Deep Learning Methods for Proximal Inference via Maximum Moment Restriction

Benjamin Kompa*
Department of Biomedical Informatics
Harvard Medical School
benjamin_kompa@hms.harvard.edu

David R. Bellamy*
Department of Epidemiology, CAUSALab
Harvard School of Public Health
david_bellamy@g.harvard.edu

Thomas Kolokotrones
Department of Epidemiology
Harvard School of Public Health
thomas_kolokotrones@hms.harvard.edu

James M. Robins
Department of Epidemiology, CAUSALab
Harvard School of Public Health
robins@hsph.harvard.edu

Andrew L. Beam
Department of Epidemiology, CAUSALab
Harvard School of Public Health
andrew_beam@hms.harvard.edu

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.

*Denotes equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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.

The assumptions are typically unverifiable for continuous data. While model misspecification 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 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 has potential applications in medical settings, where a natural question is 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 trained 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.

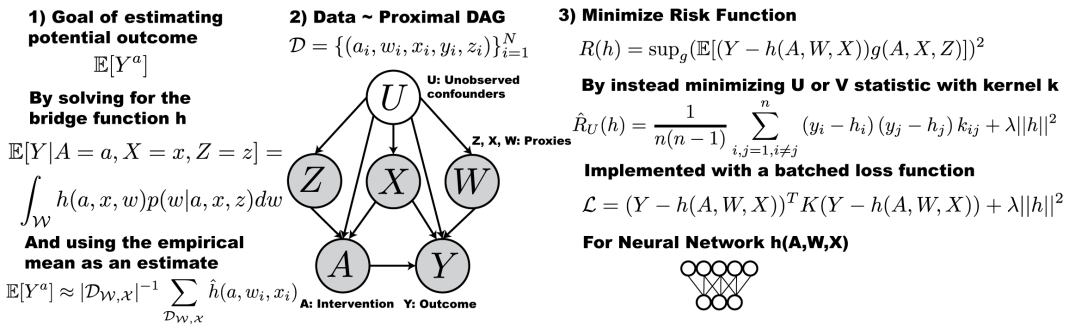


Figure 1: A summary of NMMR. Our method estimates the bridge function h, which can be used to compute the average potential outcome $\mathbb{E}[Y^a]$. We rely on structural assumptions for the causal DAG generating the data. NMMR uses a U or V statistic to train a neural network that solves a risk function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

2 Background: Proximal Inference

In this manuscript, capital, caligraphic, and lowercase letters (e.g. A, $\mathcal{A}$, and a) denote random variables and their corresponding ranges and realizations, respectively. Estimates of random variables and functions will be indicated by hats, e.g. $\hat{Y}$ and $\hat{h}$ are estimates of Y and h, respectively.

Our goal is to estimate, for each level of treatment, a, the expected potential outcome $\mathbb{E}[Y^a]$. Without loss of generality, we refer to A as a treatment, though it could refer to any (possibly continuous) intervention. Proximal inference allows us to do this, in the presence of unobserved confounders, U, provided that we have a sufficiently rich set of proxies, (W, Z), that obey certain structural assumptions. We may also include observed confounders, X. We also require the following:

Assumption 1. *Given* (A, U, W, X, Y, Z), $Y \perp\!\!\!\perp Z|A, U, X$ *and* $W \perp\!\!\!\perp (A, Z)|U, X$.

Figure 1 provides an example of a DAG that satisfies these assumptions. The following completeness conditions formalize the notion that the proxies are "sufficiently rich":

Assumption 2. *For all* $f \in L^2$ *and all* $a \in \mathcal{A}, x \in \mathcal{X}$, $\mathbb{E}[f(U)|A=a, X=x, Z=z] = 0$ *for all* $z \in \mathcal{Z}$ *if and only if* $f(U) = 0$ *almost surely.*

Assumption 3. *For all* $f \in L^2$ *and all* $a \in \mathcal{A}, x \in \mathcal{X}$, $\mathbb{E}[f(Z)|A=a, W=w, X=x] = 0$ *for all* $w \in \mathcal{W}$ *if and only if* $f(Z) = 0$ *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 zeros.

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. $k : (\mathcal{A} \times \mathcal{X} \times \mathcal{Z})^2 \to \mathbb{R}$ *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; its solution, h, is often called the "bridge function." Theorem 1 of Miao et al. [4] shows that the expected potential outcomes are given by:

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

We can obtain unbiased estimates of the expected potential outcomes, $\mathbb{E}[Y^a]$, by splitting the sample, using the first part of the data to estimate the bridge function h, by some $\hat{h}$, and using the

second part of the data to compute the empirical mean of $\hat{h}$ with a fixed to the value of interest, $\hat{\mathbb{E}}[Y^a] = \frac{1}{M}\sum_{i=1}^{M} \hat{h}(a, w_i, x_i)$. From these, we can obtain other quantities of interest, like the ACE.

In what follows, all norms will be L^2 with respect to the relevant probability measure, unless otherwise noted. If necessary, we will explicitly denote the norm of f with respect to a probability measure $\mathcal{P}_{U,V,\ldots}$ by $\|f\|_{\mathcal{P}_{U,V,\ldots}}$ or $\|f\|_{2,\mathcal{P}_{U,V,\ldots}}$.

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, W, to "recover" the distribution of U, potentially using external datasets to estimate $p(w|u)$. Tchetgen-Tchetgen and colleagues extended these results to allow for identification without recovery of U in Miao et al. [4] and Tchetgen Tchetgen et al. [5], also providing a 2-Stage Least Squares (2SLS) method to identify and estimate causal effects under the assumption that the bridge function is linear. Cui et al. [11] introduced a bridge function for Inverse Probability Weighting (IPW), which enabled IPW and Doubly Robust (DR) proximal estimators, for which they presented influence functions under binary treatment. This was extended in Ghassami et al. [12] and further explored by Kallus et al. [13], who provided alternative identification assumptions as well as results for general treatments. The latter two works also consider the use of adversarial methods for estimation, which were previously utilized by Lewis and Syrgkanis [14] and Bennett et al. [15] in the Instrumental Variable (IV) setting and Dikkala et al. [16] in conditional moment models.

Other early investigators include Deaner [17], who developed machine learning techniques for proximal inference introducing a method based 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, allowing more flexibility than the linear basis of Tchetgen Tchetgen et al. [5]. Xu et al. [7] further improved upon this with 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. Singh and colleagues also considered two stage kernel models in the IV setting [18] and RKHS techniques for proximal inference [19].

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 [20]. Zhang et al. [9] introduced the MMR framework to the IV setting, which can be considered a subset of proximal inference without outcome-inducing proxies W and with additional exclusion restrictions [5]. There are now several machine learning methods that can be applied in the IV setting [21, 14, 18, 16, 22]

Of note, Zhang et al. [9] introduced MMR-IV, which is related to work by Lewis and Syrgkanis [14] and Dikkala et al. [16]. MMR-IV involves optimizing a family of risk functions based on U or V-statistics [23]. However, Zhang et al. [9] only considered IV, rather than proximal, inference and only optimized neural networks by a loss that corresponds to the V-statistic. The V-statistic provides a biased estimate [23] 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): a method to estimate expected potential outcomes $\mathbb{E}[Y^a]$ in the presence of unmeasured confounding. We use deep neural networks due to their flexibility, scalability, and adaptability to diverse data types (e.g. using convolutions for images). By rewriting maximum moment restrictions [8] as U and V-statistics [6, 9], we show how a single-stage neural network procedure can be used to estimate the bridge function.

Following Muandet et al. [8] and Zhang et al. [9], we rewrite the integral equation (1) as a *conditional moment restriction*: $\mathbb{E}[Y - h(A, W, X)|A, X, Z] = 0$. Then, for any measurable $g : \mathcal{A} \times \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, $\mathbb{E}[(Y - h(A, W, X))g(A, X, Z)|A, X, Z] = 0$, and, thus, $\mathbb{E}[(Y - h(A, W, X))g(A, X, Z)] = 0$, so we can use unconditional expectations instead of conditional ones. This is the basis of the MMR framework of Muandet et al. [8]. Since this yields an infinite number of moment restrictions we employ a minimax strategy to estimate h by minimizing the risk $R(h)$ for the worst-case value of g:

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

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 expected potential outcome by using Equation 2.

The empirical risk $\hat{R}_{k,n}$ given data $\mathcal{D} = \{(a_i, w_i, x_i, y_i, z_i)\}_{i=1}^N$ can be written as either a U or V-statistic, respectively [23]:

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

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

where $h_i = h(a_i, w_i, x_i)$ and $k_{ij} = k((a_i, z_i, x_i), (a_j, z_j, x_j))$. $\hat{R}_{k,U,n}(h)$ is the minimum variance unbiased estimator of $R_k(h)$ [23], while $\hat{R}_{k,V,n}(h)$ is a biased estimator of $R_k(h)$. In order to prevent overfitting, we add an additional penalty to our risk function $\Lambda : \mathcal{H} \times \Theta_{\mathcal{H}} \to \mathbb{R}_+$, which is a function of h as well as, possibly, its parameters, θ_h (e.g. network weights). Specifically, we take Λ to be an L^2 penalty on the weights so $\Lambda[h, \theta_h] = \sum_i \theta_{h,i}^2$. We then denote the penalized risk functions by $\hat{R}_{k,U,\lambda,n}(h) = \hat{R}_{k,U,n}(h) + \lambda\Lambda[h, \theta_h]$ and $\hat{R}_{k,V,\lambda,n}(h) = \hat{R}_{k,V,n}(h) + \lambda\Lambda[h, \theta_h]$, respectively. In practice, $\hat{R}_{k,U,\lambda,n}(h)$ is slightly biased, but, in simulations, is much less biased than even the unpenalized $\hat{R}_{k,V,n}(h)$. 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 the regularized loss function:

$$\mathcal{L} = (Y - h(A, W, X))^t K (Y - h(A, W, X)) + \lambda\Lambda[h, \theta_h]$$

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}. We choose k to be an RBF kernel (see Appendix B). If $\mathcal{L}$ represents a V-statistic, we include the main diagonal elements of K, while if $\mathcal{L}$ represents a U-statistic, 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 expected potential outcome with data from a held-out dataset with M data points

$$D_{\mathcal{W},\mathcal{X}} = \{(w_i, x_i)\}_{i=1}^M, \quad \mathbb{E}[\hat{Y}^a] = \frac{1}{M} \sum_{i=1}^{M} \hat{h}(a, w_i, x_i)$$

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 $\hat{R}_{k,\lambda,n}(h) = \hat{R}_{k,n}(h) + \lambda\Lambda\left[h, \theta_h\right]$ (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 $\hat{h}_{k,U,\lambda,n}$ minimize $\hat{R}_{k,U,\lambda,n}(h)$ for $h \in \mathcal{H}$, $k : (\mathcal{A} \times \mathcal{X} \times \mathcal{Z})^2 \to [-M_k, M_k]$, $\Lambda : \mathcal{H} \times \Theta_h \to [0, M_\lambda]$, 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*

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

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

Also let,

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

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

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

Further, if Assumption 5 holds, so k is ISPD, then d_k is a metric on $L^2_{\mathcal{AXZ}}$ and, if the right hand side of the inequality goes to zero as n goes to infinity,
$d_k\left(\mathbb{E}\left[h^*|A, X, Z\right] - \mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right]\right) \xrightarrow{\mathrm{P}} 0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A, X, Z\right]$ *in d_k. Also,* $\left\|\mathbb{E}\left[h^*|A, X, Z\right] - \mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right]\right\|_{\mathrm{P}_{A,X,Z}} \xrightarrow{\mathrm{P}} 0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A, X, Z\right]$ *in $L^2\left(\mathrm{P}_{\mathcal{A},\mathcal{X},\mathcal{Z}}\right) - norm$.*

$$\mathcal{F}'_{a,x,z} = \left\{f_{a,x,z} \mid \exists_{h\in\mathcal{H}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}} f_{a,x,z}\left(a', w', x', z'\right) = h\left(a', w', x'\right) k\left(\left(a', x', z'\right), \left(a, x, z\right)\right)\right\}$$
$$\mathcal{F}' = \left\{f \mid \exists_{h\in\mathcal{H},a\in\mathcal{A},x\in\mathcal{X},z\in\mathcal{Z}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}} f\left(a', w', x', z'\right) = h\left(a', w', x'\right) k\left(\left(a', x', z'\right), \left(a, x, z\right)\right)\right\}$$

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. In Theorem 1 if the quadratic form converges at

a particular rate, say $n^{-\frac{1}{2}}$, $\mathbb{E}\left[\hat{h}\middle|A,X,Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A,X,Z\right]$, under the metric induced by the kernel, d_k, at half the rate, in this case $n^{-\frac{1}{4}}$. This is similar to Kallus et al. [13]'s findings in the unstabilized case.

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. 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. [24] 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, 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 (likely $\sqrt{n}$) estimator of h^*.

We can also compare the convergence of the estimated bridge function to that of its projection onto $L^2_{\mathcal{AXZ}}$. Prior literature has focused on a measure of "ill-posedness" $\tau = \sup_{h\in\mathcal{H}} \|h - h^*\|_2 \left\|\mathrm{E}\left[h - h^*|A,Z,X\right]\right\|_2^{-1}$. If τ is finite, then the rate of convergence of the estimated bridge function will be at worst τ times that of its projection (it will be slower by a factor of τ). This will be the case whether we measure convergence using $\left\|\mathrm{E}\left[h - h^*|A,Z,X\right]\right\|_2$ or the metric induced by k.

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. Structural approaches leverage causal information about the data generating process. They include Kernel Proxy Variables (KPV) [6], Proximal Maximum Moment Restriction (PMMR) [6], Deep Feature Proxy Variables (DFPV) [25], Causal Effect Variational Autoencoder[26] (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. [26]. 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, instead directly regressing A and 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 that has undergone the same architecture search as NMMR (described further in Appendix B) that is trained to predict Y directly from A and 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 and W. Least Squares with Quadratic Features (LS-QF) is the same as LS but with additional quadratic terms A^2, W^2, AW.

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 prices affect the number of tickets sold in the presence of a latent confounder: 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 leverages image-

specific models, which are rarely used in the causal inference literature. Neither task uses X. For the Demand experiment we evaluate all the methods mentioned above, whereas for the dSprite experiment, we omit 2SLS, LS, and LS-QF because of their lack of scalability to high-dimensional settings.

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

6.2 Demand Experiment

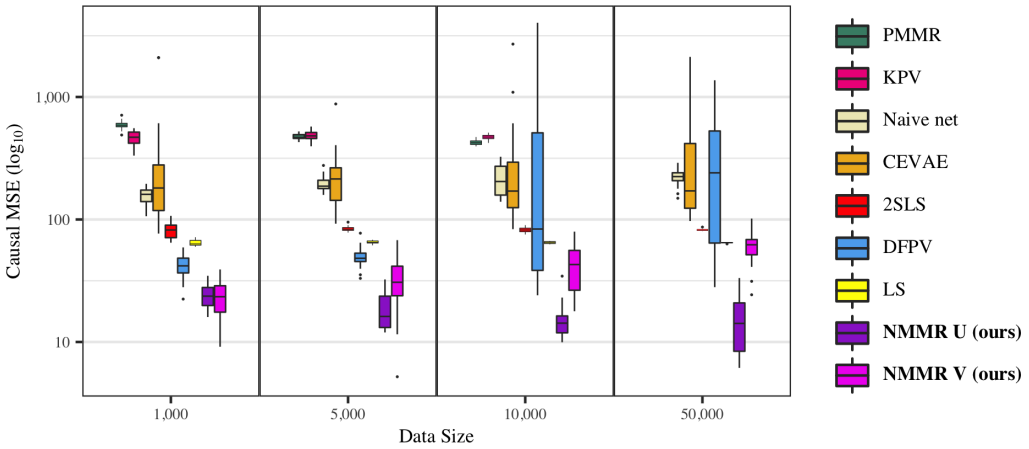


Figure 2: **NMMR-U and NMMR-V achieve state of the art performance across all sample sizes**. Causal MSE (c-MSE) of NMMR and baseline methods in the Demand experiment. Each method was replicated 20 times and evaluated on the same 10 test values of $\mathbb{E}[Y^a]$ each replicate. Each individual box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

Hartford et al. [21] 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, which is confounded by demand U (e.g. seasonal fluctuations). We use the cost of fuel, $Z = (Z_1, Z_2)$, as a treatment-inducing proxy and number of views at a ticket reservation website, W, as an outcome-inducing proxy (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 each method's estimated potential outcomes, $\hat{E}[Y^a]$, against estimates of the truth, $E[Y^a]$, obtained from Monte Carlo simulations (10,000 replicates) of the data generating process for each a. 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, predictions are computed using a heldout dataset, $\mathcal{D}_{\mathcal{W}}$ with 1,000 draws from W so $\hat{\mathbb{E}}[Y^{a_i}] = |D_{\mathcal{W}}|^{-1}\sum_j^{|D_{\mathcal{W}}|}\hat{h}(a_i, w_j)$, i.e. a sample average of the estimated bridge function over W. We performed 20 replicates for each method on each sample size, where a single 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 a close second. 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 scale to $n = 50,000$.

