CMRs with very small error does not guarantee a solution $\widehat{f}$ that is close to f_0. In our case, we demonstrate that the identification condition of a DML estimator actually implies bounded ill-posedness. Specifically, Condition 4 (d) implies that the ill-posedness is bounded, as shown by the following proposition.

Proposition 9 *For all* $\theta \in \Theta$*, if there exists a constant* $L > 0$ *such that* $\|f_\theta(x) - f_{\theta_0}(x)\|_2 \leq L\|\theta - \theta_0\|_2$ *for all* $x \in \mathcal{X}$ *and* $\theta \in \Theta$*, then Condition 4 (d), which states*

$$2\|\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]\| \geq \|J_0(\theta - \theta_0)\|$$

and the Jacobian matrix J_0 *have singular values bounded between* $c_0 > 0$ *and* $c_1 > 0$*, implies the ill-posedness is bounded by* $\nu \leq L/\sqrt{c_0}$*.*

The proof of Proposition 9 is deferred to Appendix C.2. This interesting finding explains why, in Theorem 6, there are no explicit assumptions about the ill-posedness of the problem. The only identifiability assumption for Theorem 6 is the non-singularity of the Jacobian matrix J_0, which for DML-CMR is sufficient to ensure Condition 4 (d) holds. Therefore, by Proposition 9, this ensures a bounded ill-posedness of the problem, which in turn allows us to identify the solution f_0 with small error.

5 Experimental Results

In this section, we empirically evaluate our DML-CMR estimator. We apply DML-CMR to two applications, IV regression and proximal causal learning, where details regarding these two problems are provided in Appendix D. In addition, we evaluate a computationally efficient version of DML-CMR, referred to as CE-DML-CMR, which does not apply K-fold cross-fitting. It trains $\widehat{s}$ and $\widehat{g}$ 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-CMR empirically performs as

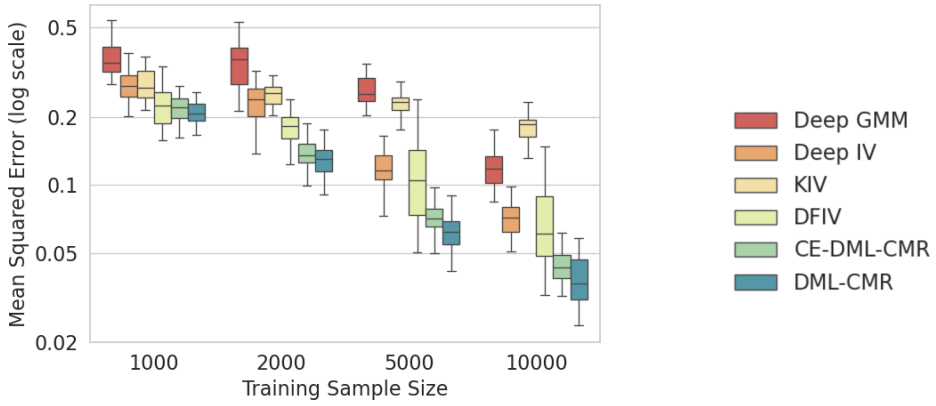

Figure 1: The mean squared error of $\widehat{f}$ on the ticket demand dataset with low-dimensional context for the IV regression task.

well as standard DML-CMR on low-dimensional datasets. We provide details and discussion regarding CE-DML-CMR in Appendix A.

Our evaluation considers both low- and high-dimensional datasets, as well as semi-synthetic real-world datasets. We ran each method 20 times and report the mean squared errors (MSE) between the estimators $\widehat{f}$ and f_0, where the median, 25th, and 75th percentiles are shown. The method employed by DML-CMR is identical to DML-IV (Shao et al., 2024) when solving the IV regression problem, and we include the results for IV regression from Shao et al. (2024) in this section. For PCL, the experimental results are new, and we implemented all algorithms using PyTorch (Paszke et al., 2019). The full code is available on GitHub[2].

5.1 IV regression

For the IV regression task (details in Appendix D.3), 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).

We use DNN estimators for both stages with network architecture and hyperparameters provided in Appendix F. Results of DML-CMR using tree-based estimators such as Random Forests and Gradient Boosting are provided in Appendix G.2, where comparable performance to DNN-based DML-CMR is demonstrated. In addition, we provide a sensitivity analysis against hyperparameter changes in Appendix G.3 and an evaluation of algorithms when the IV is weakly correlated with the treatment, representing higher ill-posedness of the CMRs, in Appendix G.1.

5.1.1 Ticket Demand Dataset

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

2. https://github.com/shaodaqian/DML-CMR

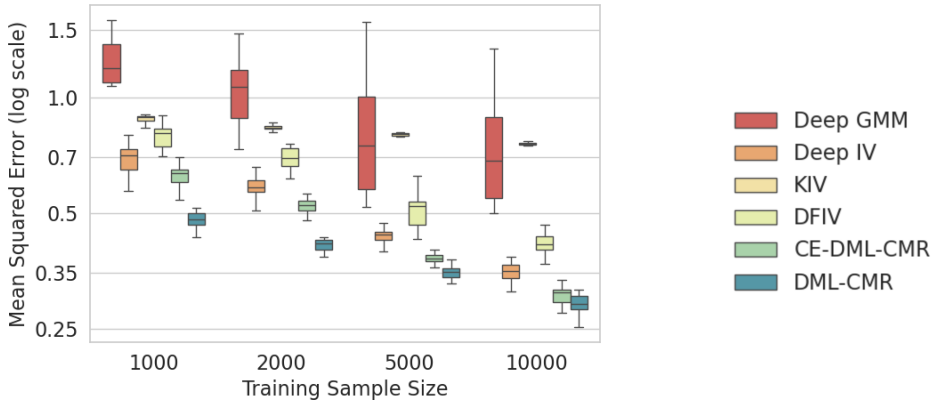

Figure 2: The mean squared error of $\widehat{f}$ on the ticket demand dataset with high-dimensional context for the IV regression task.

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 $f_0((t, s), p) = 100 + (10 + p) \cdot s \cdot \psi(t) - 2p$, 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 E.1.1.

The results for learning f_0 with this dataset of various sizes are provided in Figure 1. It can be seen that DML-CMR performs better than other IV regression methods for all dataset sizes. CE-DML-CMR, which requires significantly less computation, matches the performance of DML-CMR in this case.

5.1.2 High-Dimensional Dataset

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 and Cortes, 2010) to evaluate our methods with high-dimensional (28^2=784 dimensions) inputs. The task remains to learn f_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 2, where DML-CMR and CE-DML-CMR outperform all other methods. In these high-dimensional settings, regularisation is heavily used to avoid overfitting. DML-CMR demonstrates the benefits of using DML to reduce both the regularisation and overfitting bias caused by learning the nuisance parameters.

5.1.3 Real-World Datsets

Lastly, we test the performance of DML-CMR on real-world datasets. The true counterfactual prediction function is rarely available for real-world data. Therefore, in line with previous

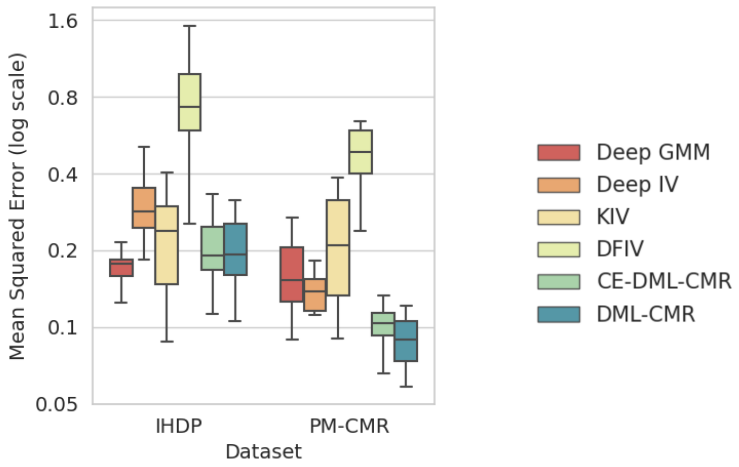

Figure 3: The mean squared error of $\widehat{f}$ on the real-world datasets IHDP and PM-CMR for the IV regression task.

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[3] (Hill, 2011) and PM-CMR[4] (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 E.1.3). As shown in Figure 3, DML-CMR and CE-DML-CMR demonstrate comparable, if not lower, MSE of fitting $\widehat{f}$ than the other methods. This shows that our algorithm is reliable when dealing with real-world data.

5.2 Proximal Causal Learning

For the PCL task (details in Appendix D.4), we compare our methods with PCL methods CEVAE (Im et al., 2021), PMMR (Mastouri et al., 2021), KPV (Mastouri et al., 2021), DFPV (Xu et al., 2021), NMMR U (Kompa et al., 2022), NMMR V (Kompa et al., 2022) and PKDR (Wu et al., 2024). We also use DNN estimators for both stages, with network architecture and hyperparameters provided in Appendix F.

5.2.1 Ticket Demand Dataset

Similarly to IV regression, we start with the ticket demand dataset (Hartford et al., 2017), which has been adapted to the PCL setting (Xu et al., 2021). We aim to understand how ticket prices affect ticket sales and learn the causal function f_0. The hidden confounder in this case is the varying demand U, while the cost of fuel V is the treatment proxy, which directly impacts the ticket price, and the number of views on the airline's reservation website W is the outcome proxy. Details of this dataset are included in Appendix E.2.1.

3. IHDP: https://www.fredjo.com/.
4. PM-CMR: https://doi.org/10.23719/1506014.

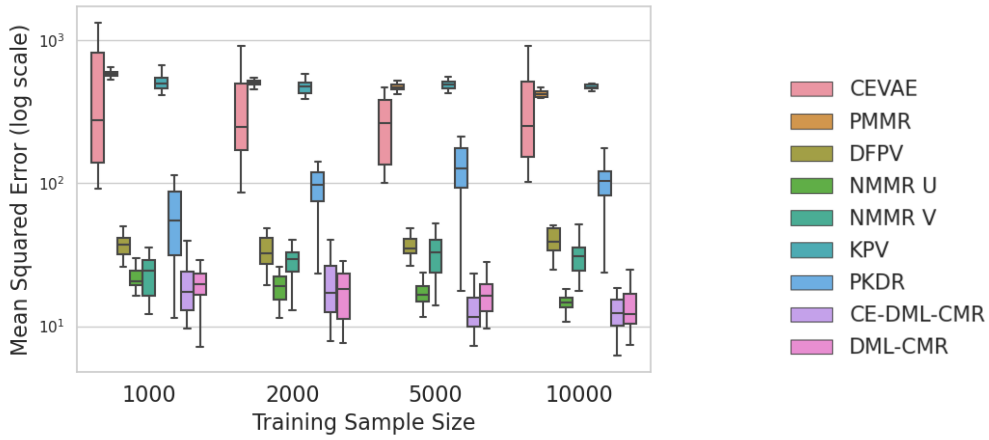

Figure 4: The mean squared error of $\widehat{f}$ on the ticket demand dataset for the PCL task.

