Deep Learning Methods for Proximal Inference via Maximum Moment Restriction

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

The No Unmeasured Confounding Assumption is widely used to identify causal effects in observational studies. Recent work on proximal inference has provided alternative identification results that succeed even in the presence of unobserved confounders, provided that one has measured a sufficiently rich set of *proxy variables*, satisfying specific structural conditions. However, proximal inference requires solving an ill-posed integral equation. Previous approaches have used a variety of machine learning techniques to estimate a solution to this integral equation, commonly referred to as the *bridge function*. However, prior work has often been limited by relying on pre-specified kernel functions, which are not data adaptive and struggle to scale to large datasets. In this work, we introduce a flexible and scalable method based on a deep neural network to estimate causal effects in the presence of unmeasured confounding using proximal inference. Our method achieves state of the art performance on two well-established proximal inference benchmarks. Finally, we provide theoretical consistency guarantees for our method.

1 Introduction

Causal inference is concerned with estimating the effect of a treatment A on an outcome Y from either observational data or the results of a randomized experiment. An estimand of primary importance is the *average causal effect* (ACE), which is the expected difference in Y caused by changing the treatment from value a to a' for each unit in the study population, and is defined as a contrast between the expected value of the potential outcomes at the two levels of the treatment: $\mathbb{E}[Y^{a'}] - \mathbb{E}[Y^a]$. However, in observational settings, the ACE is rarely equal to the observed difference in conditional means, $\mathbb{E}[Y|A = a'] - \mathbb{E}[Y|A = a]$ due to confounding. In an attempt to eliminate the influence of confounding, investigators measure putative confounders X and subsequently make adjustments for X in their analyses.

Given X, common approaches, such as standardization and inverse probability weighting (Hernán and Robins [1]), obtain valid estimates of the ACE given that the following assumptions hold: i) Positivity: $Pr[A = a|X = x] > 0$ for all x in the population, ii) Consistency: $Y^a = Y$ for all individuals with $A = a$, iii) No unmeasured confounding which results in conditional exchangeability: $Y^a \perp\!\!\!\perp A|X.$, and iv) No model misspecification.

While positivity can be empirically verified, the remaining 3 assumptions are not. While model mispecification is likely in all real-world scenarios, flexible models and doubly robust estimators have been developed to mitigate the effect of this assumption[2]. Therefore, the assumption of conditional exchangeability, or equivalently, the No Unmeasured Confounding Assumption (NUCA), is the defining characteristic of this broad set of approaches to causal effect estimation (Hernán and Robins [3]). However, in many settings, it is unrealistic to assume that we are able to measure a sufficient set of confounders for A and Y such that conditional exchangeability holds.

Proximal inference is a recently introduced framework that allows for the identification of causal effects even in the presence of unmeasured confounders [4, 5]. Proximal inference requires categorizing the measured covariates into three mutually exclusive (but not exhaustive) groups: treatment-inducing proxy variables Z, outcome-inducing proxy variables W, and "backdoor" variables X that affect both A and Y (i.e. typical confounders). See Figure 1 for an example of a directed acyclic graph (DAG) that admits identification under the assumptions of proximal inference. The proxy sets W and Z must contain sufficient information about the remaining unobserved confounders U, a condition that can be formalized by completeness assumptions. Under these and several other conditions, one can estimate average potential outcomes from data even in the presence of unmeasured confounding. Proximal inference methods have potential applications in medical settings, where a natural question is considering the effect of a treatment on an outcome in the presence of unmeasured confounding. Before applying proximal inference to real world problems, more validation is required before they can be used safely to inform medical decision-making.

Existing methods for proximal inference can be divided into two categories: two-stage regression procedures and methods that impose a maximum moment restriction (MMR). In two-stage regression procedures, the first stage aims to predict outcome-inducing proxy variables W as a function of A, X, and Z. Then, the second stage regression estimates outcomes Y as a function of the predicted $\hat{W}$ and the treatment A, and measured confounders X. Tchetgen Tchetgen et al. [5] introduced the first estimation technique for proximal inference which was a two-stage procedure that used a model based on ordinary least squares regression. Mastouri et al. [6] extended this framework by replacing simple linear regression with kernel ridge regression. Xu et al. [7] increased feature flexibility further by incorporating neural networks as feature maps instead of kernels.

In contrast, MMR methods are single-stage procedures to estimate average potential outcomes. Muandet et al. [8] introduced MMR for reproducing kernel Hilbert spaces (RKHS). MMR critically relies on the optimization of a V-statistic or U-statistic for learning a function needed to calculate the ACE. Zhang et al. [9] used an MMR method to obtain point identification of the ACE in the instrumental variable (IV) setting and incorporated neural networks into their method by training with the V-statistic as a loss function and optimized using stochastic gradient descent. Mastouri et al. [6] demonstrated that the MMR framework with kernel functions can be used for proximal inference as well as IV regression.

In this work, we introduce a new method, *Neural Maximum Moment Restriction* (NMMR) which is a flexible neural network approach that is trained to minimize a loss function derived from either a U-statistic or V-statistic to satisfy MMR in the proximal setting. The method introduced in this work makes several novel contributions to the proximal inference literature:

- We introduce a new, single stage method based on neural networks for estimating potential outcomes and the ACE in the presence of unmeasured confounding.
- We provide new theoretical consistency guarantees for our method.
- We demonstrate state-of-the-art (SOTA) performance on two well-established proximal inference benchmark tasks.
- We show for the first time how to incorporate domain-specific inductive biases using a convolutional model on a proximal inference task that uses images.
- We provide the first unbiased estimate of the MMR risk function using the U-statistic rather than V-statistic in the proximal setting.