For a more in-depth view of the potential outcome curve estimated by each method, we provide replicate-wise potential outcome prediction curves for each of the 4 sample sizes in Figures S3-S6. Least Squares estimates relatively unbiased prediction curves due to the nature of the data generating process and has very low variance. LS-QF matches some of the curvature, although its c-MSE

[2]https://github.com/beamlab-hsph/Neural-Moment-Matching-Regression

distribution (not shown) is not better than LS. Kernel-based methods, KPV and PMMR, are highly biased. DFPV is less biased, but still suffers from a lack of flexibility. Both NMMR variants demonstrate the benefit of added flexibility and have lower variance, resulting in a lower c-MSE.

Finally, we also varied the variance of 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). Figure S10, shows that NMMR-V is more robust to proxy noise than NMMR-U. This could be because U-statistics yield unbiased, but higher variance, estimators than V-statistics, so, when proxies are less reliable, the estimated risk function $R_k(h)$ is less stable. Kernel-based methods (KPV and PMMR) perform increasingly well with noisier proxies, which is likely related to the fact that they are less data-adaptive. Figures S11 through S18 show replication-wise prediction curves across all 72 noise levels, with one grid plot per method.

6.3 dSprite Experiment

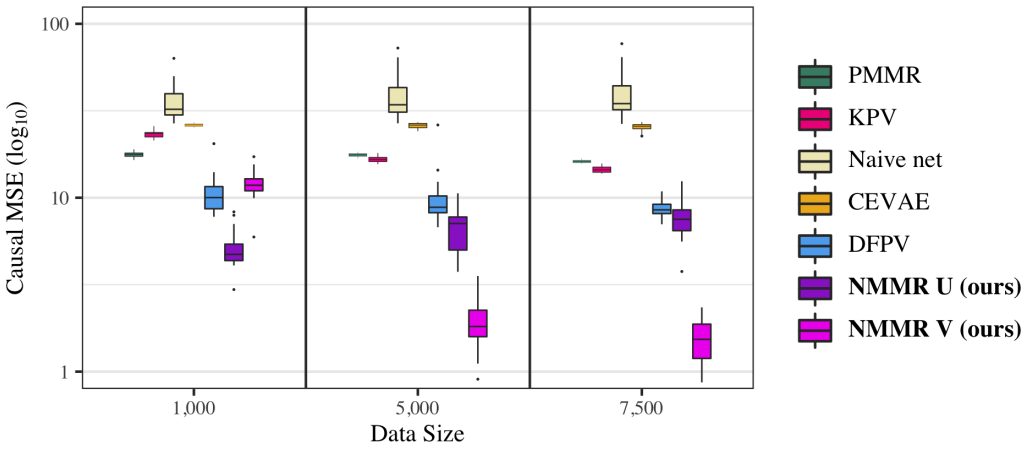


Figure 3: Causal MSE (c-MSE) of NMMR and baseline methods in the dSprite experiment. Each method was replicated 20 times and evaluated on the same 588 test images A each replicate. Each individual box plot represents 20 values of c-MSE. See Table S5 for the statistics of each boxplots

The second benchmark uses the dSprite dataset from Matthey et al. [27], which was initially adapted to instrumental variable regression in Xu et al. [25], and repurposed 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. A and W are vectors representing noised images of a heart 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

$$\mathbb{E}[Y^a] = \frac{\frac{1}{10}\left\|vec(a)^t B\right\|_2^2 - 5000}{1000}$$

where B is a 4096×10 matrix of $\mathcal{U}(0,1)$ weights from Xu et al. [7]. The observed outcome is computed as

$$Y = \frac{\frac{1}{10}\left\|vec(A)^t B\right\|_2^2 - 5000}{1000} \times \frac{(31 \times U - 15.5)^2}{85.25} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.5)$$

U is a discrete uniform random variable with

$$\mathbb{E}\left[\frac{(31 \times U - 15.5)^2}{85.25}\right] = 1$$

that dictates the vertical position of the shape in A, as well as the value of Y, making U a confounder of A, Y. 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:

$$\frac{1}{588}\sum_{i=1}^{588}\left(\mathbb{E}[Y^{a_i}] - \hat{\mathbb{E}}[Y^{a_i}]\right)^2$$

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. KPV and PMMR do not improve much with increasing sample size. 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, it has 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.

Further, the present work focuses on methods that estimate only the outcome bridge function, rather than also estimating the IPW bridge function, which would permit us to construct a doubly robust estimator, as is done in Cui et al. [11] and Kallus et al. [13]. However, our method extends naturally to this setting and we expect to explore such estimators in future work.

In summary, we provide a new single stage estimator and show how it can be trained on a U-statistic based loss in addition to existing approaches based on V-statistics. We further prove theoretical 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] Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. November 2020.

[12] 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.

[13] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. March 2021.

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

[15] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. May 2019.

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

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

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

[19] Rahul Singh. Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. December 2020.

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

[21] 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.

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

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

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

[25] 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.

[26] 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.

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

[28] 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/beamlab-hsph/Neural-Moment-Matching-Regression`
   (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/beamlab-hsph/Neural-Moment-Matching-Regression`
   (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]

A Proofs

Lemma 3. *Let X be a random variable taking values in $\mathcal{X}$ and let $\mathcal{F}$ be a family of measurable functions with $f \in \mathcal{F} : \mathcal{X}^2 \to [-M, M]$, with each $f \in \mathcal{F}$ optionally, and possibly not uniquely, (partially) parameterized by $\theta_f \in \Theta_{\mathcal{F}}$, $\Lambda : \mathcal{F} \times \Theta_{\mathcal{F}} \to [0, M_\Lambda]$, and $\lambda \geq 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, any IID sample $S = \{x_i\}_{i=1}^n$ drawn from P_X satisfies*

$$\mathbb{E}f = \mathbb{E}_{X,X'} f(X, X') \leq \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} f(x_i, x_j) + \lambda\Lambda[f, \theta_f] + 2\left(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)\right) + 2M\left(\frac{2}{n}\log\frac{1}{\delta}\right)^{\frac{1}{2}}$$

$$\frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} f(x_i, x_j) + \lambda\Lambda[f, \theta_f] \leq \mathbb{E}f + \lambda M_\Lambda + 2\left(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)\right) + 2M\left(\frac{2}{n}\log\frac{1}{\delta}\right)^{\frac{1}{2}}$$

$$\mathbb{E}f = \mathbb{E}_{X,X'} f(X, X') \leq \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} f(x_i, x_j) + \lambda\Lambda[f, \theta_f] + 2\left(\hat{\mathcal{R}}_{n-1,S}(\mathcal{F}_1) + \mathcal{R}_S(\mathcal{F}_2)\right) + 6M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$$

$$\frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} f(x_i, x_j) + \lambda\Lambda[f, \theta_f] \leq \mathbb{E}f + \lambda M_\Lambda + 2\left(\hat{\mathcal{R}}_{n-1,S}(\mathcal{F}_1) + \mathcal{R}_S(\mathcal{F}_2)\right) + 6M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$$

where $\mathcal{R}_S$ is the empirical Rademacher Complexity, given by:

$$\mathcal{R}_S(\mathcal{F}) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)$$

where ϵ is a Rademacher random vector taking values uniformly in $\{-1, 1\}^n$, $S_{-i} = \{x_j\}_{j=1, j\neq i}^n$, $\hat{\mathcal{R}}_{n-1,S}(\mathcal{F}) = n^{-1}\sum_{i=1}^n \mathcal{R}_{S_{-i}}(\mathcal{F})$, $\mathcal{F}_1 = \{g \mid \exists_{f\in\mathcal{F}, x'\in\mathcal{X}} \forall_{x\in\mathcal{X}} [g(x) = f(x', x)]\}$, $\mathcal{F}_2 = \{g \mid \exists_{f\in\mathcal{F}, x'\in\mathcal{X}} \forall_{x\in\mathcal{X}} [g(x) = f(x, x')]\}$, and $\mathcal{R}_n$ is the (expected) Rademacher complexity for a sample of size n, $\mathcal{R}_n = \mathbb{E}_S\mathcal{R}_S$, where the expectation is over all samples, S, of size n.

Finally, if the elements of $\mathcal{F}$ are symmetric, so that $\forall_{f\in\mathcal{F}, x, x'\in\mathcal{X}} (f(x, x') = f(x', x))$, $\mathcal{F}_1 = \mathcal{F}_2$.

In particular, θ_f might be the weights associated with the neural network f. Note that, as is the case for neural networks, θ_f may not be uniquely determined by f, so that multiple θs may be associated with the same f. We may also take $\Theta_{\mathcal{F}} = \emptyset$, so that f is regarded purely as an element of $\mathcal{F}$.

Proof. Let $\hat{E}_S f = \frac{1}{n(n-1)}\sum_{i,j=1,j\neq i} f(x_i, x_j)$, $\hat{E}_{S,\lambda} f = \frac{1}{n(n-1)}\sum_{i,j=1,j\neq i} f(x_i, x_j) + \lambda\Lambda[f, \theta_f]$, $\tilde{E}_S f = \frac{1}{n}\sum_{i=1}^n E_X f(x_i, X)$, $\phi(S) = \sup_{f\in\mathcal{F}}\left(\hat{\mathbb{E}}_S f - \mathbb{E}f\right)$, $\phi^{(1)}(S) = \sup_{f\in\mathcal{F}}\left(\hat{E}_S f - \tilde{E}_S f\right)$, and $\phi^{(2)}(S) = \sup_{f\in\mathcal{F}}\left(\tilde{\mathbb{E}}_S f - \mathbb{E}f\right)$. Note that $\mathbb{E}_S\hat{\mathbb{E}}_S f = \mathbb{E}_{X,X'} f(X, X') = \mathbb{E}f = \mathbb{E}_{X,X'} f(X, X') = \mathbb{E}_S\tilde{\mathbb{E}}_S f$. Also, let $S' = \{x_i'\}_{i=1}^n$ and let $S_i = \{x_{i,j}\}_{j=1}^n$ be obtained from S by replacing x_i by x_i', so that $x_{i,j} = x_j$ for $j \neq i$ and $x_{i,i} = x_i'$. In order to apply McDiarmid's Inequality, we must find bounds, c_i such that $|\phi(S_i) - \phi(S)| \leq c_i$ for $i = 1, \ldots, n$.

$$
\begin{aligned}
|\phi(S_i) - \phi(S)| &= \left| \sup_{f\in\mathcal{F}} \left(\hat{\mathbb{E}}_{S_i} f - \mathbb{E} f\right) - \sup_{f\in\mathcal{F}} \left(\hat{\mathbb{E}}_{S} f - \mathbb{E} f\right) \right| \\
&\leq \sup_{f\in\mathcal{F}} \left| \left(\hat{\mathbb{E}}_{S_i} f - \mathbb{E} f\right) - \left(\hat{\mathbb{E}}_{S} f - \mathbb{E} f\right) \right| \\
&= \sup_{f\in\mathcal{F}} \left| \hat{\mathbb{E}}_{S_i} f - \hat{\mathbb{E}}_{S} f \right| \\
&= \sup_{f\in\mathcal{F}} \left| \frac{1}{n(n-1)} \sum_{j,k=1,k\neq j}^{n} f(x_{i,j}, x_{i,k}) - \frac{1}{n(n-1)} \sum_{j,k=1,k\neq j}^{n} f(x_j, x_k) \right| \\
&\leq \frac{1}{n(n-1)} \sup_{f\in\mathcal{F}} \sum_{j,k=1,k\neq j}^{n} |f(x_{i,j}, x_{i,k}) - f(x_j, x_k)| \\
&= \frac{1}{n(n-1)} \left(\sum_{j=1,j\neq k,k=i}^{n} \sup_{f\in\mathcal{F}} |f(x_{i,j}, x_{i,i}) - f(x_j, x_i)| \right. \\
&\qquad \left. + \sum_{k=1,k\neq j,j=i}^{n} \sup_{f\in\mathcal{F}} |f(x_{i,i}, x_{i,k}) - f(x_i, x_k)| \right) \\
&= \frac{1}{n(n-1)} \left(\sum_{j=1,j\neq i}^{n} \sup_{f\in\mathcal{F}} |f(x_j, x_i') - f(x_j, x_i)| \right. \\
&\qquad \left. + \sum_{k=1,k\neq i}^{n} \sup_{f\in\mathcal{F}} |f(x_i', x_k) - f(x_i, x_k)| \right) \\
&= \frac{1}{n(n-1)} \left(\sum_{j=1,j\neq i}^{n} \sup_{f\in\mathcal{F}} 2M + \sum_{k=1,k\neq i}^{n} \sup_{f\in\mathcal{F}} 2M \right) \leq \frac{2}{n(n-1)} \cdot (n-1) \cdot 2M \\
&= 4Mn^{-1}
\end{aligned}
$$

so we can choose $c_i = c = 4Mn^{-1}$. The exponent in McDiarmid's Inequality is then $-2\epsilon^2 \left(\sum_{i=1}^n c_i^2\right)^{-1} = -2\epsilon^2 \left(n \cdot \left(4Mn^{-1}\right)^2\right)^{-1} = -2\epsilon^2 \left(16M^2 n^{-1}\right)^{-1} = -\frac{1}{8} n M^{-2} \epsilon^2$. Setting $\frac{\delta}{2} = e^{-\frac{1}{8} n M^{-2} \epsilon^2}$ gives $\epsilon = \left(-\frac{8M^2}{n} \log \frac{\delta}{2}\right)^{\frac{1}{2}} = M \left(\frac{8}{n} \log \frac{2}{\delta}\right)^{\frac{1}{2}} = 2M \left(\frac{2}{n} \log \frac{2}{\delta}\right)^{\frac{1}{2}}$. Then, McDiarmid's Inequality yields,

$$
\mathrm{P}\left[\phi(S) - \mathbb{E}_S \phi(S) \geq \epsilon\right] \leq \frac{\delta}{2}, \quad \mathrm{P}\left[\mathbb{E}_S \phi(S) - \phi(S) \geq \epsilon\right] \leq \frac{\delta}{2}
$$

so that, with probability $1 - \frac{\delta}{2}$, $\phi(S) \leq \mathbb{E}_S \phi(S) + 2M \left(\frac{2}{n} \log \frac{2}{\delta}\right)^{\frac{1}{2}}$. We now need to compute $\mathbb{E}_S \phi(S)$. However, this is difficult to do directly, so we instead compute it separately for $\phi^{(1)}$ and $\phi^{(2)}$. Let ϵ be a Rademacher random vector taking values uniformly in $\{-1, 1\}^n$. Then,

$$
\begin{aligned}
\mathbb{E}_S \phi^{(1)}(S) &= \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\hat{E}_S f - \tilde{E}_S f\right) \\
&= \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\frac{1}{n(n-1)} \sum_{i,j=1,j\neq i}^{n} f\left(x_i, x_j\right) - \frac{1}{n}\sum_{i=1}^{n} E_X f\left(x_i, X\right)\right) \\
&= \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\frac{1}{n(n-1)} \sum_{i,j=1,j\neq i}^{n} f\left(x_i, x_j\right) - \frac{1}{n}\sum_{i=1}^{n} E_{S'} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} f\left(x_i, x'_j\right)\right) \\
&= \mathbb{E}_S \sup_{f\in\mathcal{F}} E_{S'} \frac{1}{n(n-1)} \sum_{i,j=1,j\neq i}^{n} \left(f\left(x_i, x_j\right) - f\left(x_i, x'_j\right)\right) \\
&\leq \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \frac{1}{n(n-1)} \sum_{i,j=1,j\neq i}^{n} \left(f\left(x_i, x_j\right) - f\left(x_i, x'_j\right)\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \left(f\left(x_i, x_j\right) - f\left(x_i, x'_j\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{x_i, S_{-i}, S'_{-i}} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \left(f\left(x_i, x_j\right) - f\left(x_i, x'_j\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\epsilon}\mathbb{E}_{x_i, S_{-i}, S'_{-i}} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \epsilon_j \left(f\left(x_i, x_j\right) - f\left(x_i, x'_j\right)\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \Bigg[\mathbb{E}_{x_i, S_{-i}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \epsilon_j f\left(x_i, x_j\right) \\
&\qquad + \mathbb{E}_{x_i, S'_{-i}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} -\epsilon_j f\left(x_i, x'_j\right)\Bigg] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{x_i} \Bigg[\mathbb{E}_{S_{-i}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \epsilon_j f\left(x_i, x_j\right) \\
&\qquad + \mathbb{E}_{S_{-i}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \epsilon_j f\left(x_i, x_j\right)\Bigg] \\
&= \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{x_i}\mathbb{E}_{S_{-i}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1,j\neq i}^{n} \epsilon_j f\left(x_i, x_j\right) \\
&= \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{x_n}\mathbb{E}_{S_{-n}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1}^{n-1} \epsilon_j f\left(x_n, x_j\right) \\
&= 2\mathbb{E}_{x_n}\mathbb{E}_{S_{-n}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{j=1}^{n-1} \epsilon_j f\left(x_n, x_j\right) = 2\mathbb{E}_{X}\mathbb{E}_{S_{-n}, \epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n-1} \sum_{i=1}^{n-1} \epsilon_i f\left(X, x_i\right) \\
&\leq 2\mathbb{E}_{S_{-n}, \epsilon} \sup_{f\in\mathcal{F}, x\in\mathcal{X}} \frac{1}{n-1} \sum_{i=1}^{n-1} \epsilon_i f\left(x, x_i\right) = 2\mathbb{E}_{S_{-n}}\mathbb{E}_{\epsilon} \sup_{f\in\mathcal{F}, x\in\mathcal{X}} \frac{1}{n-1} \sum_{i=1}^{n-1} \epsilon_i f\left(x, x_i\right) \\
&= 2\mathcal{R}_{n-1}\left(\mathcal{F}_1\right)
\end{aligned}
$$

where, in the seventh line, we note that the inner sum depends only on $S = \{x_i\} \cup S_{-i}$ and S'_{-i}, but not x'_i, in the eighth line, we introduce Rademacher variables because reversing the order of the difference is equivalent to swapping elements between S_{-i} and S'_{-i} and, since the expectation is over

all possible pairs of samples, its value is unchanged, in the tenth line, we note that negation simply interchanges pairs of Rademacher vectors, so the expectation is unchanged, and, in the final line, $\mathcal{F}_1 = \{g \mid \exists_{f\in\mathcal{F},x'\in\mathcal{X}} \forall_{x\in\mathcal{X}} [g(x) = f(x', x)]\}$.