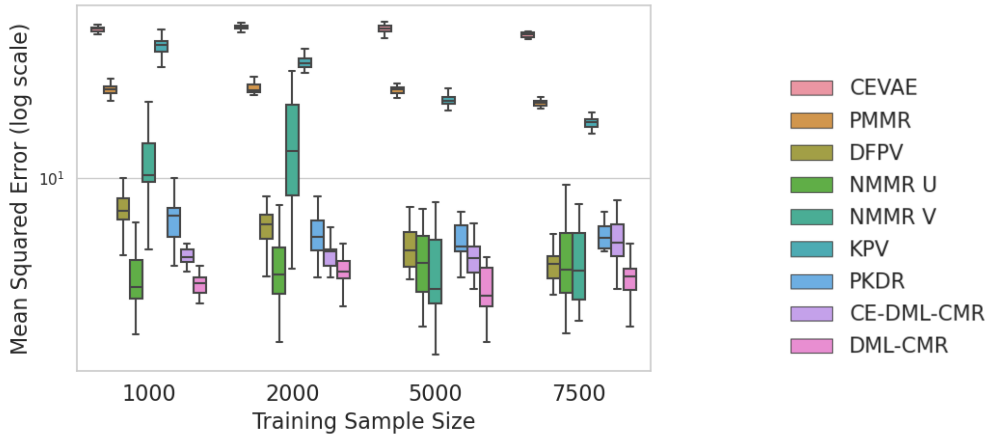

Figure 5: The mean squared error of $\widehat{f}$ on the dSprites dataset with high dimensional treatment for the PCL task.

The results for learning f_0 with this dataset of various sizes are provided in Figure 4. It can be seen that DML-CMR and CE-DML-CMR achieved state-of-the-art performance with very similar performance to each other. NMMR U can match DML-CMR at smaller dataset sizes, but DML-CMR achieves lower MSE at 5000 and 7500 sample sizes. The performance gap between CE-DML-CMR and DML-CMR is small in this case, as expected, because the variables are low-dimensional.

5.2.2 High-dimensional dSprites Dataset

For a high-dimensional dataset, we adopt the dSprites dataset (Matthey et al., 2017) for PCL, first introduced by Xu et al. (2021). dSprites is an image (64×64) dataset where each image is described by five parameters: *shape*, *scale*, *rotation*, *posX* and *poxY*. The treatments are these high-dimensional dSprites images, the hidden confounder is posY, the

proxies are noisy observations of scale, rotation, and posX, and the outcome is defined by a nonlinear causal function. Details of this dataset are provided in Appendix E.2.2.

The results are presented in Figure 5. DML-CMR achieved similar performance to the state-of-the-art methods while outperforming CE-DML-CMR. In addition, a lower variance can be observed when using DML-CMR compared to other methods, especially for smaller data sizes. The NMMR methods in some cases outperform CE-DML-CMR. This is in line with our previous observations that, without k-fold validation, performance can be worse for high-dimensional datasets and the full debiasing effect of the DML framework is required to achieve the best results.

6 Conclusion

We have proposed a novel estimator for solving CMRs, DML-CMR. Using the DML framework and our novel Neyman orthogonal score, DML-CMR can effectively estimate solutions to CMR problems with fast convergence rate guarantees by mitigating the regularisation and overfitting biases in two-stage estimations. We theoretically analysed DML-CMR and proved a convergence rate of $O(N^{-1/2})$ with high probability under mild regularity and parametric assumptions. We also demonstrated interesting connections between the notion of ill-posedness for CMRs and DML's identifiability condition. We applied DML-CMR to problems in causal inference such as IV regression and proximal causal learning, and evaluated it on corresponding benchmarks, including semi-synthetic real-world data. Our experiments demonstrated that DML-CMR achieves state-of-the-art performance against similar algorithms that are developed specifically for the IV regression and PCL problems, as well as general CMR solvers, with lower estimation error and better stability.

Future work includes considering other estimation methods for the nuisance parameters following our Neyman orthogonal score, and theoretically analysing our Neyman orthogonal score for estimating nonparametric functions of interest following (Foster and Syrgkanis, 2019).

Acknowledgments and Disclosure of Funding

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 the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme AISG Award No: AISG2-RP-2020-018, and by the Office of Naval Research (ONR) grant N00014-24 -1-2470. MK and FQ acknowledge 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).

Appendix A. Computationally Efficient CE-DML-CMR

Algorithm 2 Computationally Efficient CE-DML-CMR

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

The standard DML-CMR with K-fold cross-fitting trains $\widehat{s}$ and $\widehat{g}$ K times on different subsets of the dataset to tackle overfitting bias, but is computationally expensive. Therefore, as mentioned in Section 5, we also evaluate CE-DML-CMR, a computationally efficient version of DML-CMR that does not apply K-fold cross-fitting and trains $\widehat{s}$ and $\widehat{g}$ only once using the entire dataset. It uses the same Neyman orthogonal score as the standard DML-CMR, 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 6. CE-DML-CMR can be viewed as a trade-off between computational complexity and theoretical guarantees, and we found that CE-DML-CMR empirically performs as well as standard DML-CMR on low-dimensional datasets, where overfitting bias is not prevalent.

Appendix B. The Score Function for Standard Two-Stage CMR Estimators

In this section, we show that the learning objective, or score function, for standard two-stage CMR estimators (Angrist et al., 1996; Hartford et al., 2017; Singh et al., 2019) is not Neyman orthogonal and thus cannot be used to create a DML estimator for the CMR problem.

Proposition 2 *The score (or objective) function for standard two-stage CMR estimators* $\ell = (Y - \widehat{g}(f, c))^2$ *is not Neyman orthogonal at* (f_0, g_0).

Proof The score $\ell = (Y - \widehat{g}(f, c))^2$ is not Neyman orthogonal because, first of all, $\mathbb{E}[(Y - g_0(f_0, c))^2] = \mathbb{E}[(Y - \mathbb{E}[Y|C])^2] \neq 0$ since $\mathbb{E}[f_0(X)|C] = \mathbb{E}[Y|C]$ and $Y - \mathbb{E}[Y|C] \neq 0$ due to the noise on Y. This violates the basic condition for a Neyman orthogonal score that the score function equals zero with the true functions f_0 and g_0.

Secondly, the Gateaux derivative against small changes in g for score $\mathbb{E}[(Y - g_0(f_0, c))^2]$ at (f_0, g_0) is

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}\Big[(Y - g_0(f_0, C) - r \cdot g(f_0, C))^2\Big] \\
&\quad = \frac{\partial}{\partial r}\mathbb{E}\Big[(Y - g_0(f_0, C))^2 - 2r \cdot (Y - g_0(f_0, C))g(f_0, C) + r^2 \cdot g(f_0, C)^2\Big] \\
&\quad = \mathbb{E}\Big[2(Y - g_0(f_0, C))g(f_0, C) + 2r \cdot g(f_0, C)^2\Big],
\end{aligned}$$

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

$$\mathbb{E}[2(Y - g_0(f_0, c))g(f_0, c)] = \mathbb{E}[2(Y - \mathbb{E}[Y|C])g(f_0, c)]$$

which does not equal to 0 for general $g \in \mathcal{G}$ since generally $g(f_0, c)$ and the residual $(Y - \mathbb{E}[Y|C])$ are correlated. Therefore, this standard score function for two-stage CMR estimation is not Neyman orthogonal at (f_0, g_0). ■

Appendix C. Proofs

In this section, we restate all the conditions required to prove the $N^{-1/2}$ convergence rate guarantees for the DML-CMR estimator, and provide the omitted proofs in the main paper for Theorem 3, Lemma 5, Theorem 6 and Corollary 7.

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

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

Condition 4 [Conditions for $N^{-1/2}$ convergence of DML, Assumption 3.3 and 3.4 in Chernozhukov et al. (2018)]

For sample size $N \geq 3$:

(a) The map $(\theta, (s, g)) \mapsto \mathbb{E}[\psi(\mathcal{D}; f_\theta, (s, g))]$ is twice continuously Gateaux-differentiable.

(b) The score ψ obeys the Neyman orthogonality conditions.

(c) The true parameter θ_0 obeys $\mathbb{E}[\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))] = 0$ and Θ contains a ball of radius $c_1 N^{-1/2} \log N$ centered at θ_0.

(d) For all $\theta \in \Theta$, the identification relationship

$$2\|\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]\| \gtrsim \|J_0(\theta - \theta_0)\|$$

is satisfied, where $J_0 := \partial_{\theta'}\{\mathbb{E}[\psi(\mathcal{D}; f_{\theta'}, (s_0, g_0))]\}|_{\theta'=\theta_0}$ is the Jacobian matrix, with singular values bounded between $c_0 > 0$ and $c_1 > 0$.

(e) All eigenvalues of the matrix $\mathbb{E}[\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))^T]$ are strictly positive (bounded away from zero).

(f) Let K be a fixed integer. Given a random partition $\{I_k\}_{k=1}^K$ of indices $[N]$, each of size $n = N/K$, the nuisance parameter estimator $\widehat{s}_k$ and $\widehat{g}_k$ learnt using data with indices I_k^c belongs to shrinking realisation sets $\mathcal{S}_N$ and $\mathcal{G}_N$, respectively, and the nuisance parameters should be estimated at the $o(N^{-1/4})$ rate, e.g., $\|\widehat{s} - s_0\|_2 = o(N^{-1/4})$.

To formalise the convergence rate guarantees in relationship to the technical conditions, we have the following proposition as a direct result of Theorem 3.3 of Chernozhukov et al. (2018).

Proposition 10 (Theorem 3.3 of Chernozhukov et al. (2018)) *If all conditions in Theorem 4 hold, then the DML estimator $\widehat{\theta}$ as defined in Theorem 1 is concentrated in a $1/\sqrt{N}$ neighbourhood of θ_0, and is approximately linear and centered Gaussian:*

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

where $\bar{\psi}(\cdot) \coloneqq -\sigma^{-1} J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$ is the influence function, J_0 is the Jacobian of ψ, the approximate variance is $\sigma^2 \coloneqq J_0^{-1} \mathbb{E}[\psi(\mathcal{D}, \theta_0, \eta_0)\psi(\mathcal{D}, \theta_0, \eta_0)^T](J_0^{-1})^T$, and the size of the remainder ρ_N converges to 0.

Proof This is a direct consequence of Theorem 3.3 from Chernozhukov et al. (2018), which states the convergence properties of the DML estimator for nonlinear score functions. Assumptions 3.3 and 3.4 in Chernozhukov et al. (2018) required for Theorem 3.3 to hold are contained in Theorem 4. ■

We will show that all of these conditions are satisfied in the proof of Theorem 6. To begin with, we prove Theorem 3, which shows that our score function ψ is Neyman orthogonal.