2 Background: Proximal Inference

Throughout the rest of this manuscript we will use capital letters (e.g. A) to denote random variables taking values in measurable spaces, (e.g. $\mathcal{A}$; typically subsets of $\mathbb{R}$). Lower case letters denote realizations of these random variables (e.g. $A = a$). Estimates of random variables will be indicated using "hat" notation, e.g. $\hat{Y}$ would be an estimate for the random variable Y.

Our goal is to estimate the average potential outcome a specified level of an treatment, i.e. $\mathbb{E}[Y^a]$ is the average potential outcome under the treatment a. Without loss of generality, we refer to A as a treatment, though it could refer to any (possibly continuous) intervention. Proximal inference allows for the estimation of mean potential outcomes for Y^a in the presence of unobserved confounders U provided that we have a sufficiently rich set of proxies Z, W along with observed confounders X that obey certain structural assumptions. The assumptions needed for identification are given below:

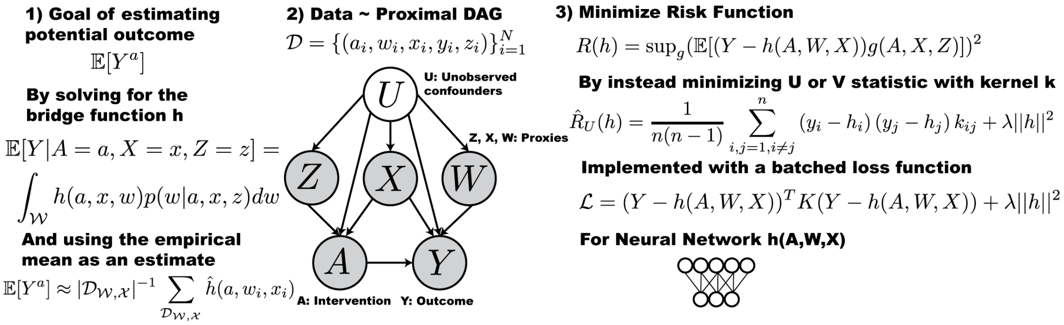
1) Goal of estimating potential outcome
$\mathbb{E}[Y^a]$
By solving for the bridge function h
$\mathbb{E}[Y|A=a,X=x,Z=z] = \int_{\mathcal{W}} h(a,x,w)p(w|a,x,z)dw$
And using the empirical mean as an estimate
$\mathbb{E}[Y^a] \approx |\mathcal{D}_{\mathcal{W},\mathcal{X}}|^{-1} \sum_{\mathcal{D}_{\mathcal{W},\mathcal{X}}} \hat{h}(a,w_i,x_i)$
2) Data ~ Proximal DAG
$\mathcal{D} = \{(a_i,w_i,x_i,y_i,z_i)\}_{i=1}^{N}$
U
U: Unobserved confounders
Z, X, W: Proxies
Z
X
W
A
Y
A: Intervention
Y: Outcome
3) Minimize Risk Function
$R(h) = \sup_g(\mathbb{E}[(Y-h(A,W,X))g(A,X,Z)])^2$
By instead minimizing U or V statistic with kernel k
$\hat{R}_U(h) = \frac{1}{n(n-1)} \sum_{i,j=1,i\neq j}^{n} (y_i-h_i)(y_j-h_j)k_{ij} + \lambda||h||^2$
Implemented with a batched loss function
$\mathcal{L} = (Y-h(A,W,X))^T K(Y-h(A,W,X)) + \lambda||h||^2$
For Neural Network h(A,W,X)

Figure 1: A summary of NMMR. Our method estimates the bridge function h, which can be used to compute the average potential outcome $\mathbb{E}[Y^a]$. We rely on structural assumptions for the causal DAG generating the data. NMMR uses a U or V statistic to train a neural network that solves a risk function that reflects a maximum moment restriction function.

Assumption 1. *A, U, W, X, Y, Z satisfy the conditional independences $Y \perp\!\!\!\perp Z | A, U, X$ and $W \perp\!\!\!\perp (A, Z) | U, X$.*

Note: Assumption 1 can be equivalently stated as the requirement that these random variables must be generated by one of 22 valid causal directed acyclic graphs (DAGs) for proximal inference [5]. Figure 1 provides an example of a DAG that satisfies these assumptions. We can also formalize the notion of "rich enough" proxies with two assumptions about the completeness of U and Z:

Assumption 2. *For any square integrable function g, $\mathbb{E}[g(U)|A = a, X = x, Z = z] = 0$, $\forall(a, x, z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$ if and only if $g(u) = 0$ almost surely.*

Assumption 3. *For any square integrable function g, $\mathbb{E}[g(Z)|A = a, W = w, X = x] = 0$, $\forall(a, w, x) \in \mathcal{A} \times \mathcal{W} \times \mathcal{X}$ if and only if $g(z) = 0$ almost surely.*

We will use two other assumptions at various points in the paper. The first guarantees the uniqueness of the bridge function, while the second ensures the risk function does not have false zeroes.