$$
\begin{aligned}
\mathbb{E}_S\phi^{(2)}(S) &= \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\tilde{\mathbb{E}}_S f - \mathbb{E}f\right) = \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\tilde{\mathbb{E}}_S f - \mathbb{E}_{S'}\tilde{\mathbb{E}}_{S'} f\right) = \mathbb{E}_S \sup_{f\in\mathcal{F}} \mathbb{E}_{S'} \left(\tilde{\mathbb{E}}_S f - \tilde{\mathbb{E}}_{S'} f\right) \\
&\leq \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \left(\tilde{\mathbb{E}}_S f - \tilde{\mathbb{E}}_{S'} f\right) \\
&= \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n E_X f(x_i, X) - \frac{1}{n}\sum_{i=1}^n E_X f(x_i', X)\right) \\
&= \mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n (E_X f(x_i, X) - E_X f(x_i', X))\right) \\
&= \mathbb{E}_\epsilon\mathbb{E}_{S,S'} \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i (E_X f(x_i, X) - E_X f(x_i', X))\right) \\
&\leq \mathbb{E}_{S,\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i E_X f(x_i, X) + \mathbb{E}_{S',\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n -\epsilon_i E_X f(x_i', X) \\
&= \mathbb{E}_{S,\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i E_X f(x_i, X) + \mathbb{E}_{S,\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i E_X f(x_i, X) \\
&= 2\mathbb{E}_{S,\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i E_X f(x_i, X) \leq 2\mathbb{E}_{S,\epsilon} \sup_{f\in\mathcal{F},x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i, x) \\
&= 2\mathcal{R}_n(\mathcal{F}_2)
\end{aligned}
$$

where, in the fifth line, we introduce Rademacher variables because changing the order of the difference is equivalent to swapping elements between S and S', and, since the expectation is over all possible pairs of samples, its value is unchanged, in the seventh line, we note that negation simply interchanges pairs of Rademacher vectors leaving the expectation unchanged, and, in the final line, $\mathcal{F}_2 = \{g \mid \exists_{f\in\mathcal{F},x'\in\mathcal{X}} \forall_{x\in\mathcal{X}} [g(x) = f(x, x')]\}$. Then,

$$
\begin{aligned}
\mathbb{E}_S\phi(S) &= \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\hat{\mathbb{E}}_S f - \mathbb{E}f\right) = \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\hat{\mathbb{E}}_S f - \tilde{\mathbb{E}}_S f + \tilde{\mathbb{E}}_S f - \mathbb{E}f\right) \\
&\leq \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\hat{\mathbb{E}}_S f - \tilde{\mathbb{E}}_S f\right) + \mathbb{E}_S \sup_{f\in\mathcal{F}} \left(\tilde{\mathbb{E}}_S f - \mathbb{E}f\right) = \mathbb{E}_S\phi^{(1)}(S) + \mathbb{E}_S\phi^{(2)}(S) \\
&= 2\mathcal{R}_{n-1}(\mathcal{F}_1) + 2\mathcal{R}_n(\mathcal{F}_2)
\end{aligned}
$$

Finally, combining the above results tells us that, with probability $1 - \frac{\delta}{2}$, $\phi(S) \leq 2(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)) + 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$ so $\hat{\mathbb{E}}_S f \leq \mathbb{E}f + 2(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)) + 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$. Replacing $\phi(S)$ by $\phi'(S) = \sup_{f\in\mathcal{F}}\left(\mathbb{E}f - \hat{\mathbb{E}}_S f\right)$ in the above proof yields $\mathbb{E}f \leq \hat{\mathbb{E}}_S f + 2(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)) + 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$. Since $\lambda \geq 0$, $0 \leq \Lambda \leq M_\Lambda$, $0 \leq \lambda\Lambda[f, \theta_f] \leq \lambda M_\Lambda$, so we also have,

$$
\begin{aligned}
\mathbb{E}f &\leq \hat{\mathbb{E}}_S f + \lambda\Lambda[f, \theta_f] + 2(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)) + 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&= \hat{\mathbb{E}}_{S,\lambda} f + 2(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)) + 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}
\end{aligned}
$$

$$
\hat{\mathbb{E}}_{S,\lambda} f = \hat{\mathbb{E}}_S f + \lambda\Lambda[f, \theta_f] \leq \mathbb{E}f + \lambda M_\Lambda + 2(\mathcal{R}_{n-1}(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)) + 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}
$$

Using 2δ in place of δ, we see that, with probability at least $1-\delta$, $\mathbb{E}f \leq \hat{\mathbb{E}}_{S,\lambda}f + 2\left(\mathcal{R}_{n-1}\left(\mathcal{F}_1\right)+\mathcal{R}_n\left(\mathcal{F}_2\right)\right)+2M\left(\frac{2}{n}\log\frac{1}{\delta}\right)^{\frac{1}{2}}$ and $\hat{\mathbb{E}}_{S,\lambda}f \leq \mathbb{E}f+\lambda M_\Lambda+2\left(\mathcal{R}_{n-1}\left(\mathcal{F}_1\right)+\mathcal{R}_n\left(\mathcal{F}_2\right)\right)+2M\left(\frac{2}{n}\log\frac{1}{\delta}\right)^{\frac{1}{2}}$, yielding the first two inequalities.

In order to obtain results in terms of the empirical Rademacher complexity, $\mathcal{R}_S$, instead of the (expected) Rademacher complexity, $\mathcal{R}_n$, we need to apply McDiarmid's Inequality a second time. Let $\mathcal{G}$ be a family of measurable functions with $g \in \mathcal{G} : \mathcal{X} \to [-M, M]$, Let $\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{G}\right) = n^{-1}\sum_{i=1}^n \mathcal{R}_{S_{-i}}\left(\mathcal{G}\right)$, so that $\mathbb{E}_S\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{G}\right) = \mathcal{R}_{n-1}\left(\mathcal{G}\right)$. Then,

$$
\begin{aligned}
&\left|\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{G}\right)-\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{G}\right)\right| \\
&\quad= \left|n^{-1}\sum_{k=1}^n \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n-1}\sum_{j=1,j\neq k}^n \epsilon_j g\left(x_{i,j}\right) - n^{-1}\sum_{k=1}^n \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n-1}\sum_{j=1,j\neq k}^n \epsilon_j g\left(x_j\right)\right| \\
&\quad= \frac{1}{n(n-1)}\left|\sum_{k=1}^n \mathbb{E}_\epsilon\left(\sup_{g\in\mathcal{G}}\sum_{j=1,j\neq k}^n \epsilon_j g\left(x_{i,j}\right) - \sup_{g\in\mathcal{G}}\sum_{j=1,j\neq k}^n \epsilon_j g\left(x_j\right)\right)\right| \\
&\quad\leq \frac{1}{n(n-1)}\sum_{k=1}^n \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}}\sum_{j=1,j\neq k}^n \left|\epsilon_j\left(g\left(x_{i,j}\right)-g\left(x_j\right)\right)\right| \\
&\quad= \frac{1}{n(n-1)}\sum_{k=1}^n \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}}\left|g\left(x_i'\right)-g\left(x_i\right)\right| I(k\neq i) \\
&\quad\leq \frac{1}{n(n-1)}\sum_{k=1}^n I(k\neq i)\mathbb{E}_\epsilon 2M = \frac{1}{n(n-1)}\sum_{k=1,k\neq i}^n 2M \\
&\quad= 2Mn^{-1}
\end{aligned}
$$

where, in the fifth line, we note that the term inside the absolute value is potentially nonzero if and only if it involves x_i', which occurs in each sum over j exactly once, except in the case in which $k \neq i$, in which case it does not occur at all.

$$
\begin{aligned}
\left|\mathcal{R}_{S_i}\left(\mathcal{G}\right)-\mathcal{R}_S\left(\mathcal{G}\right)\right| &= \left|\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} n^{-1}\sum_{j=1}^n \epsilon_j g\left(x_{i,j}\right) - \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} n^{-1}\sum_{j=1}^n \epsilon_j g\left(x_j\right)\right| \\
&\leq n^{-1}\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}}\sum_{j=1}^n \left|\epsilon_j\left(g\left(x_{i,j}\right)-g\left(x_j\right)\right)\right| = n^{-1}\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}}\left|g\left(x_i'\right)-g\left(x_i\right)\right| \\
&\leq n^{-1}\mathbb{E}_\epsilon 2M \\
&= 2Mn^{-1}
\end{aligned}
$$

Combining the above results gives,

$$
\begin{aligned}
&\left|\left(\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{F}_1\right)+\mathcal{R}_{S_i}\left(\mathcal{F}_2\right)\right)-\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)\right)\right| \\
&\quad= \left|\left(\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{F}_1\right)-\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)\right)+\left(\mathcal{R}_{S_i}\left(\mathcal{F}_2\right)-\mathcal{R}_S\left(\mathcal{F}_2\right)\right)\right| \\
&\quad\leq \left|\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{F}_1\right)-\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)\right| + \left|\mathcal{R}_{S_i}\left(\mathcal{F}_2\right)-\mathcal{R}_S\left(\mathcal{F}_2\right)\right| = 2Mn^{-1}+2Mn^{-1} \\
&\quad= 4Mn^{-1}
\end{aligned}
$$

This is the same value we obtained previously, so that the corresponding $\epsilon = 2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$. Then, McDiarmid's Inequality gives,

$$\mathrm{P}\left[\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)\right)-\left(\mathcal{R}_{n-1}\left(\mathcal{F}_1\right)+\mathcal{R}_n\left(\mathcal{F}_2\right)\right)\geq\epsilon\right]\leq\frac{\delta}{2}$$
$$\mathrm{P}\left[\left(\mathcal{R}_{n-1}\left(\mathcal{F}_1\right)+\mathcal{R}_n\left(\mathcal{F}_2\right)\right)-\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)\right)\geq\epsilon\right]\leq\frac{\delta}{2}$$

so that, with probability $1-\frac{\delta}{2}$, $\mathcal{R}_{n-1}\left(\mathcal{F}_1\right)+\mathcal{R}_n\left(\mathcal{F}_2\right)\leq\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)+2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$. Combining this with the previous results shows that, with probability at least $1-\delta$, $\mathbb{E}f\leq\hat{\mathbb{E}}_{S,\lambda}f+2\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)\right)+\left(2M+2\cdot 2M\right)\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}=\hat{\mathbb{E}}_{S,\lambda}f+2\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)\right)+6M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$ and $\hat{\mathbb{E}}_{S,\lambda}f\leq\mathbb{E}f+\lambda M_\Lambda+2\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_1\right)+\mathcal{R}_S\left(\mathcal{F}_2\right)\right)+6M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$, yielding the second pair of inequalities.

Finally, if the elements of $\mathcal{F}$ are symmetric, so that $\forall_{f\in\mathcal{F},x,x'\in\mathcal{X}}f\left(x,x'\right)=f\left(x',x\right)$, if $g\in\mathcal{F}_1$ then $\exists_{f\in\mathcal{F},x'\in\mathcal{X}}\forall_{x\in\mathcal{X}}\left(g(x)=f\left(x',x\right)=f\left(x,x'\right)\right)$, so $g\in\mathcal{F}_2$ as well and $\mathcal{F}_1\subseteq\mathcal{F}_2$. Likewise, if $g\in\mathcal{F}_2$ then $\exists_{f\in\mathcal{F},x'\in\mathcal{X}}\forall_{x\in\mathcal{X}}\left(g(x)=f\left(x,x'\right)=f\left(x',x\right)\right)$, so $g\in\mathcal{F}_1$ as well and $\mathcal{F}_2\subseteq\mathcal{F}_1$. Thus, $\mathcal{F}_1=\mathcal{F}_2$.

□

Corollary 4. *The inequalities in Lemma 3 can be strengthened to the following:*

$$\begin{aligned}\mathbb{E}f\leq&\frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}f\left(x_i,x_j\right)+\lambda\Lambda\left[f,\theta_f\right]\\&+2\mathbb{E}_X\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,X}\right)\right)+2M\left(\frac{2}{n}\log\frac{1}{\delta}\right)^{\frac{1}{2}}\\&\frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}f\left(x_i,x_j\right)+\lambda\Lambda\left[f,\theta_f\right]\leq\mathbb{E}f+\lambda M_\Lambda\\&+2\mathbb{E}_X\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,X}\right)\right)+2M\left(\frac{2}{n}\log\frac{1}{\delta}\right)^{\frac{1}{2}}\end{aligned}$$

$$\begin{aligned}\mathbb{E}f\leq&\frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}f\left(x_i,x_j\right)+\lambda\Lambda\left[f,\theta_f\right]\\&+2\mathbb{E}_X\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_S\left(\mathcal{F}_{2,X}\right)\right)+6M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}\\&\frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}f\left(x_i,x_j\right)+\lambda\Lambda\left[f,\theta_f\right]\leq\mathbb{E}f+\lambda M_\Lambda\\&+2\mathbb{E}_X\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_S\left(\mathcal{F}_{2,X}\right)\right)+6M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}\end{aligned}$$

$\mathcal{F}_{1,x}=\{g_x\mid\exists_{f\in\mathcal{F}}\forall_{x'\in\mathcal{X}}\left[g_x\left(x'\right)=f\left(x,x'\right)\right]\}$, $\mathcal{F}_{2,x}=\{g_x\mid\exists_{f\in\mathcal{F}}\forall_{x'\in\mathcal{X}}\left[g_x\left(x'\right)=f\left(x',x\right)\right]\}$
If the elements of $\mathcal{F}$ are symmetric, then $\mathcal{F}_{1,x}=\mathcal{F}_{2,x}$.

Proof. This follows directly from the proof of Lemma 3 in which we show that
$\mathbb{E}_S\phi^{(1)}(S)\leq 2\mathbb{E}_X\mathbb{E}_{S_{-n},\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n-1}\sum_{i=1}^{n-1}\epsilon_i f\left(X,x_i\right)=2\mathbb{E}_X\mathcal{R}_{n-1}\left(\mathcal{F}_{1,X}\right)$
and $\mathbb{E}_S\phi^{(2)}(S)\leq 2\mathbb{E}_{S,\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i E_X f\left(x_i,X\right)\leq 2\mathbb{E}_X\mathbb{E}_{S,\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f\left(x_i,X\right)$
$=2\mathbb{E}_X\mathcal{R}_n\left(\mathcal{F}_{2,x}\right)$. Using these sharper bounds in the expressions (obtained from McDiarmid's inequality) in Lemma 3 (and using 2δ in place of δ) yields the first pair of equations.