Theorem 3 (Neyman orthogonality) *The score function $\psi(\mathcal{D}; f, (s, g)) = (s(c) - g(f, c))^2$ obeys the Neyman orthogonality conditions at $(f_0, (s_0, g_0))$.*

Proof Firstly, by Equation (26), we have $s_0(C) = g_0(f_0, C)$, thus

$$\psi(\mathcal{D}; f_0, (s_0, g_0)) = \mathbb{E}\Big[(s_0(C) - g_0(f_0, C))^2\Big] = 0$$

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

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}\Big[(s_0(C) + r \cdot s(C) - g_0(f_0, C) - r \cdot g(f_0, C))^2\Big] \\
=&\frac{\partial}{\partial r}\mathbb{E}\Big[2r(s_0(C) - g_0(f_0, C))(s(C) - g(f_0, C)) + r^2(s(C) - g(f_0, C))^2\Big] \\
=&\mathbb{E}\Big[2(s_0(C) - g_0(f_0, C))(s(C) - g(f_0, C)) + 2r(s(C) - g(f_0, C))^2\Big],
\end{aligned}$$

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

$$\mathbb{E}\Big[2(s_0(C) - g_0(f_0, C))(s(C) - g(f_0, C))\Big] = \mathbb{E}\Big[0 \times (s(C) - g(f_0, C))\Big] = 0 \quad \forall s, g \in \mathcal{S}, \mathcal{G},$$

since $s_0(C) = \mathbb{E}[Y|C] = \mathbb{E}[f_0(X)|C] = g_0(f_0, C)$. Therefore, our moment function ψ is Neyman orthogonal at $(f_0, (s_0, g_0))$. ■

In turn, we state and prove the convergence of the nuisance parameters $\widehat{s}$ and $\widehat{g}$ with the machinery developed for the analysis of the excess risk in constrained ERM in Appendix C.3.

Lemma 5 (Convergence of nuisance parameters) *Under Assumption 2, let $\mathcal{S}_N^*$ be the star-hull of the realisation set $\mathcal{S}_N$ of function class $\mathcal{S}$,*

$$\mathcal{S}_N^* = \{C \mapsto \gamma(s(C) - s_0(C)) \ : \ s \in \mathcal{S}_N, \gamma \in [0,1]\},$$

$\mathcal{P}_N^*$ be the star-hull of the realisation set $\mathcal{P}_N$ of the function class $\mathcal{P}$,

$$\mathcal{P}_N^* = \{C \mapsto \gamma(F(\cdot|C) - F_0(\cdot|C)) \ : \ F \in \mathcal{P}_N, \gamma \in [0,1]\},$$

and $\mathcal{G}_N^*$ be the star-hull of the realisation set $\mathcal{G}_N$ of the function class $\mathcal{G}$,

$$\mathcal{G}_N^* = \{C, f \mapsto \gamma(g(C,f) - g_0(C,f)) \ : \ g \in \mathcal{G}_N, \gamma \in [0,1]\},$$

where $\mathcal{S}_N$, $\mathcal{P}_N$ and $\mathcal{G}_N$ are properly shrinking neighbourhoods of the true functions s_0, F_0 and g_0. Then, there exist universal constants c_1 and c_2, for which we have that with probability at least $1-\xi$, the estimation errors are bounded as

$$\|\widehat{s} - s_0\|_2^2 \leq c_1 \left(\delta_N(\mathcal{S}_N^*)^2 + \sqrt{\frac{\log(1/\zeta)}{N}} \right);$$
$$\|\widehat{g} - g_0\|_2^2 \leq c_2 \left(\delta_N(\mathcal{P}_N^*)^2 + \sqrt{\frac{\log(1/\zeta)}{N}} \right).$$

Proof The bound on estimation error for $\widehat{s}$ is straightforward and follows directly by Theorem 14. In order to prove that, we only need to show that the choice of loss function for the ERM estimator $\widehat{s}$ is Lipschitz in its first argument. To this end, we formulate the ERM as

$$\widehat{s} = \arg\min_{s \in \mathcal{S}} \mathcal{L}_N(s), \quad \text{where } \mathcal{L}_N(s) = \frac{1}{N} \sum_{i=1}^{N} \ell(s; y_i, c_i) \text{ and } \ell(s; y, c) = (y - s(c))^2.$$

Thus, we have,

$$\begin{aligned} |\ell(s_2; y, c_2) - \ell(s_1; y, c_1)| &= \left| (y - s_2(c_2))^2 - (y - s_1(c_1))^2 \right| \\ &= |(2y + s_2(c_2) + s_1(c_1))(s_2(c_2) - s_1(c_1))| \\ &\leq 4B |s_2(c_2) - s_1(c_1)|, \end{aligned} \tag{7}$$

where we used the fact that $y, \|s\|_\infty \leq B$ by Assumption 2. Thus, $\ell(s; y, c)$ is $4B$-Lipschitz in its argument and therefore, by Theorem 14, for a universal constant c_1,

$$\|\widehat{s} - s_0\|_2^2 \leq c_1 \left(\delta_N(\mathcal{S}_N^*)^2 + \sqrt{\frac{\log(1/\zeta)}{N}} \right).$$

For the estimation error of $\widehat{g}$, we recall that for any $F(\cdot|C) \in \mathcal{P}$,

$$g_F(f_\theta, c) \coloneqq \int f_\theta(x) F(x|C=c)dx.$$

Thus, we can connect the estimation error of $\widehat{g}_{\widehat{F}}$ to the estimation error of the conditional density estimator F by showing

$$\begin{aligned} \|\widehat{g} - g_0\|_2 &= \|\widehat{g}_{\widehat{F}} - g_{F_0}\|_2 \\ &\leq B\|\widehat{F} - F_0\|_2 \end{aligned} \tag{8}$$

To prove (8), we observe that for any C and any test function $f \in \mathcal{F}$, we have that,

$$\begin{aligned} |g_{F_1}(C,f) - g_{F_2}(C,f)| &= \left| \int f(x)[F_1 - F_2](dx|C) \right| \\ &\leq \int |f(x)||F_1 - F_2|(dx|C) \\ &\leq B \int |F_1 - F_2|(dx|C), \end{aligned} \tag{9}$$

since $\|f\|_\infty \leq B$ by Assumption 2. Thus, by integrating w.r.t. joint law of (C, f), and (9), we can show that (8) holds since

$$\begin{aligned} \|g_{F_1} - g_{F_2}\|_2^2 &= \mathbb{E}_C\Big[\big(g_{F_1}(C,f) - g_{F_2}(C,f)\big)^2\Big] \\ &\leq B^2 \mathbb{E}_C\left[\left(\int |F_1 - F_2|(dx|C)\right)^2\right] \\ &\leq B^2 \mathbb{E}_C\left[\int [F_1 - F_2]^2(dx|C)\right] \\ &= B^2 \|F_1 - F_2\|_2^2, \end{aligned} \tag{10}$$

for any $F_1, F_2 \in \mathcal{P}$, where in (10), we used Cauchy–Schwarz in dx, i.e.,

$$\left(\int |h|\right)^2 \leq \int h^2.$$

Having (8) at hand, it suffices to prove the upper bound on the estimation error of $\widehat{F}$. To this end, we observe that Squared CDF error,

$$\ell(F; x, c) = [\mathbb{1}_{\{X \leq x \mid C=c\}} - \mathscr{F}(x|c)]^2,$$

is Lipschitz by a similar argument to that for (7), where $\mathscr{F}(x|c)$ is the conditional CDF induced by the conditional density F. It is not hard to observe that *clipped* versions of other losses for density estimation such as integrated squared error (ISE), negative log-likelihood, and Hellinger-squared also satisfy the Lipschitz condition in the first argument. Thus, by Theorem 14, for a universal constant c_3,

$$\|\widehat{F} - F_0\|_2^2 \leq c_3 \left(\delta_N(\mathcal{P}_{N^*})^2 + \sqrt{\frac{\log(1/\zeta)}{N}} \right).$$

Choosing $c_2 = B^2 c_3$ and (8) completes the proof. ■

Now, we are ready to prove Theorem 6, which is the main theorem that states the $N^{-1/2}$ convergence rate guarantees for our DML estimator.

Theorem 6 (Convergence of the DML estimator for CMRs) *Let $f_{\theta_0} \in \mathcal{F}$ be a solution that satisfies the CMRs in Equation (P), let ψ be the Neyman orthogonal score defined in Equation (5) and let $J_0 := \partial_{\theta'}\{\mathbb{E}[\psi(\mathcal{D}; f_{\theta'}, (s_0, g_0))]\}|_{\theta'=\theta_0}$ be the Jacobian matrix of $\mathbb{E}[\psi]$ w.r.t. θ. Suppose that the upper bound of the critical radius $\delta_N = o(N^{-1/4})$, for $\widehat{s}$, $\widehat{g}$, and J_0 has bounded singular values. Then, if Assumption 1 and 2 hold, our DML estimator $f_{\widehat{\theta}}$ satisfies that $\widehat{\theta}$ is concentrated in a $N^{-1/2}$ neighbourhood of θ_0, and is approximately linear and centred Gaussian:*

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

where the estimator variance is given by

$$\sigma^2 := J_0^{-1}\mathbb{E}[\psi(\mathcal{D}, \theta_0, (s_0, g_0))\psi(\mathcal{D}, \theta_0, (s_0, g_0))^T](J_0^{-1})^T,$$

which is constant w.r.t. N.

Proof Following Proposition 10, we need to check whether, under Assumption 2, all of Condition 4 for DML $N^{-1/2}$ convergence rate is satisfied. Condition (a) is satisfied since $(s-g)^2$ is twice continuously differentiable with respect to s and g. Condition (b) is satisfied by Theorem 3. Condition (c) is satisfied since f_{θ_0} satisfies the CMRs and, from Theorem 3, we have that $\mathbb{E}[\psi(\mathcal{D}; f_{\theta_0}, (s_0, g_0))] = 0$. In addition, from Assumption 1, since the true parameter $\theta_0 \in \Theta$ is in the interior of Θ, Θ contains some neighbourhood centered at the true parameter θ_0.

Condition (d) is a sufficient identifiability condition, which states that the closeness of the score function at point θ to zero implies the closeness of θ to θ_0. This assumption is standard in conditional moment problems and implies that the *ill-posedness* (see Definition 8) of the CMR problem is bounded, as shown in Section 4.4. To check condition (d), 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 $\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]$ around the point θ_0,

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

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

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

Now for identifiability, we only need to check that $J_0 J_0^T$ is non-singular, which is guaranteed by bounded singular value of J_0 as stated in the Theorem.