Assumption 4. *$\mathbb{E}\left[f\left(A, W, X\right)|A, X, Z\right] = 0$ $\mathrm{P}_{A,X,Z}$-almost surely if and only if $f\left(A, W, X\right) = 0$ $\mathrm{P}_{A,W,X}$-almost surely.*

Assumption 5. *$k : \left(\mathcal{A} \times \mathcal{X} \times \mathcal{Z}\right)^2 \to \mathbb{R}$ is continuous, bounded, and Integrally Strictly Positive Definite (ISPD), so that $\int f\left(\xi\right) k\left(\xi, \xi'\right) f\left(\xi'\right) d\xi d\xi' > 0$ if and only if $f \neq 0$ $\mathrm{P}_{A,Z,X}$-almost surely.*

Assumptions 1-3 together with several regularity assumptions (see assumptions (v)-(vii) in [4]) ensure that there exists a function h such that:

$$\mathbb{E}[Y|A = a, X = x, Z = z] = \int_{\mathcal{W}} h(a, w, x)p(w|a, x, z)dw \tag{1}$$

Equation 1 is a Fredholm integral equation of the first kind and in general is difficult to solve for h. The function h is often referred to as the "bridge function" and Theorem 1 of Miao et al. [4] also showed that if one has a solution h, then average potential outcomes can be computed as:

$$\mathbb{E}[Y^a] = \int_{\mathcal{W},\mathcal{X}} h(a, w, x)p(w, x)dwdx = \mathbb{E}_{W,X}\left[h(a, W, X)\right] \tag{2}$$

Thus, under the proximal inference framework an unbiased estimate of the average potential outcome $\mathbb{E}[Y^a]$ is obtained by first estimating the bridge function h, by some $\hat{h}$, and then fixing a and taking the empirical average of $\hat{h}$ over the remaining variables using a held-out dataset with M data points $D_{\mathcal{W},\mathcal{X}} = \{(w_i, x_i)\}_{i=1}^{M}$, $\hat{\mathbb{E}[Y^a]} = \frac{1}{M}\sum_{i=1}^{M} \hat{h}(a, w_i, x_i)$. Once these mean potential outcomes have been estimated, then contrasts of interest like the ACE can be calculated in a straight-forward manner.

3 Related Work

Kuroki and Pearl [10] first established identification of a causal effect in the setting of unobserved confounders by leveraging noisy proxy variables and estimating the distribution of U using external datasets that allowed for estimation of a distribution $p(w|u)$. These results were extended to allow for identification without recovery of U in Miao et al. [4] and Tchetgen Tchetgen et al. [5]. In particular, Tchetgen Tchetgen et al. [5] set out the 22 valid causal DAGs that satisfy Assumption 1 and allow for identification provided one has access to a sufficiently rich set of proxy variables. These authors also provide a 2-Stage Least Squares (2SLS) method to identify and estimate causal effects under the assumption that the bridge function has a linear form. Ghassami et al. [11] established a doubly robust method for proximal inference using influence functions.

Earlier work applying machine learning techniques to the task of identification in the setting of proximal inference included Deaner [12], which relied on a two-stage penalized sieve distance minimization. Several later works similarly employed two-stage regressions with increasingly flexible basis functions to estimate potential outcomes. Mastouri et al. [6] developed a two-stage kernel ridge regression (Kernel Proxy Variables "KPV") to estimate the bridge function h. While a kernel basis has more flexibility than the linear basis of Tchetgen Tchetgen et al. [5], Xu et al. [7] introduced an adaptive basis derived from neural networks. Their two stage regression method, Deep Feature Proxy Variables (DFPV), established the previous SOTA performance on the proximal benchmark tasks that we consider in our work.

An alternative approach based on maximum moment restriction (MMR) uses single-stage estimators of the bridge function. MMR-based methods were established in Muandet et al. [8] as a way to enforce conditional moment restrictions [13]. Zhang et al. [9] introduced the MMR framework to the instrumental variable (IV) setting. The IV setting assumes a specific DAG where unobserved confounders can be circumvented by use of an instrument that is not confounded by U. IV DAGs can be considered a subset of proximal DAGs without outcome-inducing proxies W [5]. There are now several machine learning methods that can be applied in the IV setting [14–18]

Of note, Zhang et al. [9] introduced MMR-IV which is related to Lewis and Syrgkanis [15] and Dikkala et al. [17]. MMR-IV involves optimizing a family of risk functions based on U- or V-statistics [19]. However, Zhang et al. [9] only consider the IV setting, which significantly differs from the proximal setting because of the absence of outcome-inducing proxy variables. Additionally, Zhang et al. [9] only optimize neural networks by a loss that corresponds to the V-statistic. The V-statistic provides a biased estimate [19] of its corresponding risk function, such as $R(h)$ in Equation 3.

Finally, Mastouri et al. [6] introduced an MMR-based method for proximal inference called Proximal Maximum Moment Restriction (PMMR). PMMR extends the MMR framework to the proximal setting through the use of kernel functions and also optimizes Equation 3 via a V-statistic. For a comparison of our model to PMMR and MMR-IV, see Table 1.

4 Our Method: Neural Maximum Moment Restriction (NMMR)

In this work we propose *Neural Maximum Moment Restriction* (NMMR) as a method to estimate average potential outcomes $\mathbb{E}[Y^a]$ in the presence of unmeasured confounding. We consider the hypothesis class of h belonging to the family of deep neural networks given their well-established flexibility, ability to scale to large data sets, and for the ease with which domain-specific inductive biases can be included directly into the model (e.g. convolutions for images). Using a connection between maximum moment restrictions [8] and U- and V-statistics, we show how a single-stage neural network procedure can be used to provide estimates of the bridge function.