In order to obtain the second pair of expressions, we again need to apply McDiarmid's inequality. Using results from the proof of lemma 3,

$$\begin{aligned}
&\left|\mathbb{E}_X\left(\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_{S_i}\left(\mathcal{F}_{2,X}\right)\right)-\mathbb{E}_X\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_{S}\left(\mathcal{F}_{2,X}\right)\right)\right| \\
&\qquad=\left|\mathbb{E}_X\left(\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{F}_{1,X}\right)-\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_{1,X}\right)\right)+\mathbb{E}_X\left(\mathcal{R}_{S_i}\left(\mathcal{F}_{2,X}\right)-\mathcal{R}_{S}\left(\mathcal{F}_{2,X}\right)\right)\right| \\
&\qquad\leq \mathbb{E}_X\left|\hat{\mathcal{R}}_{n-1,S_i}\left(\mathcal{F}_{1,X}\right)-\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_{1,X}\right)\right|+\mathbb{E}_X\left|\mathcal{R}_{S_i}\left(\mathcal{F}_{2,X}\right)-\mathcal{R}_{S}\left(\mathcal{F}_{2,X}\right)\right| \\
&\qquad=\mathbb{E}_X 2Mn^{-1}+\mathbb{E}_X 2Mn^{-1}=2Mn^{-1}+2Mn^{-1} \\
&\qquad=4Mn^{-1}
\end{aligned}$$

Since this is the same value of c_i we obtained in lemma 3, McDiarmid's Inequality holds for $\epsilon=2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$, so that, with probability at least $1-\frac{\delta}{2}$, $\mathbb{E}_X\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_{n}\left(\mathcal{F}_{2,X}\right)\right)\leq \mathbb{E}_X\left(\hat{\mathcal{R}}_{n-1,S}\left(\mathcal{F}_{1,X}\right)+\mathcal{R}_{S}\left(\mathcal{F}_{2,X}\right)\right)+2M\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$. Then, the second pair of expressions is obtained by replacing δ by $\frac{\delta}{2}$ in the first pair of expressions and combining the result with the above inequality. Since each of these expressions hold with probability at least $1-\frac{\delta}{2}$, their combination will hold with probability at least $1-2\cdot\frac{\delta}{2}=1-\delta$, as claimed.

If the elements of $\mathcal{F}$ is symmetric, so that $\forall_{f\in\mathcal{F},x,x'\in\mathcal{X}} f\left(x,x'\right)=f\left(x',x\right)$, if $g\in\mathcal{F}_{1,x}$ then $\exists_{f\in\mathcal{F}}\forall_{x'\in\mathcal{X}}\left[g_x\left(x'\right)=f\left(x,x'\right)=f\left(x',x\right)\right]$, so $g\in\mathcal{F}_{2,x}$ as well and $\mathcal{F}_{1,x}\subseteq\mathcal{F}_{2,x}$. Likewise, if $g\in\mathcal{F}_{2,x}$ $\exists_{f\in\mathcal{F}}\forall_{x'\in\mathcal{X}}\left[g_x\left(x'\right)=f\left(x',x\right)=f\left(x,x'\right)\right]$, so $g\in\mathcal{F}_{1,x}$ as well and $\mathcal{F}_{2,x}\subseteq\mathcal{F}_{1,x}$. Thus, $\mathcal{F}_{1,x}=\mathcal{F}_{2,x}$.

□

Lemma 5. *Let* $h\in\mathcal{H}:\mathcal{A}\times\mathcal{W}\times\mathcal{X}\rightarrow[-M,M]$ *such that if* $h\in\mathcal{H}$, $-h\in\mathcal{H}$, $\mathcal{Y}\subseteq[-M,M]$, $k:\left(\mathcal{A}\times\mathcal{Z}\times\mathcal{X}\right)^2\rightarrow[-M_k,M_k]$, $\forall_{a,a'\in\mathcal{A},x,x'\in\mathcal{X},z,z'\in\mathcal{Z}}\left(k\left(\left(a,x,z\right),\left(a',x',z'\right)\right)=k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right)$, *and* $\Xi=\left(A,W,X,Y,Z\right)$. *Additionally, let* $f_{h,k}\left(\left(a,w,x,y,z\right),\left(a',w',x',y',z'\right)\right)=\left(y-h\left(a,w,x\right)\right)\times\left(y'-h\left(a',w',x'\right)\right)k\left(\left(a,x,z\right),\left(a',x',z'\right)\right)$ *Then,*

$$\begin{aligned}
\mathbb{E}_\Xi\mathcal{R}_n\left(\mathcal{F}_{1,\Xi}\right)&=\mathbb{E}_\Xi\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right) \\
&\leq 2M\mathbb{E}_{A,X,Z,S,\epsilon}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\qquad+\left(2\log 2\right)^{\frac{1}{2}}M^2M_kn^{-\frac{1}{2}} \\
&=2M\mathbb{E}_{A,X,Z}\mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)+\left(2\log 2\right)^{\frac{1}{2}}M^2M_kn^{-\frac{1}{2}} \\
&\leq 2M\mathbb{E}_{S,\epsilon}\sup_{h\in\mathcal{H},a\in\mathcal{A},x\in\mathcal{X},z\in\mathcal{Z}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(a,x,z\right)\right) \\
&\qquad+\left(2\log 2\right)^{\frac{1}{2}}M^2M_kn^{-\frac{1}{2}} \\
&=2M\mathcal{R}_n\left(\mathcal{F}'\right)+\left(2\log 2\right)^{\frac{1}{2}}M^2M_kn^{-\frac{1}{2}}
\end{aligned}$$

$$\mathcal{F}'_{a,x,z}=\left\{f_{a,x,z}\mid\exists_{h\in\mathcal{H}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}}f_{a,x,z}\left(a',w',x',z'\right)=h\left(a',w',x'\right)k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right\}$$
$$\mathcal{F}'=\left\{f\mid\exists_{h\in\mathcal{H},a\in\mathcal{A},x\in\mathcal{X},z\in\mathcal{Z}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}}f\left(a',w',x',z'\right)=h\left(a',w',x'\right)k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right\}$$

Proof. Let $\mathcal{F}=\{g\mid\exists_{h\in\mathcal{H}}\left(g=f_{h,k}\right)\}$. Since, for all $h\in\mathcal{H}$, $f_{h,k}$ is manifestly symmetric, we have $\mathcal{F}_1=\mathcal{F}_2$ and $\mathcal{F}_{1,\xi}=\mathcal{F}_{2,\xi}$, where these classes are defined in lemma 3 and corollary 4, respectively. We now compute $\mathbb{E}_\Xi\mathcal{R}_n\left(\mathcal{F}_{1,\Xi}\right)=\mathbb{E}_\Xi\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)$.

$$
\begin{aligned}
\mathbb{E}_{\Xi}\mathcal{R}_n\left(\mathcal{F}_{1,\Xi}\right) &= \mathbb{E}_{\Xi}\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right) = \mathbb{E}_{\Xi}\mathbb{E}_{S,\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f\left(x_i,\Xi\right) \\
&= \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f\left(\left(a_i,w_i,x_i,y_i,z_i\right),\left(A,W,X,Y,Z\right)\right) \\
&= \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(y_i-h\left(a_i,w_i,x_i\right)\right)\left(Y-h\left(A,W,X\right)\right) \\
&\qquad\qquad \times k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&= \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}\left(Y-h\left(A,W,X\right)\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(y_i-h\left(a_i,w_i,x_i\right)\right) \\
&\qquad\qquad \times k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\leq \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}Y\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad + \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}-Y\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad + \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}-h\left(A,W,X\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad + \mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}h\left(A,W,X\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right)
\end{aligned}
$$

We analyze each of these four terms separately.

$$
\begin{aligned}
&\mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}Y\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\qquad = \mathbb{E}_{A,X,Y,Z,S,\epsilon}Y\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\qquad = \mathbb{E}_{A,X,Y,Z,S}Y\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\epsilon}\epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\qquad = \mathbb{E}_{A,X,Y,Z,S}Y\cdot\frac{1}{n}\sum_{i=1}^{n}0\cdot y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) = \mathbb{E}_{Y}Y\cdot 0 \\
&\qquad = 0
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{A,W,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}-Y\cdot\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\qquad \leq \mathbb{E}_{A,X,Y,Z,S,\epsilon}\sup_{h\in\mathcal{H}}\left|Y\right|\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right)\right| \\
&\qquad \leq \mathbb{E}_{A,X,Z,S,\epsilon}\sup_{h\in\mathcal{H}}M\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right)\right| \\
&\qquad = M\mathbb{E}_{A,X,Z,S,\epsilon}\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i h\left(a_i,w_i,x_i\right)k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right)
\end{aligned}
$$

where the final equality is due to the fact that $h \in \mathcal{H}$ if and only if $-h \in \mathcal{H}$, so that the supremum of the sum will be equal to the supremum of its absolute value.

$$
\begin{aligned}
&\mathbb{E}_{A,W,X,Y,Z,S,\epsilon} \sup_{h \in \mathcal{H}} h\left(A,W,X\right) \cdot \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h\left(a_i,w_i,x_i\right) k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad \leq \mathbb{E}_{A,W,X,Z,S,\epsilon} \sup_{h,h' \in \mathcal{H}} h'\left(A,W,X\right) \cdot \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h\left(a_i,w_i,x_i\right) k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad \leq \mathbb{E}_{A,W,X,Z,S,\epsilon} \sup_{h,h' \in \mathcal{H}} \left|h'\left(A,W,X\right)\right| \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h\left(a_i,w_i,x_i\right) k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \right| \\
&\quad \leq \mathbb{E}_{A,X,Z,S,\epsilon} \sup_{h \in \mathcal{H}} M \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h\left(a_i,w_i,x_i\right) k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \right| \\
&\quad = M \mathbb{E}_{A,X,Z,S,\epsilon} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h\left(a_i,w_i,x_i\right) k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right)
\end{aligned}
$$

where the final equality follows, as above, because $h \in \mathcal{H}$ if and only if $-h \in \mathcal{H}$.

$$
\begin{aligned}
&\mathbb{E}_{A,W,X,Y,Z,S,\epsilon} \sup_{h \in \mathcal{H}} -h\left(A,W,X\right) \cdot \frac{1}{n} \sum_{i=1}^{n} \epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad \leq \mathbb{E}_{A,W,X,Z,S,\epsilon} \sup_{h \in \mathcal{H}} \left|h\left(A,W,X\right)\right| \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \right| \\
&\quad \leq \mathbb{E}_{A,X,Z,S,\epsilon} M \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \right| \\
&\quad = M \mathbb{E}_{A,X,Z,S,\epsilon} \sup_{h \in \{-1,1\}} h \cdot \frac{1}{n} \sum_{i=1}^{n} \epsilon_i y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad = M \mathbb{E}_{A,X,Z,S} \mathbb{E}_{\epsilon} \sup_{h \in \{-1,1\}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h y_i k\left(\left(a_i,x_i,z_i\right),\left(A,X,Z\right)\right) \\
&\quad \leq M \cdot \mathbb{E}_{A,X,Z,S} M M_k \left(2 \log 2\right)^{\frac{1}{2}} n^{-\frac{1}{2}} = M \cdot M M_k \left(2 \log 2\right)^{\frac{1}{2}} n^{-\frac{1}{2}} \\
&\quad = \left(2 \log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}}
\end{aligned}
$$

where the final inequality follows from Massart's Finite Lemma using $|yk| \leq M M_k$. Combining these results gives,

$$
\begin{aligned}
\mathbb{E}_{\Xi}\mathcal{R}_n\left(\mathcal{F}_{1,\Xi}\right) &= \mathbb{E}_{\Xi}\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right) \\
&\leq 0 + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&\quad + M\mathbb{E}_{A,X,Z,S,\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h\left(a_i, w_i, x_i\right) k\left(\left(a_i, x_i, z_i\right), \left(A, X, Z\right)\right) \\
&\quad + M\mathbb{E}_{A,X,Z,S,\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h\left(a_i, w_i, x_i\right) k\left(\left(a_i, x_i, z_i\right), \left(A, X, Z\right)\right) \\
&= 2M\mathbb{E}_{A,X,Z,S,\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h\left(a_i, w_i, x_i\right) k\left(\left(a_i, x_i, z_i\right), \left(A, X, Z\right)\right) \\
&\quad + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&= 2M\mathbb{E}_{A,X,Z}\mathbb{E}_{S,\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h\left(a_i, w_i, x_i\right) k\left(\left(a_i, x_i, z_i\right), \left(A, X, Z\right)\right) \\
&\quad + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&= 2M\mathbb{E}_{A,X,Z}\mathbb{E}_{S,\epsilon} \sup_{f_{A,X,Z}\in\mathcal{F}'_{A,X,Z}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_{A,X,Z}\left(a_i, w_i, x_i, z_i\right) \\
&\quad + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&= 2M\mathbb{E}_{A,X,Z}\mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right) + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&\leq 2M\mathbb{E}_{S,\epsilon} \sup_{h\in\mathcal{H}, a\in\mathcal{A}, x\in\mathcal{X}, z\in\mathcal{Z}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h\left(a_i, w_i, x_i\right) k\left(\left(a_i, x_i, z_i\right), \left(a, x, z\right)\right) \\
&\quad + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&= 2M\mathbb{E}_{S,\epsilon} \sup_{f\in\mathcal{F}'} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f\left(a_i, w_i, x_i, z_i\right) + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&= 2M\mathcal{R}_n\left(\mathcal{F}'\right) + \left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{F}'_{a,x,z} &= \left\{ f_{a,x,z} \mid \exists_{h\in\mathcal{H}} \forall_{a'\in\mathcal{A}, x'\in\mathcal{X}, z'\in\mathcal{Z}} f_{a,x,z}\left(a', w', x', z'\right) = h\left(a', w', x'\right) k\left(\left(a', x', z'\right), \left(a, x, z\right)\right) \right\} \\
\mathcal{F}' &= \left\{ f \mid \exists_{h\in\mathcal{H}, a\in\mathcal{A}, x\in\mathcal{X}, z\in\mathcal{Z}} \forall_{a'\in\mathcal{A}, x'\in\mathcal{X}, z'\in\mathcal{Z}} f\left(a', w', x', z'\right) = h\left(a', w', x'\right) k\left(\left(a', x', z'\right), \left(a, x, z\right)\right) \right\}
\end{aligned}
$$

□

Lemma 6. *Let $\mathcal{X}$ be a measurable space, μ be a σ-finite measure on $\mathcal{X}$, $\mathcal{F}_0$ be a collection of μ-measurable functions, with $f \in \mathcal{F}_0 : \mathcal{X} \to \mathbb{R}$, $k : \mathcal{X}^2 \to \mathbb{R}$ be symmetric and measurable with respect to the product measure ($\mu \times \mu$), and $\mathcal{F}$ be the quotient space of $\mathcal{F}_0$ in which functions are identified if they are equal μ-almost everywhere. For $f, g \in \mathcal{F}_0$, define the bilinear form $\langle f, g\rangle_k = \int f(x)k(x,y)g(y)d\mu(x,y)$, where μ is the product measure. Then, $\langle\rangle_k$ is an inner product on $\mathcal{F}$ if and only if k is an Integrally Strictly Positive Definite (ISPD) kernel, so that, for all $f \in \mathcal{F}_0$ such that $f \neq 0$ μ-almost everywhere, $\int f(x)k(x,y)f(y)d\mu(x,y) > 0$. Further, if k is ISPD, then it defines a metric on $\mathcal{F}$ by $d_k(f,g) = \|f-g\|_k = \langle f-g, f-g\rangle_k^{\frac{1}{2}}$.*

Proof. For $f, g \in \mathcal{F}_0$,

$$
\begin{aligned}
\langle cf+g,h\rangle_k &= \int (cf+g)(x)k(x,y)h(y)d\mu(x,y) \\
&= c\int f(x)k(x,y)h(y)d\mu(x,y) + \int g(x)k(x,y)h(y)d\mu(x,y) \\
&= c\langle f,h\rangle_k + \langle g,h\rangle_k
\end{aligned}
$$

$$
\begin{aligned}
\langle f,g\rangle_k &= \int f(x)k(x,y)g(y)d\mu(x,y) = \int g(y)k(y,x)f(x)d\mu(x,y) \\
&= \int g(x)k(x,y)f(y)d\mu(x,y) \\
&= \langle g,f\rangle_k
\end{aligned}
$$

so $\langle\rangle_k$ is a bilinear form, as claimed.