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

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

By trace trick, for each data point $\mathcal{D}$, the only eigenvalue of $\psi(\mathcal{D}; f_\theta, (s_0, g_0))\psi(\mathcal{D}; f_\theta, (s_0, g_0))^T$ is $\|\psi(\mathcal{D}; f_\theta, (s_0, g_0))\|^2 \geq 0$, with $\psi(\mathcal{D}; f_\theta, (s_0, g_0))$ as the corresponding eigenvector. Therefore, $\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))\psi(\mathcal{D}; f_\theta, (s_0, g_0))^T]$ 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 $\psi(\mathcal{D}; f_\theta, (s_0, g_0))$, which is true in our setting as the co-domain of $\psi(\mathcal{D}; f_\theta, (s_0, g_0))$ is $\mathbb{R}$.

Condition (f) is satisfied since we have the critical radius $\delta_N = o(N^{-1/4})$, and together with Lemma 5, the nuisance parameters converge sufficiently quickly to ensure $\|\widehat{s} - s_0\|_2 \leq O(\delta_N + N^{-1/2}) = O(o(N^{-1/4}) + N^{-1/2}) = o(N^{-1/4})$ and similarly $\|\widehat{g} - g_0\|_2 \leq O(\delta_N + N^{-1/2}) = o(N^{-1/4})$.

Therefore, all the conditions in Condition 4 are satisfied, which concludes the proof by Proposition 10. ■

Corollary 7 *Let $f_{\widehat{\theta}}$ be the DML estimator for CMRs. If all assumptions for Theorem 6 hold and there exists a constant $L > 0$ such that $\|f_\theta(x) - f_{\theta_0}(x)\|_2 \leq L\|\theta - \theta_0\|_2$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$, then for all $\zeta \in (0, 1]$, we have that*

$$\|f_{\widehat{\theta}} - f_{\theta_0}\|_2 = O\left(L\sqrt{\frac{\ln(1/\zeta)}{N}}\right),$$

with probability $1 - \zeta$.

Proof From theorem 6, we have that the parameters $\widehat{\theta}$ for our DML estimator $f_{\widehat{\theta}}$ learnt from a dataset of size N satisfy $(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2/N)$, where σ^2 is the DML estimator's variance. This means that, for all $\epsilon > 0$ and $\zeta \in (0, 1]$, there exists an integer $K > 0$ such that for all $N \geq K$,

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

where Φ is the CDF of a standard Gaussian distribution. If we assume L to be a constant such that $\|f_\theta(x) - f_{\theta_0}(x)\| \leq L\|\theta - \theta_0\|$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$, we have that for all $\epsilon > 0$ and $\zeta \in (0, 1]$, there exists an integer $K > 0$ such that for all $N \geq K$,

$$\begin{aligned} &\mathbb{P}\left(\|f_{\widehat{\theta}}(x) - f_{\theta_0}(x)\| > L \cdot \epsilon\right) \leq 1 - \Phi(\epsilon \cdot \sqrt{N}/\sigma) + \zeta/2 \quad \forall x \in \mathcal{X}, \\ \implies &\mathbb{P}\left(\|f_{\widehat{\theta}}(x) - f_{\theta_0}(x)\| \leq L \cdot \epsilon\right) \geq \Phi(\epsilon \cdot \sqrt{N}/\sigma) - \zeta/2 \quad \forall x \in \mathcal{X}, \end{aligned}$$

Now, for any $\zeta \in (0, 1]$, we can choose $\epsilon > 0$ such that $\Phi(\epsilon \cdot \sqrt{N}/\sigma) = 1 - \zeta/2$ since $0.5 \leq 1 - \zeta/2 < 1$, and by substituting ϵ out of the above equation, we have that

$$\mathbb{P}\left(\|f_{\widehat{\theta}} - f_{\theta_0}\|_2 \leq L \cdot \Phi^{-1}(1 - \zeta/2)\sigma/\sqrt{N}\right) \geq 1 - \zeta.$$

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 $N > K$,

$$\begin{aligned}\|f_{\widehat{\theta}} - f_{\theta_0}\|_2 \leq L \cdot \Phi^{-1}(1-\zeta/2)\sigma/\sqrt{N} &\leq L\sigma\sqrt{-2\ln(\zeta/2)}/\sqrt{N} \\ &= L\sigma\sqrt{2\ln(2/\zeta)}/\sqrt{N} \\ &= \sqrt{2}L\sigma\sqrt{\frac{\ln(2/\zeta)}{N}} \quad \text{with probability } 1-\zeta,\end{aligned}$$

which completes the proof. ■

C.2 Ill-posedness and DML Identification

Proposition 9 *For all $\theta \in \Theta$, if there exists a constant $L > 0$ such that $\|f_\theta(x) - f_{\theta_0}(x)\|_2 \leq L\|\theta - \theta_0\|_2$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$, then Condition 4 (d), which states*

$$2\|\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]\| \geq \|J_0(\theta - \theta_0)\|$$

and the Jacobian matrix J_0 have singular values bounded between $c_0 > 0$ and $c_1 > 0$, implies the ill-posedness is bounded by $\nu \leq L/\sqrt{c_0}$.

Proof Recall that our score function is $\psi(\mathcal{D}; f_\theta, (s, g)) = (s(c) - g(f, c))^2$ where $\psi(\mathcal{D}; f_\theta, (s_0, g_0)) = (s_0(c) - g_0(f, c))^2 = (\mathbb{E}[Y - f(X)|C])^2$. Under a finite-dimensional parameterised setting, we have that from $2\|\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]\| \gtrsim \|J_0(\theta - \theta_0)\|$,

$$\begin{aligned}\|\mathbb{E}[f_{\theta_0}(X) - f_\theta(X)|C]\|_2^2 =& \|\mathbb{E}[(\mathbb{E}[f_{\theta_0}(X) - f_\theta(X)|C])^2]\| \\ =& \|\mathbb{E}[(\mathbb{E}[Y - f_\theta(X)|C])^2]\| \\ =& \|\mathbb{E}[\psi(\mathcal{D}; f_\theta, (s_0, g_0))]\| \\ \geq& \frac{1}{2}\|J_0(\theta - \theta_0)\| \\ =& \sqrt{(\theta - \theta_0)^T(J_0^T J_0)(\theta - \theta_0)} \\ \geq& \frac{1}{2}\sqrt{c_0^2\|(\theta - \theta_0)\|_2^2} \geq \frac{1}{2}c_0\|(\theta - \theta_0)\|_2 \geq \frac{1}{2}c_0\|(\theta - \theta_0)\|_2^2 \qquad (11)\end{aligned}$$

for $\|(\theta - \theta_0)\| \leq 1$ and the singular value lower bound $c_0 > 0$ of J_0. With a local Lipschitz condition of f_θ around θ_0: $\|f_\theta(x) - f_{\theta_0}(x)\|_2 \leq L\|\theta - \theta_0\|_2$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$, we have that

$$\begin{aligned}\|f_{\theta_0} - f_\theta\|_2^2 &= \mathbb{E}[(f_{\theta_0}(x) - f_\theta(x))^2] \\ &\leq \mathbb{E}[(L\|\theta_0 - \theta\|)^2] \\ &\leq L^2\|\theta_0 - \theta\|^2 \\ \implies \|\theta_0 - \theta\|^2 &\geq \frac{\|f_{\theta_0} - f_\theta\|_2^2}{L^2} \qquad (12)\end{aligned}$$

Therefore, from Equation (11) and Equation (12), we have that

$$\|\mathbb{E}[f_{\theta_0}(X) - f_\theta(X)|C]\|_2^2 \geq c_0\|(\theta-\theta_0)\|_2^2 \geq \frac{c_0\|f_{\theta_0} - f_\theta\|_2^2}{L^2}$$
$$\implies \|\mathbb{E}[f_{\theta_0}(X) - f_\theta(X)|C]\|_2 \geq \sqrt{c_0}\|(\theta-\theta_0)\|_2 \geq \frac{\sqrt{c_0}\|f_{\theta_0} - f_\theta\|_2}{L}$$

which bounds the ill-posedness by

$$\nu = \sup_{f\in\mathcal{F}} \frac{\|f_{\theta_0} - f_\theta\|_2}{\|\mathbb{E}[f_{\theta_0}(X) - f_\theta(X)|C]\|_2} \leq \frac{L}{\sqrt{c_0}}. \tag{13}$$

■

C.3 Constrained Empirical Risk Minimisation Bounds

In this section, we introduce basic concepts from empirical process theory and further discuss bounds on the excess risk of general Empirical Risk Minimizer (ERM) in the style of Wainwright (2019); Foster and Syrgkanis (2019).

Definition 11 *The critical radius denoted by $\delta_N(\mathcal{H}^*)$ is defined as the minimum δ that satisfies the following upper bound on the local Gaussian complexity of a star-shaped function class $\mathcal{H}^*$[5], $\mathcal{G}(\mathcal{H}^*, \delta) \leq \delta^2/2$, where local Gaussian complexity is defined as*

$$\mathcal{G}(\mathcal{H}^*, \delta) = \mathbb{E}_\epsilon\left[\sup_{h\in\mathcal{H}^*:\|h\|_N\leq\delta} \langle\epsilon, h\rangle\right], \tag{14}$$

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

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,

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

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

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

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

Thus, the critical radius of $\mathcal{H}^*$ will be upper bounded by

$$\delta_N(\mathcal{H}^*) \lesssim \frac{\sqrt{\log \mathcal{N}(\mathcal{H}^*, L^2(\mathbb{P}_N), \epsilon)}}{\sqrt{N}} = O(d_N(\mathcal{H}^*)^{1/2} N^{-1/2}),$$

where Chernozhukov et al. (2022b, 2021) referred to

$$d_N(\mathcal{H}^*) \coloneqq \inf\left\{ d > 0 : \log \mathcal{N}(\mathcal{H}^*, L^2(\mathbb{P}_N), \epsilon) \leq d \log\left(\frac{C}{\epsilon}\right) \ \ \forall \epsilon \in (0,1) \text{ and } C \text{ is a constant.} \right\},$$

as the effective dimension of the hypothesis space. Note that this rate matches the minimax lower bound of fixed design estimation for this setting (Yang and Barron, 1999).