Following Muandet et al. [8] and Zhang et al. [9], we can rewrite the integral equation (1) as a *conditional moment restriction*: $\mathbb{E}[Y - h(A, W, X)|A, X, Z] = 0$. Then, $\mathbb{E}[(Y - h(A, W, X))g(A, X, Z)|A, X, Z] = 0$ for any measurable function g on $\mathcal{A} \times \mathcal{X} \times \mathcal{Z}$, as well. By taking the expectation over A, X, Z we obtain $\mathbb{E}[(Y - h(A, W, X))g(A, X, Z)] = 0$ eliminating the need to take conditional expectations entirely. The intuition for this result is that any solution for h in the original conditional moment restriction will cause $\mathbb{E}[Y - h(A, W, X)|A, Z, X]$ to be 0 and, thus, the product with $g(A, Z, W)$ will also be 0, as well as the unconditional expectation, regardless of which function g one chooses. This creates an infinite number of moment restrictions and serves as the basis of the MMR framework of Muandet et al. [8]. This suggests a minimax strategy of searching for an estimate of h such that risk $R(h)$ for the worst-case value of g is minimized:

$$R(h) = \sup_{\|g\| \leq 1} (\mathbb{E}[(Y - h(A, W, X))g(A, X, Z)])^2 \tag{3}$$

Following Zhang et al. [9]'s work in the IV setting, Mastouri et al. [6] (Lemma 2) showed that, if g is an element of an RKHS, $R(h)$ can be rewritten in the form $R_k(h) = \mathbb{E}[(Y - h(A, W, X))(Y' - h(A', W', X'))k((A, X, Z), (A', X', Z'))]$ where (A', W', X', Y', Z') are independent copies of the random variables (A, W, X, Y, Z) and $k : (\mathcal{A} \times \mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}$ is a continuous, bounded, and Integrally Strictly Positive Definite (ISPD) kernel. Then, if h satisfies $R_k(h) = 0$, $\mathbb{E}[Y - h(A, W, X)|A, X, Z] = 0$ $\mathrm{P}_{A,X,Z}$-almost surely. Thus, if we can find a neural network h that satisfies $R_k(h) = 0$, we will have obtained a $\mathrm{P}_{A,X,Z}$-almost sure solution to Equation 1 and can compute any desired average potential outcome with Equation 2.

We can minimize $R_k(h)$ for data $\mathcal{D} = \{(a_i, w_i, x_i, y_i, z_i)\}_{i=1}^N$ by minimizing the expectation in Mastouri et al. [6] (Lemma 2), which can be approximated by either a V-statistic [19]

$$\hat{R}_V(h) = \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - h_i)(y_j - h_j) k_{ij}$$

or with a U-statistic [19]:

$$\hat{R}_U(h) = \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j}^{n} (y_i - h_i)(y_j - h_j) k_{ij}$$

where $h_i = h(a_i, w_i, x_i)$ and $k_{ij} = k((a_i, z_i, x_i), (a_j, z_j, x_j))$. $\hat{R}_U(h)$ is the minimum variance unbiased estimator of $R_k(h)$ [19], while $\hat{R}_V(h)$ is a biased estimator of $R_k(h)$. In practice, a penalized version of $\hat{R}_U(h)$ has a slight amount of bias, but in practice is much less biased than even an unpenalized $\hat{R}_V(h)$. Previous work either did not consider the U-statistic [6], or did not utilize the U-statistic [9]. In our work, we introduce two variants of our method, NMMR-U and NMMR-V, where the former is optimized with a U-statistic and the latter a V-statistic. We train the neural networks in both variants with a regularized loss function:

$$\mathcal{L} = (Y - h(A, W, X))^T K (Y - h(A, W, X)) + \lambda ||h||^2$$

where $(Y - h(A, W, X))$ is a vector of residuals from the neural network's predictions and K is a kernel matrix with entries k_{ij}. Throughout, we choose k to be the RBF kernel (see Appendix B). If $\mathcal{L}$ is representing a V-statistic, then we include main diagonal elements of K. If $\mathcal{L}$ is representing a U-statistic, then we set the main diagonal to be 0.

Once we've obtained an optimal neural network $\hat{h}$, we can compute an estimate of the average potential outcome with data from a held-out dataset with M data points $D_{\mathcal{W},\mathcal{X}} = \{(w_i, x_i)\}_{i=1}^M$, $\mathbb{E}[\hat{Y}^a] = \frac{1}{M} \sum_{i=1}^{M} \hat{h}(a, w_i, x_i)$

In contrast to PMMR [6], which uses kernels as feature maps for proxy and treatment variables, NMMR uses adaptive feature maps from neural networks. NMMR is similar to MMR-IV [9], but MMR-IV is restricted to the instrumental variable (IV) setting rather than the proximal inference setting. Table 1 places NMMR in context with existing methods for proximal inference and IV regression.

Table 1: Comparison of the most related methods to NMMR.

Method	Setting	# of Stages	Hypothesis Class	Optimization Objective
KPV [6]	Proximal	2	Kernels	2-stage least squares
DFPV [7]	Proximal	2	Neural Networks	2-stage least squares
MMR-IV [9]	IV	1	Neural Networks	V-statistic
PMMR [6]	Proximal	1	Kernels	V-statistic
NMMR-V (ours)	Proximal	1	Neural Networks	V-statistic
NMMR-U (ours)	Proximal	1	Neural Networks	U-statistic

5 Consistency of NMMR

In this section we provide a probabilistic bound on the distance of the estimated bridge function, $\hat{h}_{k,\lambda,n}$, from the true bridge function, h^*, in terms of the Radamacher complexity $\mathcal{R}_n(\mathcal{F})$ of a class of functions $\mathcal{F}$ derived from elements of the hypothesis space $\mathcal{H}$ and the fixed kernel, k. Note that $\hat{R}_{k,\lambda,n}(h) = \hat{R}_U(h) + \lambda \|h\|_2^2$ (see Section 4). We use this bound to demonstrate that, under mild conditions, $\hat{h}_{k,\lambda,n}$ converges in probability to h^*, and that, under an additional completeness assumption, h^* is unique $\mathrm{P}_{A,W,X}$-almost surely. This provides a consistent estimate of $\mathbb{E}[Y^a]$.