To see that $\langle\rangle_k$ is well defined on $\mathcal{F}$. Note, that, if $f=f'$ a.e., then

$$
\begin{aligned}
|\langle f,g\rangle_k - \langle f',g\rangle_k| &= \left|\int f(x)k(x,y)g(y)d\mu(x,y) - \int f'(x)k(x,y)g(y)d\mu(x,y)\right| \\
&\leq \int |f(x)-f'(x)|\,|k(x,y)|\,|g(y)|\,d\mu(x,y) \\
&= \int |(f-f')(x)|\,|k(x,y)|\,|g(y)|\,d\mu(x,y) \\
&= \int\int |(f-f')(x)|\,|k(x,y)|\,|g(y)|\,d\mu(x)d\mu(y) = \int 0 d\mu(y) \\
&= 0
\end{aligned}
$$

where, in the fourth line, we use Tonelli's Theorem and the fact that $f-f'=0$ a.e., so $|f'-f|\,|k||g|=0$ μ_x-a.e. and, thus, the inner integral is 0. Thus, if $f=f'$ a.e., $\langle f,g\rangle_k = \langle f',g\rangle_k$. By the symmetry of the bilinear form, if $g=g'$ a.e. $\langle f',g\rangle_k = \langle f',g'\rangle_k$, so that, if $f=f', g=g'$ a.e. $\langle f,g\rangle_k = \langle f',g'\rangle_k$, so $\langle\rangle_k$ is well defined on $\mathcal{F}$.

If k is also ISPD, then, for $f\neq 0$ μ-almost everywhere, $\langle f,f\rangle_k = \int f(x)k(x,y)f(y)d\mu(x,y) > 0$, so that, combined with the above results, $\langle\rangle_k$ is an inner product on $\mathcal{F}$. Conversely, if $\langle\rangle_k$ is an inner product, then, for $f\neq 0$ μ-almost everywhere, $\langle f,f\rangle_k > 0$, so k is ISPD, by definition.

Since $\langle\rangle_k$ is an inner product, it defines a norm $\|\|_k$ on $\mathcal{F}$ by $\|f\|_k = \langle f,f\rangle_k^{\frac{1}{2}}$. Let $d_k(f,g) = \|f-g\|_k$. Since $\|\|_k$ is a norm, $d_k(f,g) = \|f-g\|_k = |-1|\|g-f\|_k = \|g-f\|_k = d_k(g,f)$ and, if $f\neq g$ a.e., then, $d_k(f,g) = \|f-g\|_k > 0$, while $d_k(f,f) = \|f-f\|_k = \|0\|_k = 0$. Finally, $d_k(f,h) = \|f-h\|_k = \|f-g+g-h\|_k \leq \|f-g\|_k + \|g-h\|_k = d_k(f,g)+d_k(g,h)$ by the subadditivity of the norm, so that the triangle inequality holds and d_k is a metric on $\mathcal{F}$, as claimed.

□

Theorem 1. *Let $\tilde{h}_k$ minimize $R_k(h)$ and $\hat{h}_{k,U,\lambda,n}$ minimize $\hat{R}_{k,U,\lambda,n}(h)$ for $h\in\mathcal{H}$, $k:(\mathcal{A}\times\mathcal{X}\times\mathcal{Z})^2\to[-M_k,M_k]$, $\Lambda:\mathcal{H}\times\Theta_h\to[-0,M_\lambda]$, 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*

$$
\begin{aligned}
R_k(h) &= \mathbb{E}\left[\left(Y-h\left(A,W,X\right)\right)\left(Y'-h\left(A',W',X'\right)\right)k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)\right] \\
\hat{R}_{k,U,\lambda,n}(h) &= \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}\left[\left(y_i-h\left(a_i,w_i,x_i\right)\right)\left(y_j-h\left(a_j,w_j,x_j\right)\right)\right. \\
&\qquad\qquad \left.\times\ k\left(\left(a_i,x_i,z_i\right),\left(a_j,x_j,z_j\right)\right)\right]+\lambda\Lambda[h,\theta_h]
\end{aligned}
$$

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* $1-\delta$*,*

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

Further, if Assumption 5 holds, so k *is ISPD, then* d_k *is a metric on* $L^2_{\mathcal{AXZ}}$ *and, if the right hand side of the inequality goes to zero as* n *goes to infinity,*
$d_k\left(\mathbb{E}\left[h^*|A,X,Z\right]-\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\right)\xrightarrow{\mathrm{P}}0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\xrightarrow{\mathrm{P}}\mathbb{E}\left[h^*|A,X,Z\right]$ *in* d_k*. Also,* $\left\|\mathbb{E}\left[h^*|A,X,Z\right]-\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\right\|_{\mathrm{P}_{A,X,Z}}\xrightarrow{\mathrm{P}}0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\xrightarrow{\mathrm{P}}\mathbb{E}\left[h^*|A,X,Z\right]$ *in* $L^2\left(\mathrm{P}_{\mathcal{A},\mathcal{X},\mathcal{Z}}\right)$*-norm.*

$$\mathcal{F}'_{a,x,z}=\left\{f_{a,x,z}\mid\exists_{h\in\mathcal{H}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}}f_{a,x,z}\left(a',w',x',z'\right)=h\left(a',w',x'\right)k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right\}$$
$$\mathcal{F}'=\left\{f\mid\exists_{h\in\mathcal{H},a\in\mathcal{A},x\in\mathcal{X},z\in\mathcal{Z}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}}f\left(a',w',x',z'\right)=h\left(a',w',x'\right)k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right\}$$

Proof. Let $\Xi=\{A,W,X,Y,Z\}$. Since $\tilde{h}_k$ minimizes $R_k(h)$ and $\hat{h}_{k,U,\lambda,n}$ minimizes $\hat{R}_{k,U,\lambda,n}(h)$ for $h\in\mathcal{H}$, $\hat{R}_{k,U,\lambda,n}\left(\hat{h}_{k,U,\lambda,n}\right)\leq\hat{R}_{k,U,\lambda,n}\left(\tilde{h}_k\right)$.

Taking $f\left(\left(a,w,x,y,z\right),\left(a',w',x',y',z'\right)\right)=\left(y-h\left(a,w,x\right)\right)\left(y'-h\left(a',w',x'\right)\right)$ $k\left(\left(a,x,z\right),\left(a',x',z'\right)\right)$, noting that $|f|\leq(M+M)^2\cdot M_k=(2M)^2M_k=4M^2M_k$, and applying lemma 3 and corollary 4 to $R_k(h)=\mathbb{E}f$ and $\hat{R}_{k,U,\lambda,n}(h)=\hat{E}_{S,\lambda}f$ tells us that, with probability at least $1-\frac{\delta}{2}$,

$$R_k(h)\leq\hat{R}_{k,U,\lambda,n}(h)+2\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$$
$$\hat{R}_{k,U,\lambda,n}(h)\leq R_k(h)+\lambda M_\lambda+2\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$$

so, with probability at least $1-\delta$,

$$\begin{aligned}
R_k\left(\hat{h}_{k,U,\lambda,n}\right) &\leq\hat{R}_{k,U,\lambda,n}\left(\hat{h}_{k,U,\lambda,n}\right)+2\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}\\
&\leq\hat{R}_{k,U,\lambda,n}\left(\tilde{h}_k\right)+2\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}\\
&\leq R_k\left(\tilde{h}_k\right)+\lambda M_\lambda+4\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}
\end{aligned}$$

Applying lemma 5 yields,

$$
\begin{aligned}
R_k\left(\hat{h}_{k,U,\lambda,n}\right) &\leq R_k\left(\tilde{h}_k\right) + \lambda M_\lambda + 4\left(2M\mathbb{E}_{A,X,Z}\mathcal{R}_{n-1}\left(\mathcal{F}'_{A,X,Z}\right) + 2M\mathbb{E}_{A,X,Z}\mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)\right) \\
&\quad + 4\left((2\log 2)^{\frac{1}{2}} M^2 M_k (n-1)^{-\frac{1}{2}} + (2\log 2)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}}\right) \\
&\quad + 16M^2 M_k \left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq R_k\left(\tilde{h}_k\right) + \lambda M_\lambda + 8M\mathbb{E}_{A,X,Z}\left(\mathcal{R}_{n-1}\left(\mathcal{F}'_{A,X,Z}\right) + \mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)\right) \\
&\quad + 4\left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}}\left(\left(\frac{n}{n-1}\right)^{\frac{1}{2}} + 1\right) + 16M^2 M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq R_k\left(\tilde{h}_k\right) + \lambda M_\lambda + 8M\mathbb{E}_{A,X,Z}\left(\mathcal{R}_{n-1}\left(\mathcal{F}'_{A,X,Z}\right) + \mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)\right) \\
&\quad + 16M^2 M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} + 4\left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \cdot \frac{5}{2} \\
&= R_k\left(\tilde{h}_k\right) + \lambda M_\lambda + 8M\mathbb{E}_{A,X,Z}\left(\mathcal{R}_{n-1}\left(\mathcal{F}'_{A,X,Z}\right) + \mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)\right) \\
&\quad + 16M^2 M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} + 10\left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}} \\
&\leq R_k\left(\tilde{h}_k\right) + \lambda M_\lambda + 8M\left(\mathcal{R}_{n-1}\left(\mathcal{F}'\right) + \mathcal{R}_n\left(\mathcal{F}'\right)\right) + 16M^2 M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\quad + 10\left(2\log 2\right)^{\frac{1}{2}} M^2 M_k n^{-\frac{1}{2}}
\end{aligned}
$$

By assumption, h^* satisfies $\mathbb{E}\left[Y - h^*(A,X,X)|A,X,Z\right] = 0$ $\mathrm{P}_{A,X,Z}$-almost surely, so that,

$$
\begin{aligned}
\mathbb{E}\left[Y - h(A,W,X)|A,X,Z\right] &= \mathbb{E}\left[Y - h^*(A,W,X) + h^*(A,W,X) - h(A,W,X)|A,X,Z\right] \\
&= \mathbb{E}\left[Y - h^*(A,W,X)|A,X,Z\right] + \mathbb{E}\left[h^*(A,W,X) - h(A,W,X)|A,X,Z\right] \\
&= 0 + \mathbb{E}\left[h^*(A,W,X) - h(A,W,X)|A,X,Z\right] \\
&= \mathbb{E}\left[h^*(A,W,X) - h(A,W,X)|A,X,Z\right]
\end{aligned}
$$

$\mathrm{P}_{A,X,Z}$-almost surely. Then,

$$
\begin{aligned}
R_k(h) &= \mathbb{E}\left[\left(Y - h\left(A,W,X\right)\right)\left(Y' - h\left(A',W',X'\right)\right)k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y - h\left(A,W,X\right)\right)\left(Y' - h\left(A',W',X'\right)\right)\right.\right. \\
&\qquad\qquad \left.\left.\times\ k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)|\left(A,X,Z\right),\left(A',X',Z'\right)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y - h\left(A,W,X\right)|A,X,Z\right]\right. \\
&\qquad\qquad \times\ \mathbb{E}\left[Y' - h\left(A',W',X'\right)|A',X',Z'\right] \\
&\qquad\qquad \left.\times\ k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[h^*\left(A,W,X\right) - h\left(A,W,X\right)|A,X,Z\right]\right. \\
&\qquad\qquad \times\ \mathbb{E}\left[h^*\left(A',W',X'\right) - h\left(A',W',X'\right)|A',X',Z'\right] \\
&\qquad\qquad \left.\times\ k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(h^*\left(A,W,X\right) - h\left(A,W,X\right)\right)\left(h^*\left(A',W',X'\right) - h\left(A',W',X'\right)\right)\right.\right. \\
&\qquad\qquad \left.\left.\times\ k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)|\left(A,X,Z\right),\left(A',X',Z'\right)\right]\right] \\
&= \mathbb{E}\left[\left(h^*\left(A,W,X\right) - h\left(A,W,X\right)\right)\left(h^*\left(A',W',X'\right) - h\left(A',W',X'\right)\right)\right. \\
&\qquad\qquad \left.\times\ k\left(\left(A,X,Z\right),\left(A',X',Z'\right)\right)\right] \\
&= d_k^2\left(h^*,h\right)
\end{aligned}
$$

Thus,

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

If the right hand side of this expression goes to zero as n goes to infinity, then, for any $\delta, \epsilon > 0$, we can find n such that the right hand side is less than ϵ. Further, since we can do this for any value of δ, we can choose a sequence of δ_ns decreasing in n, so that $\lim_{n\to\infty}\delta_n = 0$, so that the left hand side converges in probability. If k is ISPD, by Lemma 6, d_k is a metric on $L^2_{\mathcal{AXZ}}$. Thus, $d_k\left(h^*, \hat{h}_{k,U,\lambda,n}\right) \xrightarrow{\mathrm{P}} 0$, so $\mathbb{E}\left[\hat{h}_{k,U,\lambda,n}\middle| A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A, X, Z\right]$, in d_k. Further, the fact that k is Integrally Strictly Positive Definite, implies that $\left\|\mathbb{E}\left[h^*|A, X, Z\right] - \mathbb{E}\left[\hat{h}_{k,U,\lambda,n}\middle| A, X, Z\right]\right\|_{\mathrm{P}_{A,X,Z}} \xrightarrow{\mathrm{P}} 0$, so that $\mathbb{E}\left[\hat{h}_{k,U,\lambda,n}\middle| A, X, Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A, X, Z\right]$, in $L^2\left(\mathrm{P}_{\mathcal{AXZ}}\right)$-norm, as well.

□

Lemma 7. *Let $f : \mathcal{X}^2 \to [-M, M]$, $\forall_{x\in\mathcal{X}} f(x, x) \geq 0$, $\hat{U}_n[f] = \frac{1}{n(n-1)}\sum_{i,j=1,j\neq i}^n f(x_i, x_j)$, and $\hat{V}_n[f] = n^{-2}\sum_{i,j=1}^n f(x_i, x_j)$. Then, $(n-1)\hat{U}_n[f] \leq n\hat{V}_n[f] \leq (n-1)\hat{U}_n[f] + M$.*

Proof.

$$
\begin{aligned}
(n-1)\hat{U}_n[f] &= n^{-1}\sum_{i,j=1,j\neq i}^n f(x_i, x_j) \leq n^{-1}\sum_{i,j=1}^n f(x_i, x_j) \\
&= n\hat{V}_n[f] = n^{-1}\left(\sum_{i,j=1,j\neq i}^n f(x_i, x_j) + \sum_{i,j=1,j=i}^n f(x_i, x_j)\right) \\
&\leq n^{-1}n(n-1)\hat{U}_n[f] + n^{-1}\sum_{i=1}^n f(x_i, x_i) \leq (n-1)\hat{U}_n[f] + M
\end{aligned}
$$

so

$$
(n-1)\hat{U}_n[f] \leq n\hat{V}_n[f] \leq (n-1)\hat{U}_n[f] + M
$$

□

Corollary 8. *Let $\tilde{h}_k$ minimize $R_k(h)$ and $\hat{h}_{k,V,\lambda,n}$ minimize $\hat{R}_{k,V,\lambda,n}(h)$ 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*

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

$$
\begin{aligned}
\hat{R}_{k,V,\lambda,n}(h) = n^{-2}\sum_{i,j=1}^n &\left(y_i - h\left(a_i, w_i, x_i\right)\right)\left(y_j - h\left(a_j, w_j, x_j\right)\right)k\left(\left(a_i, x_i, z_i\right), \left(a_j, x_j, z_j\right)\right) \\
&+ \lambda\Lambda[f, \theta_f]
\end{aligned}
$$

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* $1-\delta$*,*

$$\begin{aligned}
d_k^2\left(h^*,\hat{h}_{k,V,\lambda,n}\right) &\leq d_k^2\left(h^*,\tilde{h}_k\right)+\lambda M_\lambda+8M\mathbb{E}_{A,X,Z}\left(\mathcal{R}_{n-1}\left(\mathcal{F}'_{A,X,Z}\right)+\mathcal{R}_n\left(\mathcal{F}'_{A,X,Z}\right)\right)\\
&\quad+16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}+\left(4M^2M_k+\lambda M_\lambda\right)(n-1)^{-1}\\
&\quad+10\left(2\log 2\right)^{\frac{1}{2}}M^2M_kn^{-\frac{1}{2}}\\
&\leq d_k^2\left(h^*,\tilde{h}_k\right)+\lambda M_\lambda+8M\left(\mathcal{R}_{n-1}\left(\mathcal{F}'\right)+\mathcal{R}_n\left(\mathcal{F}'\right)\right)+16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}\\
&\quad+\left(4M^2M_k+\lambda M_\lambda\right)(n-1)^{-1}+10\left(2\log 2\right)^{\frac{1}{2}}M^2M_kn^{-\frac{1}{2}}
\end{aligned}$$