Given the dataset $\mathcal{D} = \{z_i \in \mathcal{Z}\}_{i=1}^N$ consisting of i.i.d. data points z_i drawn from distribution $\mathbb{P}$, and a function class $\mathcal{H}$, we define the realisation of a function space by subscript N, e.g., $\mathcal{H}_N$ is the realisation of $\mathcal{H}$ in the N observed data points. Since the definition of critical radius is for star-shaped function classes, we equip ourselves with the star-hull notation, where $\mathcal{H}_N^*$ is the star-hull of the function class $\mathcal{H}_N$ centred at the true function h_0, defined as

$$\mathcal{H}_N^* \coloneqq \{Z \mapsto \gamma(h(X) - h_0(Z)) \ : \ f \in \mathcal{H}_N, \gamma \in [0,1]\},$$

and denote its critical radius by $\delta_N(\mathcal{H}_N^*)$, or simply $\delta_N(\mathcal{H})$. In statistical learning, we are given a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, and we have the risk and empirical risk defined accordingly as:

$$\mathcal{L}(h) = \mathbb{E}_z[\ell(h; z)] \quad \text{and} \quad \mathcal{L}_N(h) = \frac{1}{N}\sum_{i=1}^N \ell(h; z_i).$$

Let us denote the ground truth function by h_0, i.e., the minimizer of the risk,

$$h_0 = \arg\min_{h \in \mathcal{H}} \mathcal{L}(h).$$

The ERM algorithm proposes the estimator $\widehat{h}$ that minimises the empirical risk of the observed N data points,

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \mathcal{L}_N(h).$$

Theorem 12 (Foster and Syrgkanis (2019, Lemma 7)) *Consider a function class $\mathcal{H}$ and its star-hull $\mathcal{H}^*$, with $\sup_{h \in \mathcal{H}^*} \|h\|_\infty \leq 1$, critical radius $\delta_N(\mathcal{H}^*)$, and any choice of δ such that,*

$$\delta^2 \geq \max\left\{ \delta_N(\mathcal{H}^*)^2, \frac{4 \log(41 \log(2 c_1 N))}{c_1 N} \right\},$$

for a constant c_1. Moreover, assume that the loss function ℓ is L-Lipschitz in its first argument with respect to ℓ_2 norm. Then there exist universal constants c_2 and c_3 such that with probability at least $1 - c_2 \exp\{c_3 N \delta^2\}$,

$$|(\mathcal{L}_N(h) - \mathcal{L}_N(h_0)) - (\mathcal{L}(h) - \mathcal{L}(h_0))| \leq 18 L \delta \{\|h - h_0\|_2 + \delta\}, \qquad \forall h \in \mathcal{H}.$$

We mention an immediate corollary of Theorem 12 for functions h that are bounded by B, which follows by rescaling arguments.

Corollary 13 *Consider a function class $\mathcal{H}$ and its star-hull $\mathcal{H}^*$, with $\sup_{h\in\mathcal{H}^*}\|h\|_\infty \le B$, critical radius $\delta_N(\mathcal{H}^*)$, and any choice of δ such that,*

$$\delta^2 \ge \max\left\{\delta_N(\mathcal{H}^*)^2, \frac{4\log(41\log(2c_1N))}{c_1N}\right\},$$

for a constant c_1. Moreover, assume that the loss function ℓ is L-Lipschitz in its first argument with respect to ℓ_2 norm. Then there exist universal constants c_2 and c_3 such that with probability at least $1 - c_2\exp\{c_3N\delta^2/\max\{1,B\}^2\}$,

$$|(\mathcal{L}_N(h) - \mathcal{L}_N(h_0)) - (\mathcal{L}(h) - \mathcal{L}(h_0))| \le \frac{18L\delta}{\max\{1,B\}}\{\|h-h_0\|_2 + \delta\}, \qquad \forall h \in \mathcal{H}.$$

Proof If $B \le 1$, the proof follows trivially from Theorem 12, thus let us assume otherwise. Define the B-scaled function class $\widetilde{\mathcal{H}}$,

$$\widetilde{\mathcal{H}} \coloneqq \left\{\widetilde{h} \;:\; B\widetilde{h} \in \mathcal{H}\right\}.$$

Then, the loss function,

$$\ell(h,z) = \ell(B\widetilde{h}, z),$$

is LB-Lipschitz in $\widetilde{h}$ and by homogeneity of the local Gaussian averages, i.e.,

$$\mathcal{G}(\widetilde{\mathcal{H}}^*, r) = \frac{1}{B}\mathcal{G}(\mathcal{H}^*, Br),$$

wee see that $\widetilde{\delta} \coloneqq \delta/B$ satisfies the fixed point condition (14). Putting these pieces together and invoking Theorem 12 for $\widehat{\mathcal{H}}$ completes the proof. ■

We can equivalently write Corollary 13 in the failure probability format. That is, for a target failure probability $0 < \xi < 1$, define

$$\delta(\xi) \coloneqq \delta_N(\mathcal{H}^*) + \max\{1,B\}\sqrt{\frac{1}{c_3N}\log(1/\xi)},$$

then with probability at least $1-\xi$,

$$|(\mathcal{L}_N(h) - \mathcal{L}_N(h_0)) - (\mathcal{L}(h) - \mathcal{L}(h_0))| \le \frac{18L\delta(\xi)}{\max\{1,B\}}\{\|h-h_0\|_2 + \delta(\xi)\}, \qquad \forall h \in \mathcal{H}. \quad (15)$$

Now, we are ready to state and prove the following master theorem for the excess risk of the constrained ERM. For the analysis, in line with Chernozhukov et al. (2021), we require that the population risk has positive curvature for identifiability purposes. Then, the generalisation bound in terms of excess risk can be converted into estimation error.

Theorem 14 (Estimation Error of Constrained ERM) *Assume that the population risk $\mathcal{L}$ has a positive curvature, i.e., for a positive number λ,*

$$\mathcal{L}(h) - \mathcal{L}(h_0) \geq \frac{\lambda}{2}\|h - h_0\|_2^2 \qquad \forall h \in \mathcal{H}, \tag{16}$$

and the loss function is bounded,

$$|\ell(h; z)| \leq M,$$

and L-Lipschitz in its first argument w.r.t. ℓ_2 norm. Then, the solution to the ERM algorithm:

$$\widehat{h} := \arg\min_{h \in \mathcal{H}} \mathcal{L}_N(h), \tag{17}$$

has the following estimation error with probability at least $1 - \xi$,

$$\|\widehat{h} - h_0\|_2^2 \leq C\left[(\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \delta_N(\mathcal{H}^*)^2 + \sqrt{\frac{\log(1/\xi)}{N}}\right].$$

for a universal constant C, where

$$h^* = \arg\inf_{h \in \mathcal{H}} \mathcal{L}(h).$$

In addition, if h_0 is realisable, i.e., $h_0 \in \mathcal{H}$, then,

$$\|\widehat{h} - h_0\|_2^2 \leq C\left[\delta_N(\mathcal{H}^*)^2 + \sqrt{\frac{\log(1/\xi)}{N}}\right].$$

Proof First, we want to upper bound the population excess risk,

$$\mathcal{L}(\widehat{h}) - \mathcal{L}(h_0),$$

and then, by the curvature of $\mathcal{L}$, we can convert this bound into an upper bound on the estimation error $\|\widehat{h} - h_0\|_2$.

To begin with, we have,

$$\mathcal{L}(\widehat{h}) - \mathcal{L}(h_0) = \underbrace{\left(\mathcal{L}(\widehat{h}) - \mathcal{L}_N(\widehat{h})\right)}_{\text{(I)}} + \underbrace{\left(\mathcal{L}_N(\widehat{h}) - \mathcal{L}_N(h^*)\right)}_{\text{(II)}} + \underbrace{\left(\mathcal{L}_N(h^*) - \mathcal{L}(h_0)\right)}_{\text{(III)}}.$$

By ERM algorithm (17), we know that,

$$\mathcal{L}_N(\widehat{h}) \leq \mathcal{L}_N(h^*),$$

Hence, (II) ≤ 0. For the term (I),

$$\begin{aligned}\text{(I)} &= \mathcal{L}(\widehat{h}) - \mathcal{L}_N(\widehat{h}) \\ &= (\mathcal{L} - \mathcal{L}_N)(\widehat{h} - h_0) + (\mathcal{L} - \mathcal{L}_N)(h_0),\end{aligned}$$

and for the term (III), we know that,

$$\begin{aligned}(\text{III}) &= \mathcal{L}_N(h^*) - \mathcal{L}(h_0) \\ &= (\mathcal{L}(h^*) - \mathcal{L}(h_0)) + (\mathcal{L}_N - \mathcal{L})(h^*) \\ &= (\mathcal{L}(h^*) - \mathcal{L}(h_0)) + (\mathcal{L}_N - \mathcal{L})(h^* - h_0) + (\mathcal{L}_N - \mathcal{L})(h_0).\end{aligned}$$

Summing these terms from (I), (II) and (III) yields,

$$\mathcal{L}(\widehat{h}) - \mathcal{L}(h_0) \leq (\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \underbrace{(\mathcal{L} - \mathcal{L}_N)(\widehat{h} - h_0)}_{\epsilon_1(\widehat{h})} + \underbrace{(\mathcal{L}_N - \mathcal{L})(h^* - h_0)}_{\epsilon_2}.$$

From Corollary 13, especially the formulation in (15), we know that with probability at least $1 - \xi/2$,

$$\sup_{h \in \mathcal{H}} |(\mathcal{L}_N - \mathcal{L})(h - h_0)| \leq \frac{18L\delta(\xi/2)}{\max\{1, B\}}(\|h - h_0\|_2 + \delta(\xi/2)),$$

where,

$$\delta(\xi/2) = \delta_N(\mathcal{H}^*) + \max\{1, B\}\sqrt{\frac{1}{c_3 N}\log(2/\xi)}. \tag{18}$$

Thus, we can upper bound $\epsilon_1(\widehat{h})$ as,

$$\epsilon_1(\widehat{h}) = (\mathcal{L} - \mathcal{L}_N)(\widehat{h} - h_0) \leq \frac{18L\delta(\xi/2)}{\max\{1, B\}}\Big(\|\widehat{h} - h_0\|_2 + \delta(\xi/2)\Big).$$

To bound the term ϵ_2, define the random variable,

$$X_i = \ell(h^*; z_i) - \ell(h_0; z_i),$$

where X_i is bounded as $|X_i| \leq 2M$. Then, by Hoeffding's inequality, with probability at least $1 - \xi/2$, we infer that

$$\epsilon_2 = (\mathcal{L}_N - \mathcal{L})(h^* - h_0) \leq 2\sqrt{2}M\sqrt{\frac{\log(4/\xi)}{N}}.$$