Theorem 1. *Let $\tilde{h}_k$ minimize $R_k(h)$ and $\hat{h}_{k,\lambda,n}$ minimize $\hat{R}_{k,\lambda,n}(h)$ for $h \in \mathcal{H}$ and let $h^* : \mathcal{A} \times \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ satisfy $\mathbb{E}\left[Y - h^*(A, W, X) | A, X, Z\right] = 0$ $\mathrm{P}_{A,X,Z}$-almost surely, where*

$$R_k(h) = \mathbb{E}\left[\left(Y - h\left(A, W, X\right)\right)\left(Y' - h\left(A', W', X'\right)\right) k\left(\left(A, X, Z\right), \left(A', X', Z'\right)\right)\right]$$

$$\hat{R}_{k,\lambda,n}(h) = \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} \left[\left(y_i - h\left(a_i, w_i, x_i\right)\right)\left(y_j - h\left(a_j, w_j, x_j\right)\right) \right.$$
$$\left. \times\ k\left(\left(a_i, x_i, z_i\right), \left(a_j, x_j, z_j\right)\right)\right] + \lambda \|h\|_2^2$$

Also let,

$$d_k^2\left(h, h'\right) = \mathbb{E}\left[\left(h\left(A, W, X\right) - h'\left(A, W, X\right)\right)\left(h\left(A', W', X'\right) - h'\left(A', W', X'\right)\right) \right.$$
$$\left. \times\ k\left(\left(A, X, Z\right), \left(A', X', Z'\right)\right)\right]$$

Then, $d_k^2\left(h^*, h\right) = R_k(h)$ and, with probability at least $1 - \delta$,

$$\begin{aligned} d_k^2\left(h^*, \hat{h}_{k,\lambda,n}\right) &\leq d_k^2\left(h^*, \tilde{h}_k\right) + \lambda M^2 + 8M\mathbb{E}_{A,X,Z}\left(\mathcal{R}_{n-1}\left(\mathcal{F}'_{A,X,Z}\right) + \mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)\right) \\ &\quad + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} + 10\left(2\log 2\right)^{\frac{1}{2}} M^2M_kn^{-\frac{1}{2}} \\ &\leq d_k^2\left(h^*, \tilde{h}_k\right) + \lambda M^2 + 8M\left(\mathcal{R}_{n-1}\left(\mathcal{F}'\right) + \mathcal{R}_n\left(\mathcal{F}'\right)\right) \\ &\quad + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} + 10\left(2\log 2\right)^{\frac{1}{2}} M^2M_kn^{-\frac{1}{2}} \end{aligned}$$

$\mathcal{F}'_{a,x,z} = \{f_{a,x,z} \mid \exists_{h\in\mathcal{H}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}} f_{a,x,z}\left(a', w', x', z'\right) = h\left(a', w', x'\right) k\left(\left(a', x', z'\right), \left(a, x, z\right)\right)\}$,
$\mathcal{F}' = \{f \mid \exists_{h\in\mathcal{H},a\in\mathcal{A},x\in\mathcal{X},z\in\mathcal{Z}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}} f\left(a', w', x', z'\right) = h\left(a', w', x'\right) k\left(\left(a', x', z'\right), \left(a, x, z\right)\right)\}$.
Further, if Assumption 5 holds, so k is ISPD, then d_k is a metric and, if the right hand side goes to zero as n goes to infinity,
$\left\|\mathbb{E}\left[h^*|A, X, Z\right] - \mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right]\right\|_{\mathrm{P}_{A,X,Z}} \xrightarrow{\mathrm{P}} 0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A, X, Z\right]$.

Corollary 8 provides a similar result for V-statistic estimators of $R(h)$, meaning we can choose to use either U or V-Statistics and have similar guarantees.

Theorem 2. *Under Assumption 4, h^* is the unique solution to the integral equation $\mathrm{P}_{A,W,X}$-almost surely. Further, if $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A, X, Z\right]$, $\hat{h}_n \xrightarrow{\mathrm{P}} h^*$.*

See Appendix A for proofs of Theorems 1 and 2 and Corollary 8. Taken together, these results tell us that, as long as our optimization algorithm is successful in estimating $\hat{h}_{k,\lambda,n}$, it will asymptotically approach the true bridge function, h^*. In order for this to occur, the right hand side of the inequalities in Theorem 1 must go to zero, which requires not only that the Rademacher terms vanish, but also that $\tilde{h}_k$ must approach h^* arbitrarily closely as n increases. In practice, this means increasing the complexity of the neural network, but doing so slowly enough the Rademacher complexity terms still decrease with sample size. Following Xu et al. [7], we note that recent results from Neyshabur et al. [20] suggest that the Rademacher complexity of a fixed network scales like $n^{-\frac{1}{2}}$ (similar to many other popular hypothesis classes) and that, although we cannot compute the scaling of the Rademacher terms directly due to the presence of the kernel function, we expect that they will decline with sample size and that, as the neural network becomes more complex, their scaling will more closely resemble terms derived from a pure neural network. Finally, we require that the regularization parameter decrease as sample size grows, which will, again, depend on the balance between increasing sample size, which tends to decrease the need for regularization, and increasing complexity, which tends to increase its importance. Thus, while it is difficult to determine an optimal rate of convergence, by choosing an appropriate growth rate for the network complexity, we expect the aforementioned terms to vanish as n increases to infinity, and, with them, the entire right hand side, making $\hat{h}_{k,\lambda,n}$ a consistent estimator of h^*.