Further, if Assumption 5 holds, so k *is ISPD, then* d_k *is a metric on* $L^2_{\mathcal{AXZ}}$ *and, if the right hand side of the inequality goes to zero as* n *goes to infinity,*
$d_k\left(\mathbb{E}\left[h^*|A,X,Z\right]-\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\right)\xrightarrow{\mathrm{P}}0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\xrightarrow{\mathrm{P}}\mathbb{E}\left[h^*|A,X,Z\right]$ *in* d_k*. Also,* $\left\|\mathbb{E}\left[h^*|A,X,Z\right]-\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\right\|_{\mathrm{P}_{A,X,Z}}\xrightarrow{\mathrm{P}}0$ *so* $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right]\xrightarrow{\mathrm{P}}\mathbb{E}\left[h^*|A,X,Z\right]$ *in* $L^2\left(\mathrm{P}_{\mathcal{A},\mathcal{X},\mathcal{Z}}\right)$*-norm.*

$\mathcal{F}'_{a,x,z}=\left\{f_{a,x,z}\mid\exists_{h\in\mathcal{H}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}}f_{a,x,z}\left(a',w',x',z'\right)=h\left(a',w',x'\right)k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right\}$
$\mathcal{F}'=\left\{f\mid\exists_{h\in\mathcal{H},a\in\mathcal{A},x\in\mathcal{X},z\in\mathcal{Z}}\forall_{a'\in\mathcal{A},x'\in\mathcal{X},z'\in\mathcal{Z}}f\left(a',w',x',z'\right)=h\left(a',w',x'\right)k\left(\left(a',x',z'\right),\left(a,x,z\right)\right)\right\}$

Proof. Defining f and Ξ as in Theorem 1, then $\hat{R}_{k,U,n}(h)=\frac{1}{n(n-1)}\sum_{i,j=1,j\neq i}^n f_h\left(\xi_i,\xi_j\right)$, $\hat{R}_{k,U,\lambda,n}(h)=\frac{1}{n(n-1)}\sum_{i,j=1,j\neq i}^n f_h\left(\xi_i,\xi_j\right)+\lambda\Lambda[f,\theta_f]$, $\hat{R}_{k,V,n}(h)=n^{-2}\sum_{i,j=1}^n f_h\left(\xi_i,\xi_j\right)$, $\hat{R}_{k,V,\lambda,n}(h)=n^{-2}\sum_{i,j=1}^n f_h\left(\xi_i,\xi_j\right)+\lambda\Lambda[f,\theta_f]$, Noting that $|f|\leq 4M^2M_k$ and applying Lemma 7 to $\hat{R}_{k,U,\lambda,n}$ and $\hat{R}_{k,V,\lambda,n}$ yields,

$$\begin{aligned}
\hat{R}_{k,\lambda,n}(h)=\hat{R}_{k,U,n}(h)+\lambda\Lambda[f,\theta_f]&\leq\frac{n}{n-1}\hat{R}_{k,V,n}(h)+\lambda\Lambda[f,\theta_f]\\
&\leq\frac{n}{n-1}\hat{R}_{k,V,n}(h)+\frac{n}{n-1}\lambda\Lambda[f,\theta_f]=\frac{n}{n-1}\hat{R}_{k,V,\lambda,n}(h)\\
&\leq\hat{R}_{k,U,n}(h)+(n-1)^{-1}4M^2M_k+\frac{n}{n-1}\lambda\Lambda[f,\theta_f]
\end{aligned}$$

From the proof of Theorem 1, with probability at least $1-\frac{\delta}{2}$, we have,

$$R_k(h)\leq\hat{R}_{k,\lambda,n}(h)+2\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$$
$$\hat{R}_{k,\lambda,n}(h)\leq R_k(h)+\lambda M^2+2\mathbb{E}_\Xi\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right)+\mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right)+8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}$$

Recalling that $\hat{h}_{k,V,\lambda,n}$ minimizes $\hat{R}_{k,V,\lambda,n}(h)$ and $\tilde{h}_k$ minimizes $R_k(h)$ over $\mathcal{H}$, so that $\hat{R}_{k,V,\lambda,n}\left(\hat{h}_{k,V,\lambda,n}\right)\leq\hat{R}_{k,V,\lambda,n}\left(\tilde{h}_k\right)$ and combining the above expressions gives,

$$
\begin{aligned}
R_k\left(\hat{h}_{k,V,\lambda,n}\right) &\leq \hat{R}_{k,U,\lambda,n}\left(\hat{h}_{k,V,\lambda,n}\right) + 2\mathbb{E}_{\Xi}\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right) + \mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right) + 8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq \frac{n}{n-1}\hat{R}_{k,V,\lambda,n}\left(\hat{h}_{k,V,\lambda,n}\right) + 2\mathbb{E}_{\Xi}\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right) + \mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right) + 8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq \frac{n}{n-1}\hat{R}_{k,V,\lambda,n}\left(\tilde{h}_k\right) + 2\mathbb{E}_{\Xi}\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right) + \mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right) + 8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq \hat{R}_{k,n}\left(\tilde{h}_k\right) + (n-1)^{-1}4M^2M_k + \frac{n}{n-1}\lambda\Lambda[f,\theta_f] + 2\mathbb{E}_{\Xi}\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right) + \mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right) \\
&\quad + 8M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq R_k\left(\tilde{h}_k\right) + (n-1)^{-1}4M^2M_k + \frac{n}{n-1}\lambda M_\lambda + 4\mathbb{E}_{\Xi}\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right) + \mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right) \\
&\quad + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\leq R_k\left(\tilde{h}_k\right) + \left(4M^2M_k + \lambda M_\lambda\right)(n-1)^{-1} + \lambda M_\lambda + 4\mathbb{E}_{\Xi}\left(\mathcal{R}_{n-1}\left(\mathcal{F}_{1,\Xi}\right) + \mathcal{R}_n\left(\mathcal{F}_{2,\Xi}\right)\right) \\
&\quad + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}}
\end{aligned}
$$

Using Lemma 5 and results from the proof of Theorem 1,

$$
\begin{aligned}
R_k\left(\hat{h}_{k,V,\lambda,n}\right) &\leq R_k\left(\tilde{h}_k\right) + \left(4M^2M_k + \lambda M_\lambda\right)(n-1)^{-1} + \lambda M_\lambda \\
&\quad + 8M\left(\mathcal{R}_{n-1}\left(\mathcal{F}'\right) + \mathcal{R}_n\left(\mathcal{F}'\right)\right) + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\quad + 10\left(2\log 2\right)^{\frac{1}{2}} M^2M_kn^{-\frac{1}{2}} \\
&= R_k\left(\tilde{h}_k\right) + \lambda M_\lambda + 8M\left(\mathcal{R}_{n-1}\left(\mathcal{F}'\right) + \mathcal{R}_n\left(\mathcal{F}'\right)\right) + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\quad + \left(4M^2M_k + \lambda M_\lambda\right)(n-1)^{-1} + 10\left(2\log 2\right)^{\frac{1}{2}} M^2M_kn^{-\frac{1}{2}}
\end{aligned}
$$

Recalling that $d_k^2\left(h^*, h\right) = R_k(h)$, we have,

$$
\begin{aligned}
d_k^2\left(h^*, \hat{h}_{k,V,\lambda,n}\right) &\leq d_k^2\left(h^*, \tilde{h}_k\right) + \lambda M^2 + 8M\left(\mathcal{R}_{n-1}\left(\mathcal{F}'\right) + \mathcal{R}_n\left(\mathcal{F}'\right)\right) + 16M^2M_k\left(\frac{2}{n}\log\frac{2}{\delta}\right)^{\frac{1}{2}} \\
&\quad + \left(4M^2M_k + \lambda M_\lambda\right)(n-1)^{-1} + 10\left(2\log 2\right)^{\frac{1}{2}} M^2M_kn^{-\frac{1}{2}}
\end{aligned}
$$

□

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

Proof. Let $h^*, h^{*\prime}$ both be zeros of $\mathbb{E}\left[Y - h\left(A,W,X\right)|A,X,Z\right]$, $\mathrm{P}_{A,X,Z}$-almost surely, then

$$\begin{aligned}
&\mathbb{E}\left[\left(h^* - h^{*\prime}\right)(A,W,X)|A,X,Z\right] = \mathbb{E}\left[\left(h^* - Y + Y - h^{*\prime}\right)(A,W,X)|A,X,Z\right] \\
&\quad = \mathbb{E}\left[Y - h^{*\prime}(A,W,X)|A,X,Z\right] - \mathbb{E}\left[Y - h^*(A,W,X)|A,X,Z\right] = 0 - 0 \\
&\quad = 0
\end{aligned}$$

$\mathrm{P}_{A,X,Z}$-almost surely, so $h^* = h^{*\prime}$ $\mathrm{P}_{A,W,X}$-almost surely.

If $\mathbb{E}\left[\hat{h}_{k,\lambda,n}\middle|A,X,Z\right] \xrightarrow{\mathrm{P}} \mathbb{E}\left[h^*|A,X,Z\right]$, $\left\|\mathbb{E}\left[\left(\hat{h}_n - h^*\right)(A,W,X)\middle|A,X,Z\right]\right\|_{\mathrm{P}_{A,X,Z}} \xrightarrow{\mathrm{P}} 0$, meaning that the convergence occurs $\mathrm{P}_{A,X,Z}$-almost surely, so, by assumption, $\left\|\hat{h}_n - h^*\right\|_{\mathrm{P}_{A,W,X}} \xrightarrow{\mathrm{P}} 0$, so $\hat{h}_n \xrightarrow{\mathrm{P}} h^*$.

□

B Hyperparameter Tuning & Model Architecture

We tuned the architectures of the Naive Net and NMMR models on both the Demand and dSprite experiments. The Naive Net used MSE loss to estimate Y^a, while NMMR relied on either the U-statistic or V-statistic.

Within each experiment, the Naive Net and NMMR models used similar architectures. In the Demand experiment, both models consisted of 2-5 ("Network depth" in Table S1) fully connected layers with a variable number ("Network width") of hidden units.

In the dSprite experiment, each model had two VGG-like heads [28] that took in A and W images and applied two blocks of {`Conv2D`, `Conv2D`, `MaxPool2D`} with 3 by 3 kernels. Each `Conv2D` layer had 64 filters in the first block, then 128 filters in the second block. The output of the second block was flattened, then projected to 256 dimensions. Two subsequent fully connected layers were used, with their number of units determined by the "layer width decay" factor in Table S1. For example, if this factor was 0.5, then the two layers would have 128 and 64 units, respectively.

We performed a grid search over the following parameters:

Table S1: Grid of hyperparameters for our naive neural network and NMMR models.

Hyperparameter	Demand	dSprite
Learning rate	{3e-3, 3e-4, 3e-5}	{3e-4, 3e-5, 3e-6}
L2 penalty	{3e-5, 3e-6, 3e-7}	{3e-6, 3e-7}
# of epochs	3000	500
Batch size	1000	256
Layer width decay		{0.25, 0.5}
Network width	{10, 40, 80}	
Network depth	{2, 3, 4, 5}	

We selected the final hyperparameters by considering the lowest average U-statistic or V-statistic on held-out validation sets for NMMR or the MSE for the Naive Net. For the Demand experiment, we repeated this process 10 times with different random seeds and averaged the statistics. For the dSprite experiment, we had 3 repetitions.

Full hyperparameter choices for all methods used in this work are available in our code. The hyperparameters selected for NMMR were tuned for each dataset:

Table S2: Optimal hyperparameters for NMMR methods

Hyperparameter	NMMR-U Demand	NMMR-U dSprite	NMMR-V Demand	NMMR-V dSprite
Learning rate	3e-3	3e-5	3e-3	3e-5
L2 penalty	3e-6	3e-6	3e-6	3e-7
# of epochs	3,000	500	3,000	500
Batch size	1,000	256	1,000	256
Layer width decay	—	0.25	—	0.5
Network width	80	—	80	—
Network depth	4	—	3	—

The hyperparameters for Naive Net for each dataset were:

Table S3: Optimal hyperparameters for Naive Net model

Hyperparameter	Naive Net Demand	Naive Net dSprite
Learning rate	3e-3	3e-5
L2 penalty	3e-6	3e-6
# of epochs	3,000	500
Batch size	1,000	256
Layer width decay	—	0.25
Network width	80	—
Network depth	2	—

Another hyperparameter of note is the choice of kernel in the loss function of NMMR. Throughout, we relied on the RBF kernel:

$$k(x_i, x_j) = e^{\frac{-||x_i - x_j||_2^2}{2\sigma^2}}$$

with $\sigma = 1$. For future work, we could consider other choices of the kernel or tune the length scale parameter σ. In the dSprite experiment, the kernel function is applied to pairs of Z and A data. Since A is an 64×64 image, we chose to concatenate $(z_i, 0.05a_i)$ as input to the kernel for the i-th data point. One could also consider tuning this multiplicative factor, but in practice we found that it allowed for both Z and A to impact the result of the kernel function.

C Experiment Details

C.1 Demand Data Generating Process

The Demand experiment has the following structural equations:

- Demand: $U \sim \mathcal{U}(0, 10)$
- Fuel cost: $[Z_1, Z_2] = [2\sin(2\pi U/10) + \epsilon_1, 2\cos(2\pi U/10) + \epsilon_2]$
- Web page views: $W = 7g(U) + 45 + \epsilon_3$
- Price: $A = 35 + (Z_1 + 3)g(U) + Z_2 + \epsilon_4$
- Sales: $Y = A \times \min(\exp(\frac{W-A}{10}), 5) - 5g(U) + \epsilon_5$
- where $g(u) = 2\left(\frac{(u-5)^4}{600} + e^{-4(u-5)^2} + \frac{u}{10} - 2\right)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$

C.2 Demand causal DAG

C.3 Demand Exploratory Data Analysis

In Figure S2, Panels A and B show that W is a more informative proxy for U than Z, although neither relationship is one-to-one. Panel C shows that the true potential outcome curve, denoted by the black curve $a \mapsto E[Y^a]$, deviates from the observed (A, Y) distribution due to confounding. In

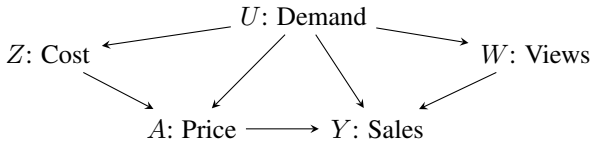


Figure S1: Causal DAG for the Demand experiment.

particular, the largest deviation occurs at smaller values of A. The goal of each method is to recover this average potential outcome curve given data on A, Z, W, and Y.

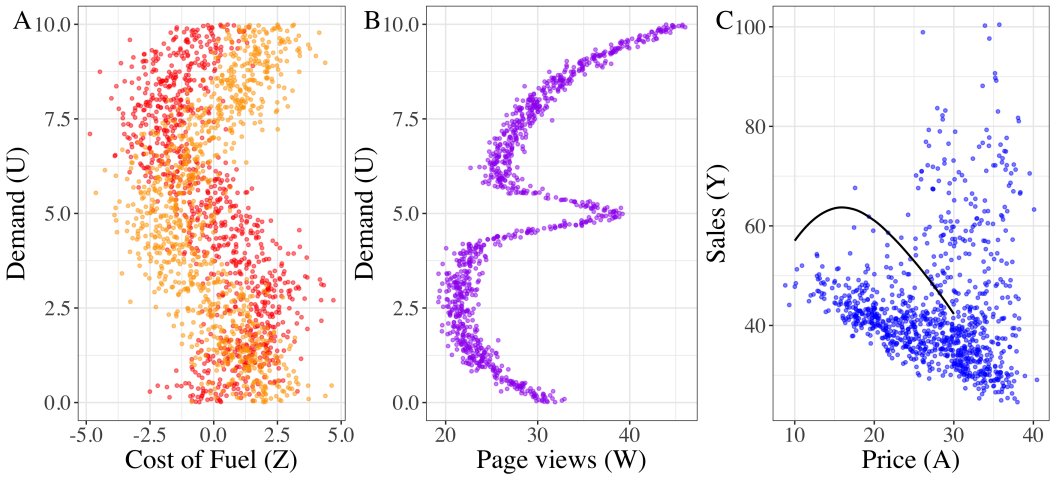


Figure S2: Views of the A, Z, W, Y, U relationships. (A) Z_1 (red) and Z_2 (orange) have sinusoidal relationships with U, (B) W has a far less noisy relationship with U, and (C) the observed distribution (A, Y) (blue) deviates from the true average potential outcome curve (black).