Thus, by union bound and (16), we conclude that with probability at least $1 - \xi$

$$\begin{aligned}\frac{\lambda}{2}\|\widehat{h} - h_0\|_2^2 &\leq (\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \frac{18L\delta(\xi/2)}{\max\{1, B\}}\Big(\|\widehat{h} - h_0\|_2 + \delta(\xi/2)\Big) + 2\sqrt{2}M\sqrt{\frac{\log(4/\xi)}{N}} \\ &= (\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \frac{18L\delta(\xi/2)}{\max\{1, B\}}\|\widehat{h} - h_0\|_2 + \frac{18L\delta(\xi/2)^2}{\max\{1, B\}} + 2\sqrt{2}M\sqrt{\frac{\log(4/\xi)}{N}} \\ &\leq (\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \frac{\lambda}{4}\|\widehat{h} - h_0\|_2^2 + \frac{1}{\lambda}\left(\frac{18L\delta(\xi/2)}{\max\{1, B\}}\right)^2 \\ &\quad + \frac{18L\delta(\xi/2)^2}{\max\{1, B\}} + 2\sqrt{2}M\sqrt{\frac{\log(4/\xi)}{N}}\end{aligned} \tag{19}$$

where (19) follows by Young's inequality, i.e., for any positive number $\lambda > 0$,

$$ax \leq \frac{\lambda}{4}x^2 + \frac{a^2}{\lambda}.$$

Therefore, by cancelling out $\frac{\lambda}{4}\|\widehat{h} - h_0\|_2^2$ from both sides,

$$\|\widehat{h} - h_0\|_2^2 \leq C_1 \left[\frac{1}{\lambda}(\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \left(\frac{L^2}{\lambda \max\{1, B\}^2} + \frac{L}{\max\{1, B\}}\right)\delta(\xi/2)^2 + M\sqrt{\frac{\log(1/\xi)}{N}}\right],$$

for a constant C_1. In turn, by (18) and Young's inequality, we know that,

$$\delta(\xi/2)^2 \leq C_2 \left[\delta_N(\mathcal{H}^*)^2 + \frac{\max\{1, B\}^2}{N}\log(1/\xi)\right],$$

for a constant C_2. Hence,

$$\begin{aligned}
\|\widehat{h} - h_0\|_2^2 \leq C_3 \Bigg[&\frac{1}{\lambda}(\mathcal{L}(h^*) - \mathcal{L}(h_0)) \\
&+ \left(\frac{L^2}{\lambda \max\{1, B\}^2} + \frac{L}{\max\{1, B\}}\right)\left[\delta_N(\mathcal{H}^*)^2 + \max\{1, B\}^2\frac{\log(1/\xi)}{N}\right] \\
&+ M\sqrt{\frac{\log(1/\xi)}{N}}\Bigg] \\
\leq C_4 \Bigg[&\frac{1}{\lambda}(\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \frac{\max\{L^2, L\}}{\min\{\lambda, 1\}\max\{1, B\}}\delta_N(\mathcal{H}^*)^2 \\
&+ \frac{\max\{L^2, L\}\max\{1, B\}}{\min\{\lambda, 1\}}\frac{\log(1/\xi)}{N} + M\sqrt{\frac{\log(1/\xi)}{N}}\Bigg], \qquad (20)
\end{aligned}$$

for constants C_3 and C_4.

Therefore, by hiding dependence on the constants λ, B, L and M in a universal constant C, we can summarise the result (20) and conclude the proof,

$$\|\widehat{h} - h_0\|_2^2 \leq C\left[(\mathcal{L}(h^*) - \mathcal{L}(h_0)) + \delta_N(\mathcal{H}^*)^2 + \sqrt{\frac{\log(1/\xi)}{N}}\right].$$

■

Appendix D. Examples of CMR Problems

There are many concrete problems in statistical estimation, causal inference, and econometrics that are, in fact, CMR problems (see Carrasco et al. (2007), Section 1.3, for nine concrete CMR problems). In this section, we introduce in detail two CMR problems in causal inference, IV regression and proximal causal learning, which we evaluated experimentally in Section 5. To begin with, we provide a brief introduction of hidden confounders and structural causal models.

D.1 Hidden Confounders

Hidden confounders (Pearl, 2000) are unobserved variables that influence both the *actions* (or *interventions*) and the *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 and Pearl, 2008). To formalise the concept of hidden confounders and provide a framework for specifying the underlying causal mechanisms in a data-generating process, we introduce structural causal models (SCMs).

D.2 Structural Causal Model

Definition 15 (Structural Causal Model) *An SCM M is a tuple $(U, V, F, P(U))$, where U is a set of exogenous (i.e. outside the model) random variables, which are typically unobserved; V is a set of endogenous (i.e. inside the model) variables; $F = f_i$ is a set of deterministic functions where, for each $V_i \in V$, $f_i(pa_i, u_i) = v_i$ (pa_i denotes the parent of V_i and U_i are exogenous variables linked to V_i). $P(U)$ is the joint distribution of exogenous variables.*

In this definition, endogenous variables are variables on which we would like to study the causal relationships (e.g., between price and revenue). Exogenous variables are external sources of noise (e.g., seasonality) that can confound the causal relationships between endogenous variables. Next, we introduce the concept of causal interventions, which are tools that allow us to study causal effects between variables. Interventions are defined through a mathematical operator called $do(x)$ (Pearl, 2000). An intervention, denoted by $do(X = x)$, simulates a physical intervention by removing the natural dependencies of X on its parent variables in the SCM and forcing it to take a specific value x, while keeping the rest of the model unchanged. The resulting causal model after the intervention is denoted M_x. The post-intervention distribution resulting from the intervention $do(X = x)$ is given by the equation

$$\mathbb{P}_M(y|do(x)) = \mathbb{P}_{M_x}(y), \tag{21}$$

where the post-intervention distribution of some variable Y is defined as the distribution of Y in the intervened model M_x.

For example, in a causal model where A (treatment) affects Y (outcome), an observational study may show correlation, but an intervention $do(A = a)$ would simulate a randomised experiment, ensuring that changes in Y are due to A and not other confounders.

In the SCM formulation, an exogenous random variable is considered a hidden confounder if it affects two or more endogenous variables, e.g., V_i and V_j, and is unobserved. Consider a SCM that specifies two endogenous variables, the outcome $Y \in \mathcal{Y}$ and the treatment $A \in \mathcal{A}$:

$$Y = f(A, U), \tag{22}$$

where $U \in \mathcal{U}$ is a hidden confounder that affects both A and Y, as illustrated in the causal graph depicted in Figure 6. Due to the presence of this hidden confounder, with

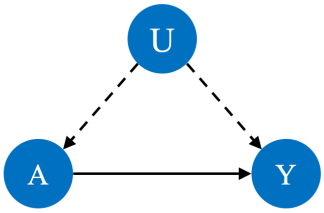

Figure 6: The causal graph of outcome Y, treatment A and hidden confounder U.

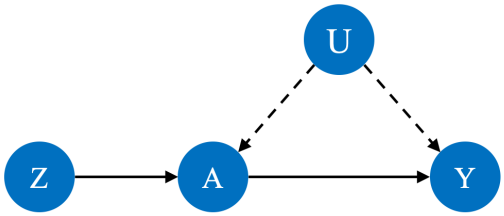

Figure 7: The causal graph of outcome Y, treatment A, hidden confounder U and an instrumental variables Z.

only observational data, standard regressions (e.g., ordinary least squares) generally fail to produce consistent estimates of the causal relationship (also known as the average treatment effect) between A and Y (Pearl, 2000), i.e., $\mathbb{E}[Y \mid do(A)]$, where $do(\cdot)$ is the interventional operator. Therefore, the ability to identify the causal relationship between A and Y requires additional assumptions. Two classic techniques are IV regression (Newey and Powell, 2003) and proximal causal learning, each with the explicit assumption to observe additional variables in the model that help the identification of $\mathbb{E}[Y \mid do(A)]$, which we will introduce next.

D.3 Instrumental Variables

We first introduce the concept of Instrumental Variables (IVs). Under the SCM of outcome $Y \in \mathcal{Y}$, treatment $A \in \mathcal{A}$ and hidden confounder $U \in \mathcal{U}$ defined in Equation (22), an IV $Z \in \mathcal{Z}$ is an observable variable that satisfies the following conditions (Newey and Powell, 2003):

- *Unconfounded Instrument*: $Z \perp\!\!\!\perp U$;
- *Relevance*: $\mathbb{P}(A|Z)$ is not constant in Z;
- *Exclusion*: Z does not directly affect Y: $Z \perp\!\!\!\perp Y \mid (A, U)$,

where a causal graph with an instrumental variable Z is depicted in Figure 7.

Furthermore, in order to identify the causal effect $\mathbb{E}[Y \mid do(A)]$, an additional assumption of *additive noise* is required, where we assume that

$$Y = f(A) + \epsilon(U) \quad \text{with} \quad \mathbb{E}[\epsilon(U)] = 0. \tag{23}$$

Specifically, since the hidden confounder U affects both A and Y, it is generally the case that $\mathbb{E}[\varepsilon(U) \mid A] \neq 0$, which makes standard regression methods such as ordinary least squares

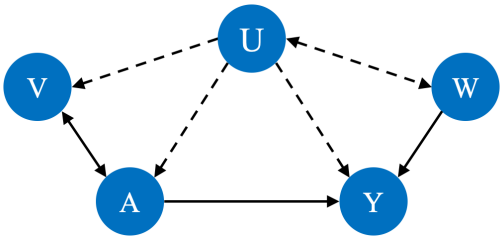

Figure 8: The causal graph of outcome Y, treatment A, hidden confounder U and proxies V and W

fail to estimate the correct causal effect. The additive noise assumption in conjunction with the IV conditions is standard for the IV settings (Newey and Powell, 2003; Xu et al., 2020; Shao et al., 2024) and allows the minimal condition to identify the causal effect. Another observation we can make here is that, following Equation (23),

$$\mathbb{E}[Y \mid do(A)] = \mathbb{E}[f(A) \mid do(A)] + \mathbb{E}[\epsilon(U) \mid do(A)] \tag{24}$$
$$= f(A) + \mathbb{E}[\epsilon(U)] = f(A), \tag{25}$$

so the task of identifying the *causal effect* $\mathbb{E}[Y \mid do(A)]$ is the same as learning the *causal function* $f(A)$.

In order to identify $f(A)$, a key observation (Newey and Powell, 2003) is that, by taking the expectation on both sides of Equation (23) conditional on Z, we have

$$\begin{aligned}\mathbb{E}[Y|Z] &= \mathbb{E}\big[f(A) + \epsilon(U)|Z\big] \\ &= \mathbb{E}[f(A)|Z] + \mathbb{E}[\epsilon(U)] \\ &= \mathbb{E}[f(A)|Z] = \int f(A)\mathbb{P}(A|Z)dA,\end{aligned} \tag{26}$$

where the expectation $\mathbb{E}[Y|Z]$ and the distribution $\mathbb{P}(A|Z)$ are both observable. Therefore, the problem of estimating the causal effect $\mathbb{E}[Y \mid do(A)]$ in the IV setting can be reduced to the CMR:

$$\mathbb{E}[Y - f(A)|Z] = 0. \tag{27}$$

D.4 Proximal Causal Learning

Next, we introduce proximal causal learning (PCL). Under the SCM of outcome $Y \in \mathcal{Y}$, treatment $A \in \mathcal{A}$ and hidden confounder $U \in \mathcal{U}$ defined in Equation (22), PCL uses two proxy variables to identify the causal effect of the treatment A on the outcome Y, i.e., $\mathbb{E}[Y \mid do(A)]$. The first proxy $V \in \mathcal{V}$ is a treatment-inducing proxy, and the second proxy $W \in \mathcal{W}$ is an outcome-inducing proxy. For V and W to be valid proxies, they need to satisfy the following *conditional independence conditions*:

- $Y \perp\!\!\!\perp V \mid (A, U)$;
- $W \perp\!\!\!\perp (A, V) \mid U$,

where a causal graph with proxies V and W is depicted in Figure 8.