6 Experiments

6.1 Overview of Baseline Models

We compare the performance of NMMR-U and NMMR-V to that of several previous approaches, which we describe briefly here. The baselines can be divided into two categories: structural and naive. The structural approaches all leverage some kind of causal information about the data generating process. The structural methods include Kernel Proxy Variables (KPV) [6], Proximal Maximum Moment Restriction (PMMR) [6], Deep Feature Proxy Variables (DFPV) [21], Causal Effect Variational Autoencoder[22] (CEVAE) and the two-stage least squares model (2SLS) from Miao et al. [4]. For a review of KPV, PMMR, and DFPV, see Section 3. CEVAE is an autoencoder approach derived by Xu et al. [7] from Louizos et al. [22]. 2SLS is a two-stage least squares method which assumes that the bridge function h is linear [5].

The naive approaches serve as baselines and do not use causal information and instead attempt to directly regress A, Z, W on the outcome Y. These methods include a naive neural network (Naive net), ordinary least squares regression (LS), and ordinary least squares with quadratic features (LS-QF). Naive Net is a neural network with the same architecture as NMMR (described further in Appendix B) that is trained to predict Y directly from A, Z, W by minimizing observational MSE, $\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2$. Least Squares (LS) is the standard linear regression model that predicts Y using a linear combination of A, Z, W. Least Squares with Quadratic Features (LS-QF) is the same as LS but with additional quadratic terms $A^2, W^2, Z^2, AW, AZ, WZ$.

We evaluate NMMR-U, NMMR-V and baseline methods on two synthetic benchmark tasks from Xu et al. [7]. The first is a simulation of how ticket price set by an airline affects the number of ticket sold in the presence of a latent confounder of demand for travel (the Demand experiment). The second is an experiment where the goal is to recover a property of an image that is influenced by an unobserved confounder (the dSprite experiment). The Demand experiment is a low-dimensional estimation problem, whereas dSprite is high-dimensional as A and W are 64x64=4096-dimensional. dSprite also offers the opportunity to leverage image-specific models that are rarely used in the causal inference literature. We note here that both tasks take the measured confounders X to be the null set. For the Demand experiment we evaluate all the methods mentioned above, whereas for the dSprite experiment, 2SLS, LS, and LS-QF were omitted because of their lack of scalability to high-dimensional settings.

The experiments are conducted in PyTorch 1.9.0 (Python 3.9.7), using an A100 40GB or TitanX 12GB GPU and CUDA version 11.2. The experiments can be run in minutes for simpler models (LS, LS-QF, 2SLS) and within several hours for the larger experiments and more complex models (DFPV, NMMR). Implementations of PMMR, KPV, CEVAE, and DFPV and the code to reproduce our experiments can be accessed on GitHub. [1]

6.2 Demand Experiment

Hartford et al. [14] introduced a data generating process for studying instrumental variable regression, and Xu et al. [7] adapted it to the proximal setting. The goal is to estimate the effect of airline ticket price A on sales Y, where these are confounded by demand U (e.g. seasonal fluctuations). For a treatment-inducing proxy, we have information on the cost of fuel $Z = (Z_1, Z_2)$, and for an outcome-inducing proxy, we have information on the number of views on the ticket reservation website W (Figure S1). Additional simulation details and the structural equations underlying the causal DAG can be found in Appendix C.1.

Each method was trained on simulated datasets with sample sizes of 1000, 5000, 10,000, and 50,000. To assess the performance of each method, we evaluated a at 10 equally-spaced intervals between 10 and 30. We compared the difference between each method's estimated average potential outcome $\hat{E}[Y^a]$ against the true average potential outcome $E[Y^a]$, which was obtained by a Monte Carlo simulation (10,000 replicates) from the data generating process with $A = a$ fixed. The evaluation metric is the causal mean squared error (c-MSE) across the 10 evaluation points of a: $\frac{1}{10}\sum_{i=1}^{10}(\mathbb{E}[Y^{a_i}] - \hat{\mathbb{E}}[Y^{a_i}])^2$. For MMR-based methods, their predictions are computed on a heldout dataset, $\mathcal{D}_{\mathcal{W}}$ with 1,000 draws from W where $\hat{\mathbb{E}}[Y^{a_i}] = |D_{\mathcal{W}}|^{-1}\sum_{j}^{|D_{\mathcal{W}}|}\hat{h}(a_i, w_j)$, i.e. a sample average of the estimated bridge function over W. We performed 20 replicates for each method on