C.4 Demand Boxplot Statistics

Table S4 contains the median and interquartile ranges (in parentheses) of the c-MSE values compiled in the boxplots shown in Figure 2. NMMR demonstrated state of the art performance on the Demand benchmark. We also extended the benchmark to include training set sizes of 10,000 and 50,000 data points, whereas Xu et al. [7] originally included 1,000 and 5,000 data points. We observed that NMMR-U had strong performance across all dataset sizes. PMMR and KPV were unable to run on 50,000 training points due to computational limits on their kernel methods, while DFPV exhibited a large increase in c-MSE as training set size increased.

Table S4: Demand Boxplot Median & (IQR) values

	Training Set Size			
Method	1,000	5,000	10,000	50,000
PMMR	587.51 (40.35)	466.5 (33.47)	423.1 (29.26)	—
KPV	469.94 (97.07)	481.32 (54.8)	470.62 (29.3)	—
Naive Net	160.35 (33.78)	186.97 (30.22)	204.36 (113.71))	224.09 (33.17)
CEVAE	180.8 (161.26)	214.98 (120.88)	170.58 (176.1)	171.98 (293.27)
2SLS	82.08 (18.82)	83.16 (4.51)	82.1 (5.55)	82.01 (2.22)
DFPV	41.83 (11.78)	48.22 (7.73)	87.14 (471.59)	242.15 (464.38)
LS	63.19 (5.82)	65.14 (2.64)	64.98 (2.44)	64.65 (0.74)
NMMR-U (ours)	23.68 (8.02)	**16.21 (10.55)**	**14.25 (4.46)**	**14.27 (12.47)**
NMMR-V (ours)	**23.41 (11.26)**	30.74 (17.73)	42.88 (29.45)	62.18 (16.97)

C.5 Demand Prediction Curves

Figures S3-S6 provide the individual predicted potential outcome curves of each method in the Demand experiment. While Figure 2 provides a summary of the c-MSE, this does not give a picture of the model's actual estimate of $\mathbb{E}[Y^a]$. These figures give an insight into ranges of A for which each model provides particularly accurate or inaccurate estimates of the potential outcomes. Across all training set sizes, methods are most accurate in the region where the training observations of Y are closest to the ground truth (see Figure S2).

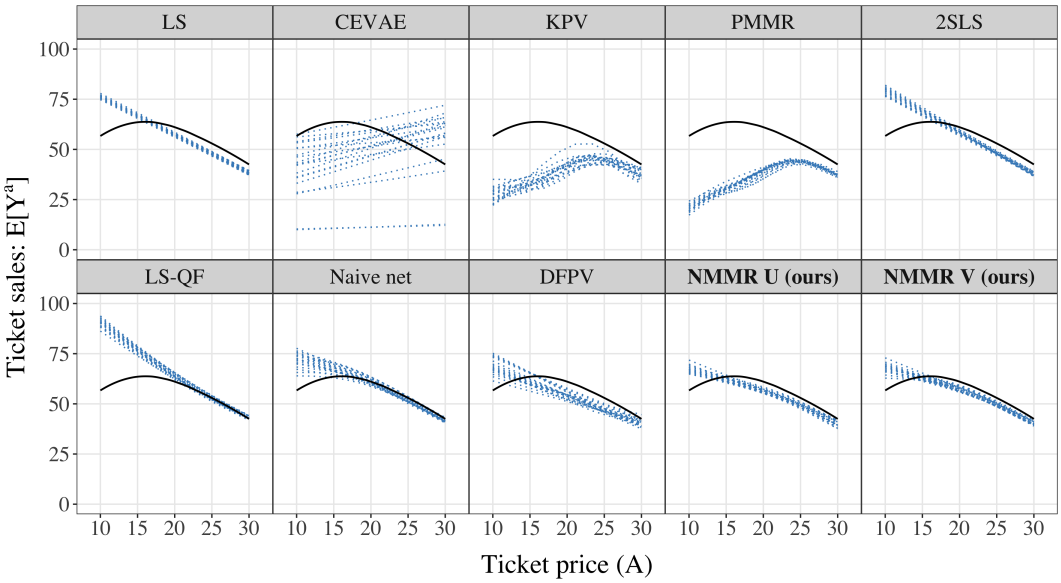


Figure S3: Demand experiment with 1,000 training data points with the true average potential outcome curves (black) and each method's predicted potential outcome curves (blue). Each method was replicated 20 times, generating one predicted curve per replicate. Note that with only a limited amount of data, most methods only recover the true curve in the later half of the range of A. See Figure S2.

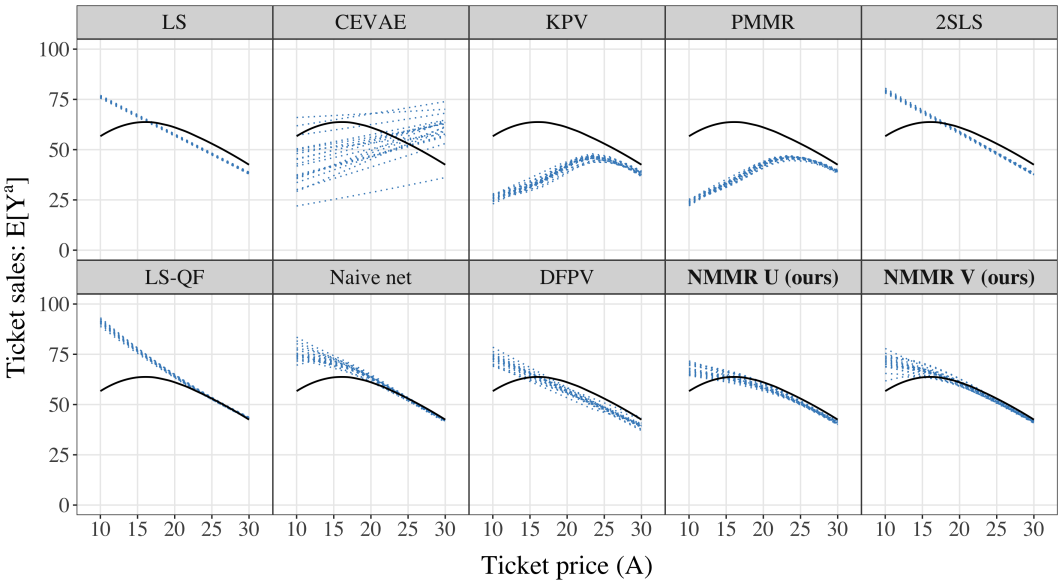


Figure S4: Demand experiment with 5,000 training data points with the true average potential outcome curves (black) and each method's predicted potential outcome curves (blue). Each method was replicated 20 times, generating one predicted curve per replicate. Note that now, NMMR begins to adjust in the range of $A \in [10, 20]$ and bend down towards the true curve. NMMR is empirically accounting for the unobserved confounder U through the proxy variables.

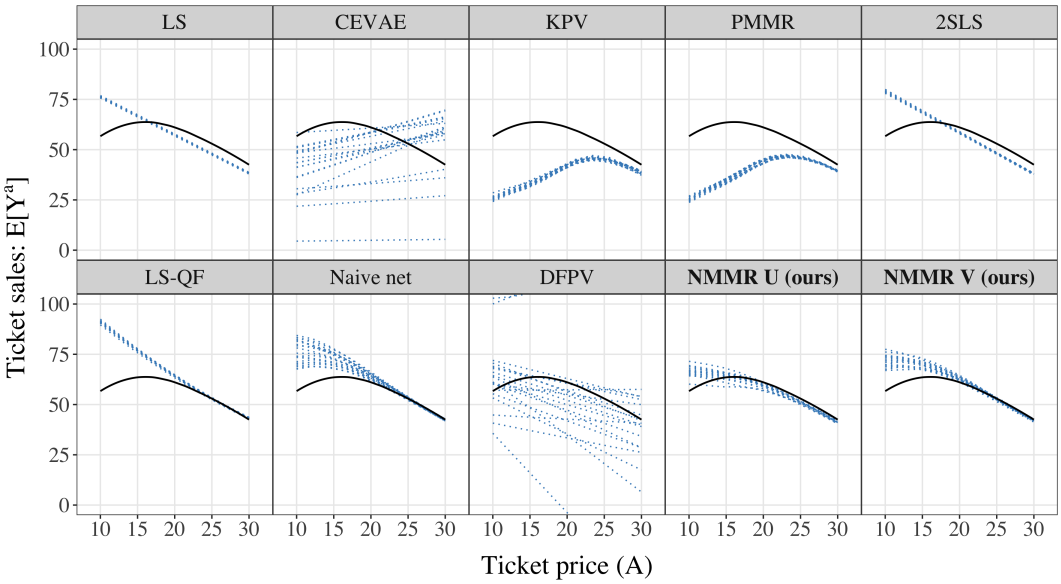


Figure S5: Demand experiment with 10,000 training data points with the true average potential outcome curves (black) and each method's predicted potential outcome curves (blue). Each method was replicated 20 times, generating one predicted curve per replicate. We observed some additional curvature to NMMR prediction curves

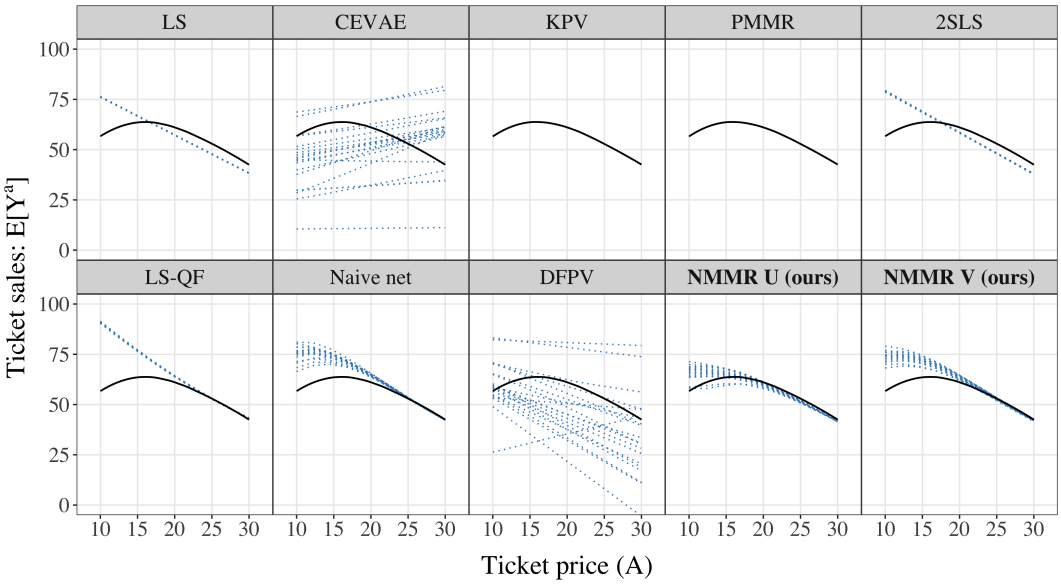


Figure S6: Demand experiment with 50,000 training data points with the true average potential outcome curves (black) and each method's predicted potential outcome curves (blue). Each method was replicated 20 times, generating one predicted curve per replicate. KPV and PMMR timed out due to computational requirements of their kernel methods.

C.6 dSprite Data Generating Process

The dSprite experiment has a unique data generating mechanism, given that the images A and W are queried from an existing dataset rather than generated on the fly. The dataset is indexed by parameters: shape, color, scale, rotation, posX, and posY. As mentioned in the paper, this experiment fixes shape = heart, color = white. Therefore, to simulate data from this system, we follow the steps:

1. Simulate values for scale, rotation, posX, posY [†].
2. Set $U =$ posY.
3. Set $Z =$ (scale, rotation, posX).
4. Set A equal to the dSprite image with the corresponding (scale, rotation, posX and posY) as found in Z and U, then add $\mathcal{N}(0, 0.1)$ noise to each pixel.
5. Set W equal to the dSprite image with (scale=0.8, rotation=0, posX=0.5) and posY from U, then add $\mathcal{N}(0, 0.1)$ noise to each pixel.
6. Compute $Y = \frac{\frac{1}{10}||vec(A)^T B||_2^2 - 5000}{1000} \times \frac{(31 \times \mathrm{U} - 15.5)^2}{85.25} + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.5)$

[†] Let $\mathcal{DU}(a, b)$ denote a Discrete Uniform distribution from a to b. Scale is a Discrete Uniform random variable taking values [0.5, 0.6, 0.7, 0.8, 0.9, 1.0] with equal probability. Rotation $\sim \mathcal{DU}(0, 2\pi)$ with 40 equally-spaced values. And both posX, posY $\sim \mathcal{DU}(0, 1)$ with 32 equally-spaced values.

C.7 dSprite causal DAG

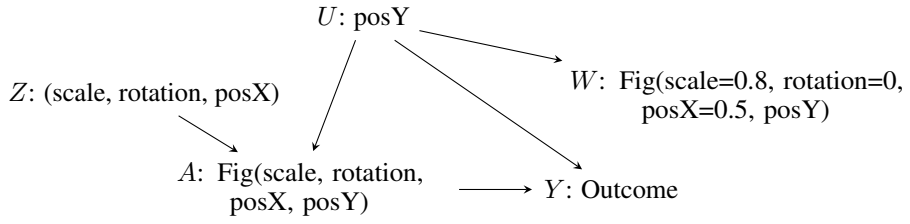
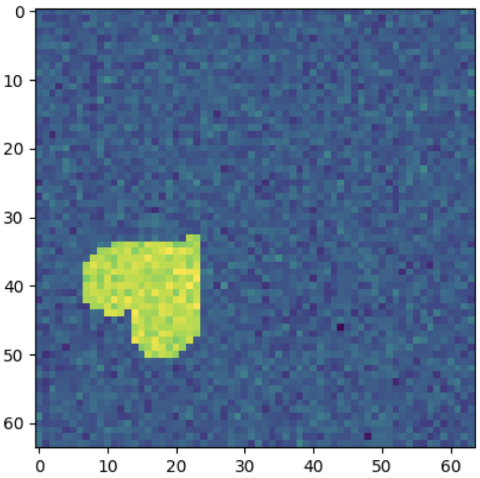
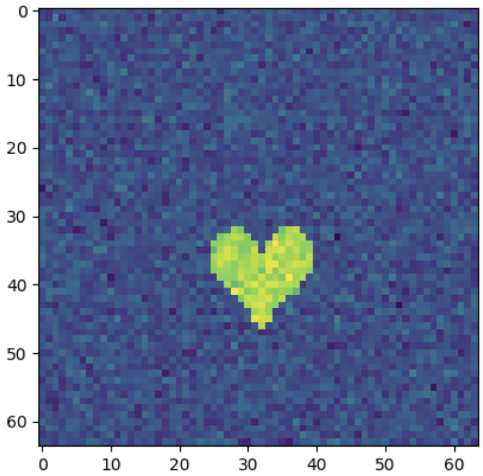


Figure S7: DAG for the dSprite experiment

C.8 dSprite Exemplar A and W



(a) Example of A in dSprite corresponding (scale, rotation, posX and posY) determined from Z and U.



(b) Example of W in dSprite, which is always a centered, vertical heart with a fixed scale. The only thing that changes is posY, which is determined by U.

Figure S8: Examples of the image based treatment A and outcome-inducing proxy W in the dSprite experiment. Previous proximal inference methods did not take advantage of the inductive bias of image convolutions, which NMMR naturally incorporates into its neural network architecture for the dSprite benchmark.

C.9 dSprite Test Set

The dSprite test set consists of 588 images A spanning the following grid of parameters:

- posX $\in [0, \frac{5}{31}, \frac{10}{31}, \frac{15}{31}, \frac{20}{31}, \frac{25}{31}, \frac{30}{31}]$
- posY $\in [0, \frac{5}{31}, \frac{10}{31}, \frac{15}{31}, \frac{20}{31}, \frac{25}{31}, \frac{30}{31}]$
- scale $\in [0.5, 0.8, 1.0]$
- rotation $\in [0, 0.5\pi, \pi, 1.5\pi]$

The labels for each test image A are computed as $\mathbb{E}[Y^a] = \frac{\frac{1}{10}||vec(a)^T B||_2^2 - 5000}{1000}$.

C.10 dSprite Boxplot Statistics

Table S5 contains the median and interquartile ranges (in parentheses) of the c-MSE values compiled in the boxplots shown in Figure 3. NMMR demonstrated state of the art performance on the dSprite benchmark. We also extended the benchmark to include training set sizes of 7,500 data points, whereas Xu et al. [7] originally included 1,000 and 5,000 data points. We observed that NMMR-V had strong performance across all dataset sizes and particularly excelled when training data increased. Most other methods remained relatively consistent as the amount of data increased.