In addition, for identifiability of the causal effect, proxies should satisfy *completeness assumptions*. Let $l : \mathcal{U} \to \mathbb{R}$ be any square integrable function, that is, $\|l\|_2 \leq \infty$. The following conditions hold for any $a \in \mathcal{A}$:

$$\mathbb{E}[l(U)|A = a, W = w] = 0 \quad \forall w \in \mathcal{W} \iff l(u) = 0 \text{ a.e. on } \mathbb{P}(U) \tag{28}$$

$$\mathbb{E}[l(U)|A = a, V = v] = 0 \quad \forall v \in \mathcal{V} \iff l(u) = 0 \text{ a.e. on } \mathbb{P}(U) \tag{29}$$

It has been shown that when the conditional independence conditions and the completeness assumptions are satisfied, it is possible to identify $\mathbb{E}[Y|do(A)]$ by solving CMRs.

Proposition 16 (Miao et al. (2018)) *Let the conditional independence and completeness assumptions hold, then there exists at least one solution to the following CMR*

$$\mathbb{E}[Y|A, V] = \mathbb{E}[h(A, W)|A, V] \tag{30}$$

$$= \int h(A, W)\mathbb{P}(A, W|A, V)dW, \tag{31}$$

for all $(A, V) \in \mathcal{A} \times \mathcal{V}$. *Let* h^* *be a solution of Equation* (30)*, then the causal effect* $\mathbb{E}[Y \mid do(A)]$ *can be estimated by* $\mathbb{E}_W[h(A, W)]$.

From this proposition, we can see that the problem of estimating the causal effect $\mathbb{E}[Y \mid do(A)]$ can be reduced to estimating h^*, which we denote as the bridge function following Miao et al. (2018). Therefore, estimating the causal effect in the PCL setting can be reduced to the CMR:

$$\mathbb{E}[Y - h(A, W)|A, V] = 0 \tag{32}$$

Remark 17 *For both IV and PCL, it is possible to include additional observed confounders* X *that affect both the treatment and the outcome as additional information or context.* X *can also be confounded by* U*, and the resulting CMRs would be* $\mathbb{E}[Y - f(A, X)|Z, X] = 0$ *for IV regression and* $\mathbb{E}[Y - h(A, W, X)|A, V, X] = 0$ *for PCL.*

Appendix E. Datasets Details

In this section, we provide details of the datasets considered in this paper for IV regression and proximal causal learning tasks.

E.1 IV Regression

We first provide the details for IV regression benchmarking datasets. Recall that we denote A as the action, Y as the outcome, Z as the instrument, and X as additional observed context and the CMR we are trying to solve is $\mathbb{E}[Y - f(A, X)|Z, X] = 0$.

E.1.1 Ticket Demand Dataset

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

$$r = f_0((t,s),p) + \epsilon, \quad \mathbb{E}[\epsilon|t,s,p] = 0;$$
$$p = 25 + (z+3)\psi(t) + \omega,$$

where r is the ticket sales (as the outcome variable Y) and p is the ticket price (as the action variable A). (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 $\rho = 0.9$. f_0 is the true causal effect function, defined as

$$f_0((t,s),p) = 100 + (10+p)\cdot s \cdot \psi(t) - 2p,$$
$$\psi(t) = 2\left(\frac{(t-5)^4}{600} + \exp(-4(t-5)^2) + \frac{t}{10} - 2\right),$$

where $\psi(t)$ is a complex non-linear function of t plotted in Figure 9. The offline dataset is sampled with the following distributions:

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

From the observations (r,p,t,s,z), we estimate $\widehat{h}$ using IV regression methods, and the mean squared error between $\widehat{h}$ and the true causal function f_0 is 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.

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.

E.1.2 Ticket Demand 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 and 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.

E.1.3 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

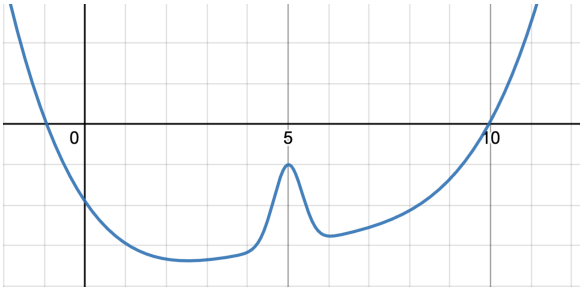

Figure 9: A graph of the nonlinear function $\psi(t)$ in the ticket demand dataset for IV regression.

IHDP[6] (Hill, 2011) and PM-CMR[7] (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 X.

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 X.

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

$$
\begin{aligned}
&Z \sim \mathbb{P}(Z = z) = 1/K, \quad z \in [1..K];\\
&A = \sum_{z=1}^{K} 1_{Z=z} \sum_{i=1}^{d_X} w_{iz}(X_i + 0.2\epsilon + f_z(z)) + \delta_A, \quad w_{iz} \sim \mathrm{Unif}(-1, 1);\\
&Y = 9A^2 - 1.5A + \sum_{i=1}^{d_X} \frac{X_i}{d_X} + |X_1 X_2| - \sin\left(10 + X_2 X_3\right) + 2\epsilon + \delta_Y,
\end{aligned}
$$

where X_i denotes the i-th variable in X, f_z is a function that returns different constants depending on the input z, $\delta_Y, \delta_A \sim \mathcal{N}(0, 1)$ and $\epsilon \sim \mathcal{N}(0, 0.1)$ are the unobserved confounders. 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).

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

E.2 Proximal Causal Learning

Next, we provide details for benchmarking datasets of proximal causal learning. Recall that we denote A as the treatment, Y as the outcome, W as the outcome proxy, and V as the treatment proxy. The CMR problem we are solving for PCL is $\mathbb{E}[Y - f(A, W)|A, V]$.

E.2.1 Ticket Demand Dataset

The ticket demand dataset (Hartford et al., 2017) is also extended to the PCL setting, which is first introduced by Xu et al. (2021). The data generating process is described by the following model:

$$U \sim \text{Unif}(0, 10) \tag{33}$$
$$[V_1, V_2] = [2\sin(2\pi U/10) + \epsilon_1, 2\cos(2\pi U/10) + \epsilon_2] \tag{34}$$
$$W = 7g(U) + 45 + \epsilon_3 \tag{35}$$
$$A = 35 + (V_1 + 3)g(U) + V_2 + \epsilon_4 \tag{36}$$
$$Y = A \cdot \min(\exp(\frac{W - A}{10}, 5) - 5g(U) + \epsilon_5 \tag{37}$$
$$\text{with} \quad g(u) = 2(\frac{(u-5)^4}{600} + \exp(-4(u-5)^2) + u/10 - 2) \tag{38}$$
$$\text{and} \quad \epsilon_i \sim \mathcal{N}(0, 1), \tag{39}$$

where U is the demand, which acts as the hidden confounder, V_1, V_2 are fuel prices which act as treatment proxy, W is the web page views which act as outcome proxy, A is the price and Y is the sale. Here, we can see that the outcome proxy W and the treatment proxy V are both affected by U, where W directly affects the outcome and V directly affects the treatment A.

E.2.2 dSprites high-dimensional Dataset

The dSprites dataset (Matthey et al., 2017) is a high-dimensional (64×64) image dataset described by five latent parameters: *shape*, *scale*, *rotation*, *posX* and *poxY*. It is proposed by Xu et al. (2021) to adopt it as a benchmark for PCL where the treatment is each figure and the hidden confounder is *posY*. For the experiments, we fix the *shape* to be heart.

The data-generating process can be described by the following steps:

1. Randomly generate values for *scale, rotation, posX and posY*: *scale* $\sim$ Unif$\{0.5, 0.6, ..., 1.0\}$, *rotation* $\sim$ Unif$(0, 2\pi)$, *posX, posY* $\sim$ Unif$\{0, ..., 31\}$.

2. Set U=*posY*

3. Set V=*(scale,rotation,posX)*

4. Set A as the dSprites image with features *(scale, rotation, posX, posY)* and add Gaussian noise $\mathcal{N}(0, 0.1)$ to each pixel.

5. Set W as posY with Gaussian noise $\mathcal{N}(0, 1)$.

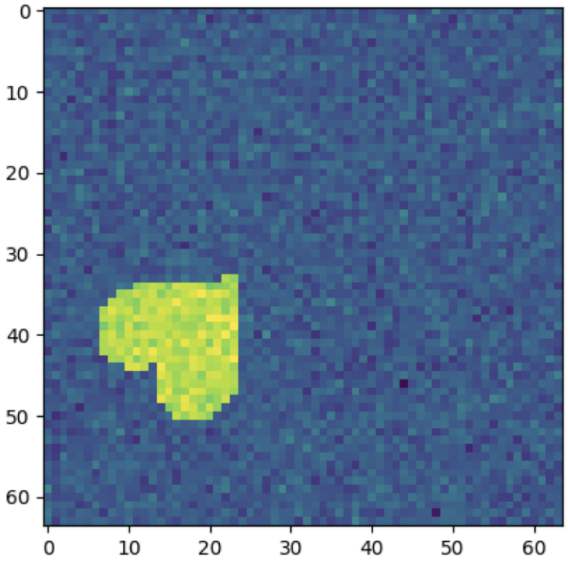

Figure 10: An example of dSprites image, which is used as the treatment A in PCL experiments. Its scale, rotation, x position and y position are randomly generated.

6. $Y = \frac{0.1\|vec(A)^T B\|_2^2 - 5000}{1000} \times \frac{(31 \times U - 15.5)^2}{85.25} + \epsilon, \epsilon \sim \mathcal{N}(0, 0.5)$, where the matrix $B \in \mathbb{R}^{64 \times 64}$ is given by $B_{i,j} = |32 - j|/32$.

For the test dataset, a fixed grid of image parameters is chosen:

$$posX \in [0, 5, 10, 15, 20, 25, 30] \tag{40}$$
$$posY \in [0, 5, 10, 15, 20, 25, 30] \tag{41}$$
$$scale \in [0.5, 0.8, 1.0] \tag{42}$$
$$rotation \in [0, 0.5\pi, \pi, 1.5\pi], \tag{43}$$

which consists of 588 images to reliably evaluate different PCL algorithms.

Appendix F. Network Architecture and Hyperparameters

Here, we describe the network architecture and hyperparameters of all experiments. Unless otherwise specified, all neural network algorithms are optimised using AdamW (Loshchilov and 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-CMR. In addition, all hyperparameter choices for methods and datasets used in this work are available in our code.

F.1 IV Regression

We first introduce details for the IV regression experiments.

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

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

Layer Type	Configuration
Input	C
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

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

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

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

Table 1: Network architecture for DML-CMR and CE-DML-CMR for the ticket demand low-dimensional dataset for IV regression. 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.