[1] https://tinyurl.com/ykatxtys https://tinyurl.com/ymd549bh

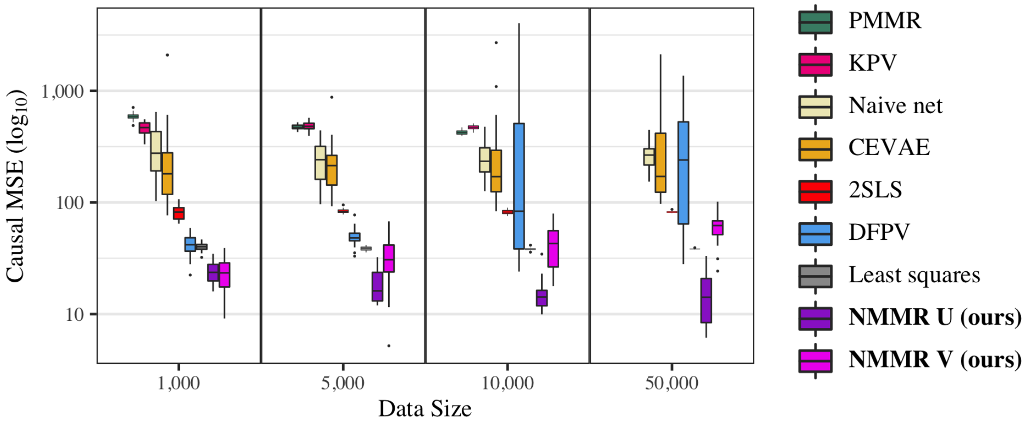

Figure 2: **NMMR-U and NMMR-V achieve state of the art performance across all sample sizes**. Causal MSE (c-MSE) of NMMR and baseline methods in the Demand experiment. Each method was replicated 20 times and evaluated on the same 10 test values of $\mathbb{E}[Y^a]$ each replicate. Each individual box plot represents 20 values of c-MSE. See Table S4 for the statistics of each boxplots

each sample size, where a single method replicate yields one c-MSE value. Figure 2 summarizes the c-MSE distribution for each method across the four sample sizes. NMMR-U has the lowest c-MSE across all sample sizes, with NMMR-V as a close second. Importantly, Least Squares consistently outperforms all other methods besides NMMR. It should also be noted that DFPV encounters difficulties with the larger sample sizes of 10,000 and 50,000, potentially due to convergence issues with its feature maps. Similarly, PMMR and KPV could not be scaled to $n = 50,000$.

For a more in-depth view of the potential outcome curve that each method estimates, we provide the replicate-wise potential outcome prediction curves for each of the 4 sample sizes in Figures S3-S6. We can see that Least Squares estimates relatively unbiased prediction curves due to the nature of the data generating process and further benefits from having very low variance. LS-QF manages to match some of the curvature, although its c-MSE distribution (not shown) is not better than LS. Kernel-based methods, KPV and PMMR, are highly biased. DFPV is less biased, though it still suffers from a lack of flexibility. Both NMMR variants demonstrate the benefit of added flexibility and additionally has relatively lower variance, which results in a lower c-MSE.

Finally, we also varied the magnitude of the variance on the Gaussian noise terms in the structural equations for Z and W in order to examine how each method performs with varying quality proxies for U (see Appendix E). In Figure S9, we can see that NMMR-V is more robust to noised proxies than NMMR-U. This could derive from the fact that the U-statistic is an unbiased, but higher variance, estimator than the V-statistic, so when the proxies are less reliable, the estimate of the risk function $R_k(h)$ is correspondingly less stable. We can also see that the kernel-based methods (KPV and PMMR) rank increasingly well with noisier proxies, which is likely related to the fact that they are less data-adaptive methods. Figures S10 through S17 show the replication-wise prediction curves across all 72 noise levels, with one grid plot per method.

6.3 dSprite Experiment

The second benchmark uses the dSprite dataset from Matthey et al. [23], which was also initially adapted for instrumental variable regression in Xu et al. [21], followed by repurposing for proximal inference in Xu et al. [7]. This image dataset consists of 2D shapes procedurally generated from 6 independent parameters: color, shape, scale, rotation, posX, and posY. All possible combinations of these parameters are present exactly once, generating 737,280 total images. In this experiment, we fix shape = heart, color = white, resulting in 245,760 images, each of which contains 64x64=4096 pixels. The causal DAG for this problem is shown in Figure S7. The structural equations and detailed data generating mechanism underlying the causal DAG can be found in Appendix C.6.

In the DAG, $Fig(\cdot)$ represents the act of retrieving the image from the dSprite dataset with the given arguments. In this experiment, A and W are vectors representing noised images of a heart

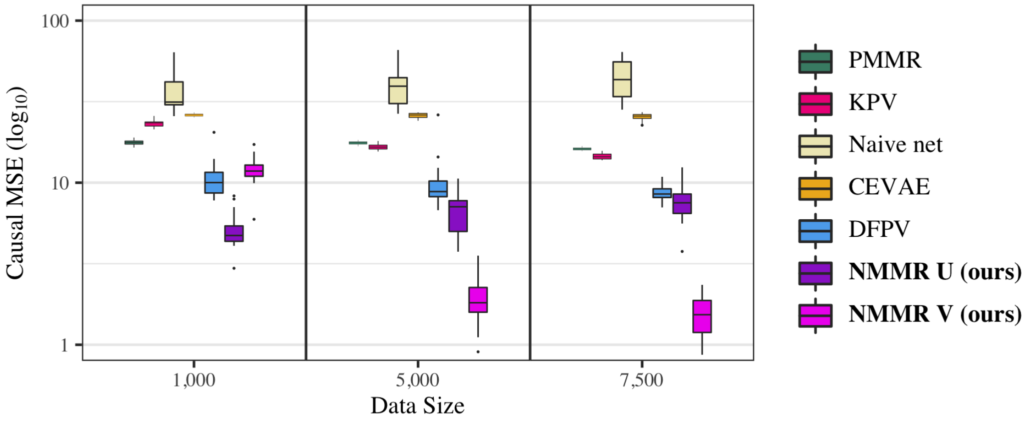

Figure 3: Causal MSE (c-MSE) of NMMR and baseline methods in the dSprite experiment. Each method was replicated 20 times and evaluated on the same 588 test images A each replicate. Each individual box plot represents 20 values of c-MSE. See Table S5 for the statistics of each boxplots