F.1.1 Ticket Demand Dataset

For DML-CMR and CE-DML-CMR, 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-CMR with dropout $= 0.1$ and the same learning rate as DML-CMR. 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 architecture as action network and stage 2 network for DML-CMR, 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.

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 2: Network architecture of the feature extractor used for the ticket demand dataset with MNIST for IV regression. 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*.

F.1.2 Ticket Demand with MNIST

For DML-CMR and CE-DML-CMR, 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-CMR, 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-CMR, 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.1.3 IHDP and PM-CMR

For the two real-world datasets, we use the same network architecture 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-CMR with dropout $= 0.1$. For DFIV, we also use the same network architecture 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 architecture as the action network and stage 2 network of DML-CMR, 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.

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

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

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

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

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

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

Table 3: Network architecture for DML-CMR and CE-DML-CMR for the ticket demand dataset with MNIST for IV regression. 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.

F.2 Proximal Causal Learning

Next, we introduce details for the proximal causal learning experiments.

F.2.1 Ticket Demand Dataset

For DML-CMR and CE-DML-CMR, we use the network architecture described in Table 4. We use a learning rate of 0.0001 with a weight decay of 0.001 (L2 regularisation) for the f network and 0.0001 for the s and g network. The dropout rate is $\frac{400}{4000+N}$, which depends on the data size N. For the comparison methods, we use the default parameter values proposed in their original papers. For CEVAE, we use 1000 epochs, 0.0001 weight decay, 10 learning samples, 20 hidden dimensions, and 5 early stopping. For DFPV, $\lambda 1$, $\lambda 2$ both set to 0.1, weight decay is 0.01 for all networks, stage 1 iteration is 20, and stage 2 iteration is 1. For KPV, we set $\lambda 1$ and $\lambda 2$ to be 0.001 with data split ratio 0.5. For NMMR, learning rate is 0.003, L2 penalty is 0.000003 with network depth 4 and width 80 for the U statistics estimator. For the V statistics estimator, network depth is 3 and width is 80 while other hyperparameters remain the same. For PKDR, the number of components is 50, gamma is 50, alpha is 35, and cross validation is 5. For PMMR, $\lambda 1$ and $\lambda 2$ are 0.01, with scale 0.5.

F.2.2 dSprites dataset

For the dSprites dataset, we adopt a CNN feature extractor to handle the image inputs for DML-CMR. The architecture of this feature extractor is provided in Table 5. We use 10 components for the mixture of Gaussian model, dropout is 0.2, batch size is 100, weight decay is 0.05, learning rate is 0.001 with Adam, and the number of epochs $int(1000000./N) + 100$ depends on the sample size N. For CEVAE, DFPV, KPV, NMMR and PMMR, we follow

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

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

Layer Type	Configuration
Input	C
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

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

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

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

Table 4: Network architecture for DML-CMR and CE-DML-CMR for the ticket demand dataset for PCL. 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.

the hyperparameters and network architecture used in Kompa et al. (2022) to generate the experimental results. For PKDR, we follow the high-dimensional dataset hyperparameters used in the original paper (Wu et al., 2024) with weight decay 0.0001, 4 layers, learning rate 0.0001, and 500 epochs.

F.3 Validation and Hyper-Parameter Tuning

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

Layer Type	Configuration
Input	64×64
Conv + ReLU	$5 \times 5 \times 64$, s:1, p:0
Max Pooling	2×2, s:2
Dropout	-
Conv + ReLU	$5 \times 5 \times 128$, s:1, p:0
Max Pooling	2×2, s:2
Dropout	-
Conv + ReLU	$5 \times 5 \times 128$, s:1, p:0
Dropout	-
Max Pooling	2×2, s:2
FC + ReLU	in: 2048, out:128

Table 5: Network architecture of the feature extractor used for the dSprites image dataset for PCL. 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 0.2. We denote this feature extractor as *ImageFeature*.

Appendix G. Additional Experimental Results

In this section, we provide additional experimental results including the effects of high ill-posedness (e.g., 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 inducing bias in the second stage estimator, especially for non-linear IV regressions. In theory, DML-CMR should be more resistant to biases in the first stage thanks to the DML framework, as long as the causal effect is identifiable under the weak instrument. This identifiability condition is captured in Condition 4 for DML, and is connected to the ill-posedness for CMR problems in general as discussed in Section 4.4. With identifiability, Theorem 6 and Corollary 7 all hold, and the convergence rate guarantees still apply. Intuitively, as the ill-posedness increases, worse empirical performance will be observed.

Experimentally, for the ticket demand dataset, we alter the instrument strength by changing how much the instrument z affects the price p. Recall from Appendix E.1.1 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 6, we present the mean and standard deviation of the MSE of $\widehat{h}$ for various IV strengths ϱ from 0.01 to 1 and sample size $N = 5000$. It is very interesting to see that DML-CMR indeed performs significantly better than SOTA nonlinear IV regression methods under weak instruments.

IV Strength	1.0	0.8	0.6	0.4	0.2	0.01
DML-CMR	**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-CMR	**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 6: Results for the low-dimensional ticket demand dataset when the IV is weakly correlated with the action, plotted against IV strength. The results from this paper are shown in boldface.

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

The DML-CMR 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-CMR. 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-CMR, CE-DML-CMR, 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 $\widehat{h}$.

In Table 7, we present the mean and standard deviation of the MSE of $\widehat{h}$ with Random Forests and Gradient Boosting estimators on the 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.

IV Strength	Dataset Size	DNNs	Random Forests	Gradient Boosting
DML-CMR	2000	**0.1308(0.0206)**	0.1689(0.0172)	**0.1301(0.0112)**
CE-DML-CMR	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-CMR	5000	**0.0676(0.0129)**	0.1067(0.0131)	**0.0632(0.0107)**
CE-DML-CMR	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-CMR	10000	**0.0378(0.0094)**	0.0657(0.0062)	**0.0482(0.0079)**
CE-DML-CMR	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 7: Results for the low-dimensional ticket demand dataset comparing the use of tree-based estimators with DNN estimators. The results from this paper are shown in boldface.

Learning Rate	Weight Decay	Dropout	DNN Width	DML-CMR	CE-DML-CMR
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 8: 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. The bold results are the best performing hyperparameters.

G.3 Sensitivity analysis for different Hyperparameters

The tunable hyperparameters in DML-CMR 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-CMR estimator $\widehat{h}$ with different hyperparameter values for both the low-dimensional and high-dimensional datasets with sample size N=5000 in Table 8 and Table 9. Overall, we see that DML-CMR is not very sensitive to small changes in the hyperparameters.

Learning Rate	Weight Decay	Dropout	CNN Channels	DML-CMR	CE-DML-CMR
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 9: 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. The bold results are the best performing hyperparameters.

References

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

I. Andrews, J. H. Stock, and L. Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753, 8 2019.

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

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

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

A. Bennett and N. Kallus. The variational method of moments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85:810–841, 12 2020.

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

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

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732, 8 2009. ISSN 0090-5364. doi: 10.1214/08-AOS620.

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

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric IV estimation of shape-invariant engel curves. *Econometrica*, 75:1613–1669, 11 2007.

J. Bound, D. A. Jaeger, and R. M. Baker. 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.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

M. Carrasco, J. P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751, 1 2007.

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

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80:277–321, 1 2012.

X. Chen and H. White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

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

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

V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 4 2021.

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

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

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

A. Christmann and I. Steinwart. *Support vector machines*. Springer, 2008.

Y. Cui, H. Pu, X. Shi, W. Miao, and E. T. Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119:1348–1359, 11 2023.

S. Darolles, Y. Fan, J. P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79:1541–1565, 9 2011.

N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, volume 2020-December. Advances in Neural Information Processing Systems, 6 2020.

M. A. Domínguez and I. N. Lobato. Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72:1601–1615, 9 2004.

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

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *Annals of Statistics*, 51(3): 879–908, 1 2019.

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

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029, 7 1982.

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

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240, 3 2011.

N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence rates for gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022.

J. Honerkamp and J. Weese. Tikhonovs regularization method for ill-posed problems - a comparison of different methods for the determination of the regularization parameter. *Continuum Mechanics and Thermodynamics*, 2:17–30, 3 1990.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13:29–61, 1 2022.

D. J. Im, K. Cho, and N. Razavian. Causal effect variational autoencoder with uniform treatment. *arXiv preprint arXiv:2111.08656*, 11 2021.

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

B. Kompa, D. R. Bellamy, T. Kolokotrones, J. M. Robins, and A. L. Beam. Deep learning methods for proximal inference via maximum moment restriction. *Advances in Neural Information Processing Systems*, 35, 5 2022.

H. Kremer and B. Schölkopf. Geometry-aware instrumental variable regression. *Proceedings of the International Conference on Machine Learning*, 5 2024.

R. Kress. *Linear integral equations*, volume 82. Springer, 1999.

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

L. Liao, Y. L. Chen, Z. Yang, B. Dai, Z. Wang, and M. Kolar. Provably efficient neural estimation of structural equation model: An adversarial approach. *Advances in Neural Information Processing Systems*, 2020-December, 7 2020.

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

Y. Luo, M. Spindler, and J. Kück. High-dimensional l_2 boosting: Rate of convergence. *arXiv preprint arXiv:1602.08927*, 2016.

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

A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. J. Kusner, A. Gretton, and K. Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dSprites: disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

W. Miao, Z. Geng, and E. J. T. Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105:987–993, 9 2018.

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

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

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 9 2003.

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

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

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

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56:931, 7 1988.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875, 2020.

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

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

D. Shao, A. Soleymani, F. Quinzan, and M. Kwiatkowska. Learning decision policies with instrumental variables through double machine learning. *Proceedings of the International Conference on Machine Learning*, 2024.

X. Shi, W. Miao, J. C. Nelson, and E. J. T. Tchetgen. Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82:521–540, 8 2020.

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

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

V. Syrgkanis and M. Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on learning theory*, pages 3453–3454. PMLR, 2020.

E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An introduction to proximal causal learning. *Statistical Science*, 39:375–390, 8 2024.

A. W. Van Der Vaart, J. A. Wellner, A. W. van der Vaart, and J. A. Wellner. *Weak convergence*. Springer, 1996.

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

V. N. Vapnik. *Statistical learning theory*. Wiley, 1998.

M. J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *Cambridge University Press*, pages 1–552, 1 2019.

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

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

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

Y. Wu, Y. Fu, S. Wang, and X. Sun. Doubly robust proximal causal learning for continuous treatments. *Proceedings of the International Conference on Learning Representations*, 9 2024.

L. H. Wyatt, G. C. L. Peterson, T. J. Wade, L. M. Neas, and A. G. Rappold. 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.

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

L. Xu, H. Kanagawa, and A. Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 31:26264–26275, 6 2021.

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

D. Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

Q. Zhao, P. Sur, and E. J. Candes. The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.