shape, where the heart has a size (*scale*), orientation (*rotation*), horizontal position (*posX*) and vertical position (*posY*). For an exemplar image A and W, see Figure S8. The benchmark computes $\mathbb{E}[Y^a] = \frac{\frac{1}{10}||vec(a)^T B||_2^2 - 5000}{1000}$ where B is a 4096×10 matrix of $\mathcal{U}(0, 1)$ weights from Xu et al. [7]. The observed outcome is computed as $Y = \frac{\frac{1}{10}||vec(A)^T B||_2^2 - 5000}{1000} \times \frac{(31 \times U - 15.5)^2}{85.25} + \epsilon, \epsilon \sim \mathcal{N}(0, 0.5)$. So U dictates the vertical position of the shape in A, as well as the value of Y, making U a confounder of A, Y. U is a discrete uniform random variable with $\mathbb{E}[\frac{(31 \times U - 15.5)^2}{85.25}] = 1$. We hypothesized that a convolutional neural network would be exceptionally strong at recovering this information about U from the images A and W.

Similar to the Demand experiment, we trained each method on simulated datasets with sizes 1,000, 5,000, and 7,500, followed by an evaluation on the same test set as Xu et al. [7]. This test set contains 588 images A that span the range of scale, rotation, posX and posY values (see Appendix C.9) and the 588 corresponding values of $\mathbb{E}[Y^a]$. The evaluation metric is again c-MSE: $\frac{1}{588} \sum_{i=1}^{588} (\mathbb{E}[Y^{a_i}] - \hat{\mathbb{E}}[Y^{a_i}])^2$ and we performed 20 replicates for each method on each sample size. Figure 3 shows that NMMR-U or NMMR-V is consistently lowest in c-MSE, with NMMR-V showing substantial improvement with increasing sample size. Due to the high dimensionality of the images A and W, we could not evaluate Least Squares, LS-QF or 2SLS on this experiment. We can also see that KPV and PMMR do not improve much with increasing sample size. Finally, we also highlight that the Naive net, which uses the same underlying convolutional neural network architecture as NMMR but is trained using observational MSE, performs second-to-worst, with a much larger c-MSE than NMMR-U or NMMR-V. This reinforces the need to use causal knowledge in scenarios where it is available.

7 Conclusion

In this work we have presented a novel method to estimate potential outcomes in the presence of unmeasured confounding using deep neural networks.Though our method is promising, there are several limitations. For very high dimensional data, calculating the kernel matrix K in the loss function can be computationally intensive (see Appendix D). Additionally, mapping real world scenarios to DAGs that satisfy Assumption 1 is non-trivial and technically unverifiable (e.g. we cannot be truly sure that W has no impact on A), though unverifiable assumptions are inherent to causal inference.

In summary, we provide a new single stage estimator and show how it can be trained on a U-statistic based loss in addition existing approaches based on V-statstics. We further prove theoretic convergence properties of our method. On established proximal inference benchmarks, our method achieves state of the art performance in estimating causal quantities. Finally, since our approach is a single-stage neural network, it potentially unlocks new domains for causal inference where deep learning has had success, such as imaging.

References

[1] Miguel A. Hernán and James M. Robins. *Casual inference: What if.* CRC Press, 2021.

[2] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761–767, 03 2011. ISSN 0002-9262. doi: 10.1093/aje/kwq439. URL https://doi.org/10.1093/aje/kwq439.

[3] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[4] Wang Miao, Zhi Geng, and Eric Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, December 2018.

[5] Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. September 2020.

[6] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-Stage estimation and moment restriction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7512–7523. PMLR, 2021.

[7] Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Adv. Neural Inf. Process. Syst.*, 34:26264–26275, December 2021.

[8] Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 41–50. PMLR, 2020.

[9] Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. October 2020.

[10] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, March 2014.

[11] Amiremad Ghassami, Andrew Ying, Ilya Shpitser, and Eric Tchetgen Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. 151:7210–7239, 2022.

[12] Ben Deaner. Proxy controls and panel data. September 2018.

[13] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, September 2003.

[14] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423. PMLR, 2017.

[15] Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. March 2018.

[16] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[17] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Adv. Neural Inf. Process. Syst.*, 33:12248–12262, 2020.

[18] Masahiro Kato, Masaaki Imaizumi, Kenichiro McAlinn, Shota Yasui, and Haruo Kakehi. Learning causal models from conditional moment restrictions by importance weighting. September 2021.

[19] Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, September 1980.

[20] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. September 2018.

[21] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

[22] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[23] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for Large-Scale image recognition. September 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The claims made in the abstract and introduction are reflected in Sections 4, 5, and 6
   (b) Did you describe the limitations of your work? [Yes] See Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1 where we discuss potential medical impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 5 and Appendix A
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Available here `https://github.com/neurips-2022-11442/anonymous-NMMR`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B and C as well as our code
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported results across 20 random seeds in all our Figures and Tables.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See details in Section 6
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We described the fork of assets from Xu et al. [7] in Section 6
   (b) Did you mention the license of the assets? [Yes] See 6. MIT license
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Available here `https://github.com/neurips-2022-11442/anonymous-NMMR`
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Synthetic data
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Synthetic data

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]