Table S5: dSprite Boxplot Median & (IQR) values

	Training Set Size		
Method	1,000	5,000	7,500
PMMR	17.7 (0.78)	17.74 (0.64)	16.2 (0.48)
KPV	23.4 (1.37)	16.58 (0.93)	14.46 (1.01)
Naive Net	32.25 (9.72)	34.24 (11.95)	34.76 (11.95)
CEVAE	26.34 (0.82)	26.16 (1.51)	25.77 (1.45)
DFPV	10.02 (2.95)	8.81 (2.04)	8.52 (1.06)
NMMR-U (ours)	**4.72 (1.1)**	7.1 (2.74)	7.52 (2.05)
NMMR-V (ours)	11.8 (1.88)	**1.82 (0.67)**	**1.53 (0.68)**

C.11 dSprite DFPV vs. NMMR Evaluation

In order to assess whether our improved performance on dSprite was due to the fact that NMMR leveraged convolutional neural networks while DFPV relied on multi-layer perceptrons with a spectral-norm regularization. We modified DFPV to include the same VGG-like heads mentioned in Appendix B. We performed a grid search over the same learning rates and L2 penalties in Table S1. Figure S9 shows that DFPV with CNNs actually performed slightly worse than the original, published-version of DFPV. We report several different results for DFPV CNN since the results were so close after cross validation. Figure S9 shows results for only the 1,000 data point evaluation – DFPV had trouble scaling in practice as dataset size increased.

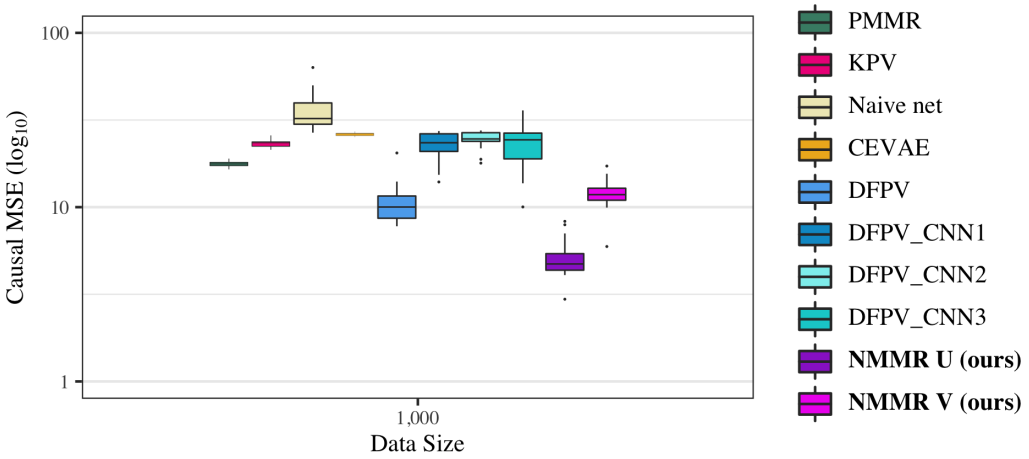


Figure S9: Performance of DFPV with CNNs compared to other evaluated methods on the dSprite benchmark. DFPV with CNNs performed worse compared to the published version of DFPV across a variety of hyperparameters

Model	Learning Rate	Weight Decay
DFPV_CNN1	3e-6	3e-6
DFPV_CNN2	3e-6	3e-7
DFPV_CNN3	3e-5	3e-7

Table S6: Hyperparameters for reported DFPV_CNN models

D Batched Loss Function

When computing the unregluarized version of the loss function of NMMR:

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

we either had to compute the kernel matrix K for all points in the training set once, or dynamically calculate this matrix per batch. The latter approach would require many, many more calculations since we'd be repeating this process every batch and every epoch.

Our solution relied on batching the V-statistic and U-statistic. Recall we can write the V-statistic as:

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

We can vectorize this double sum as a series of vector and matrix multiplciations:

$$(y_1 - h(a, w_1, x_1), \ldots, y_n - h(a, w_n, x_n)) \begin{pmatrix} k_{1,1} & \ldots & k_{1,n} \\ \vdots & \ddots & \vdots \\ k_{n,1} & \ldots & k_{n,n} \end{pmatrix} \begin{pmatrix} y_1 - h(a, w_1, x_1) \\ \ldots \\ y_n - h(a, w_n, x_n) \end{pmatrix}$$

and for the U-statistic, we can simply set the main diagonal of K to be 0 to eliminate $i = j$ terms from this double sum.

However, calculating K for large datasets in a tensor-friendly manner resulted in enormous GPU allocation requests, on the order of 400GBs in the dSprite experiments. We implemented a batched version of the matrix multiplication above to circumvent this issue:

```
def NMMR_loss_batched(model_output, target, kernel_inputs,
    kernel_function, batch_size: int, loss_name: str):
    residual = target - model_output
    n = residual.shape[0]

    loss = 0
    for i in range(0, n, batch_size):
        # return the i-th to i+batch_size rows of K
        partial_kernel_matrix = calculate_kernel_matrix_batched(
            kernel_inputs,
            (i, i+batch_size),
            kernel_function)
        if loss_name == "V_statistic":
            factor = n ** 2
        if loss_name == "U_statistic":
            factor = n * (n-1)
            # zero out the main diagonal of the full matrix
            for row_idx in range(partial_kernel_matrix.shape[0]):
                partial_kernel_matrix[row_idx, row_idx+i] = 0
        # partial matrix multiplication
        temp_loss = residual[i:(i+batch_size)].T @
            partial_kernel_matrix @ residual / factor
        loss += temp_loss[0, 0]
    return loss
```

E Noise Figures

In the Demand noise experiment, we tested each method's ability to estimate $\mathbb{E}[Y^a]$ given varying levels of noise in the proxies Z and W. Specifically, we varied the variance on the Gaussian noise terms ϵ_1, ϵ_2, and ϵ_3 from the Demand structural equations described in Appendix C.1. We will refer to these variances as $\sigma^2_{Z_1}$, $\sigma^2_{Z_2}$, and σ^2_W, respectively, and we set $\sigma^2_{Z_1} = \sigma^2_{Z_2}$ throughout. We refer to the pair $(\sigma^2_{Z_1}, \sigma^2_{Z_2})$ as "Z noise" and σ^2_W as "W noise". In Xu et al. [7], these variances were all equal to 1. We evaluated each method†on 5000 samples from the Demand data generating process with the following Z and W noise levels:

$$\sigma^2_{Z_1}, \sigma^2_{Z_2} \in \{0, 0.01, 0.1, 0.5, 1, 2, 4, 8, 16\}$$
$$\sigma^2_W \in \{0, 0.01, 0.1, 0.5, 1, 16, 64, 150\}$$

In total there are 9 x 8 = 72 noise levels. From Appendix Figure S2, Panels A and B, we can see that Z_1 and Z_2 lie approximately within the interval $[-4, 4]$, whereas W lies approximately in the interval $[20, 45]$. Accordingly, we chose the maximum value of Z and W noise to be the square of half of the variable's range. So for W, half of this range is approximately 12.5 units, therefore the maximum value for σ^2_W is $12.5^2 \approx 150$. Similarly, half of the range of Z is 4, and so the maximum value of both $\sigma^2_{Z_1}$ and $\sigma^2_{Z_2}$ is $4^2 = 16$. This maximum level of noise is capable of completely removing any information on U contained in Z and W.

Intuitively, as the noise on Z and W is increased, they become less informative proxies for U. We would expect that greater noise levels will degrade each method's performance in terms of c-MSE. This experiment provides a way of evaluating how efficient each method is at recovering information about U, given increasingly corrupted proxies. It also provides some initial insights into the relative importance of each proxy, Z and W.

Figure S10 contains a 72-window grid plot with 1 window for each combination of Z and W noise and Figures S11-S18 show each method's individual potential outcome prediction curves at each of the 72 noise levels. We can see that NMMR-V is notably more robust to noise than NMMR-U, and also appears to be the most efficient method at higher noise levels. NMMR also consistently outperforms Naive Net, supporting the utility of the U- and V-statistic loss functions. However, we also note that kernel-based methods, such as KPV and PMMR, rank increasingly well with increased noise level, likely due to their lack of data adaptivity. We also observe that less data adaptive methods are less prone to large errors. Finally, we see a surprising trend that as the Z noise is increased, several methods achieve lower c-MSE. We believe this stems from the fact that W is a more informative proxy, so it is possible that noising Z aids methods in relying more strongly on the better proxy for U.

†We used the optimal hyperparameters for NMMR-U, NMMR-V and Naive Net found through tuning, as described in Appendix B.

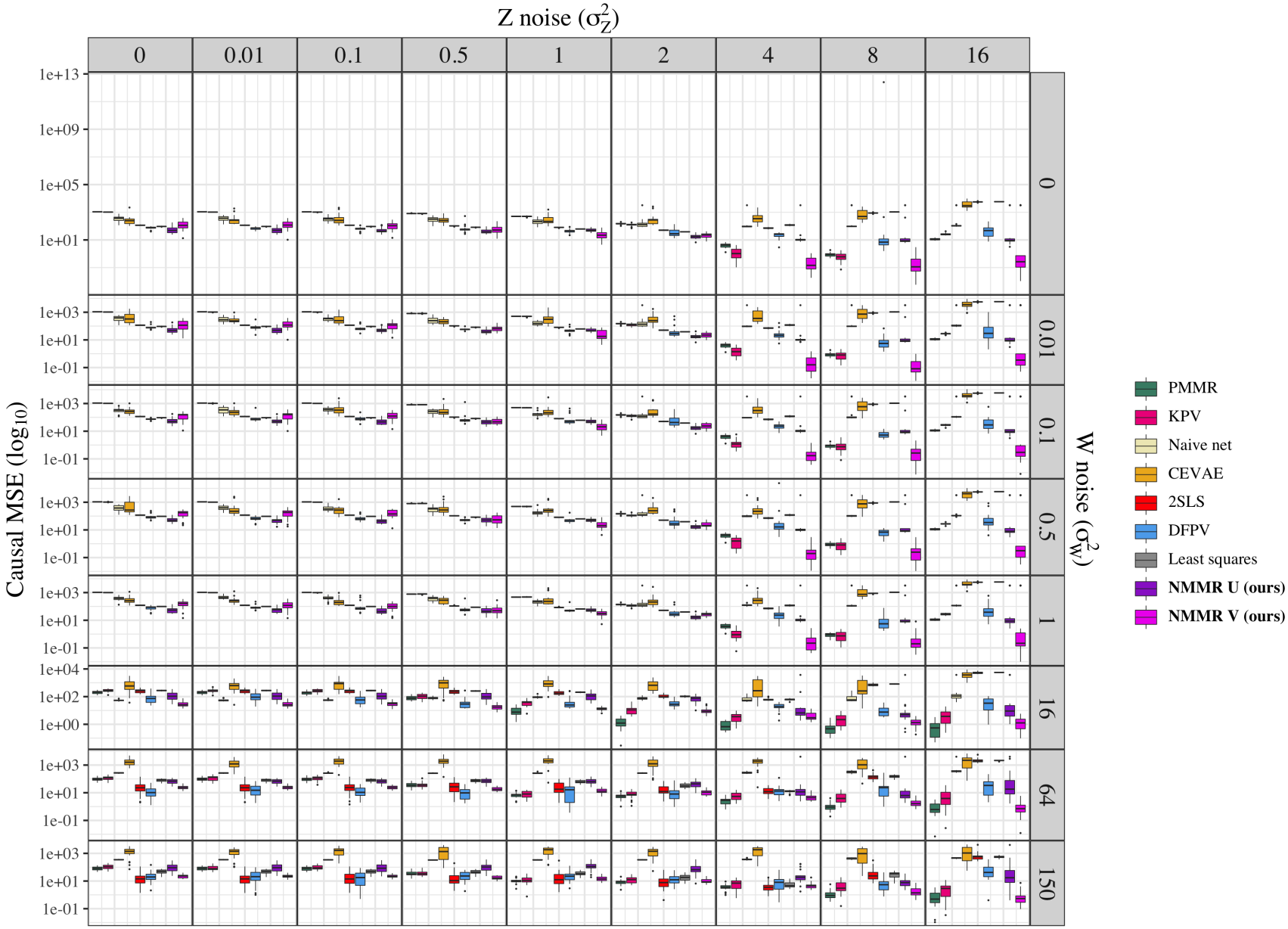


Figure S10: Causal MSE (c-MSE) of NMMR and baseline methods across 72 noise levels in the Demand experiment. Each method was replicated 20 times and evaluated on the same 10 test values of $\mathbb{E}[Y^a]$ each replicate. Each individual box plot represents 20 values of c-MSE.

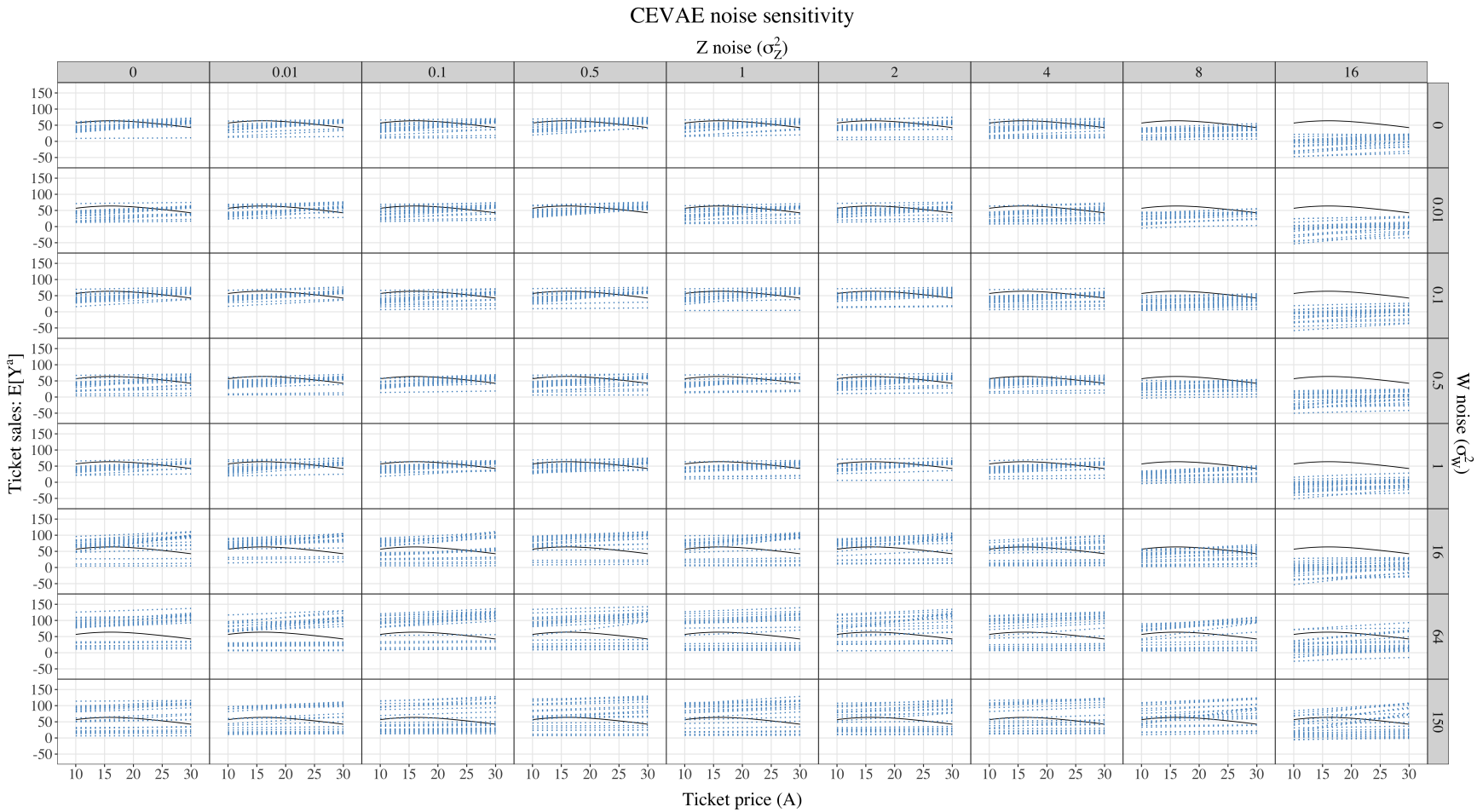


Figure S11: Predicted potential outcome curves for 20 replicates of CEVAE. Black curve is the ground truth $\mathbb{E}[Y^a]$.

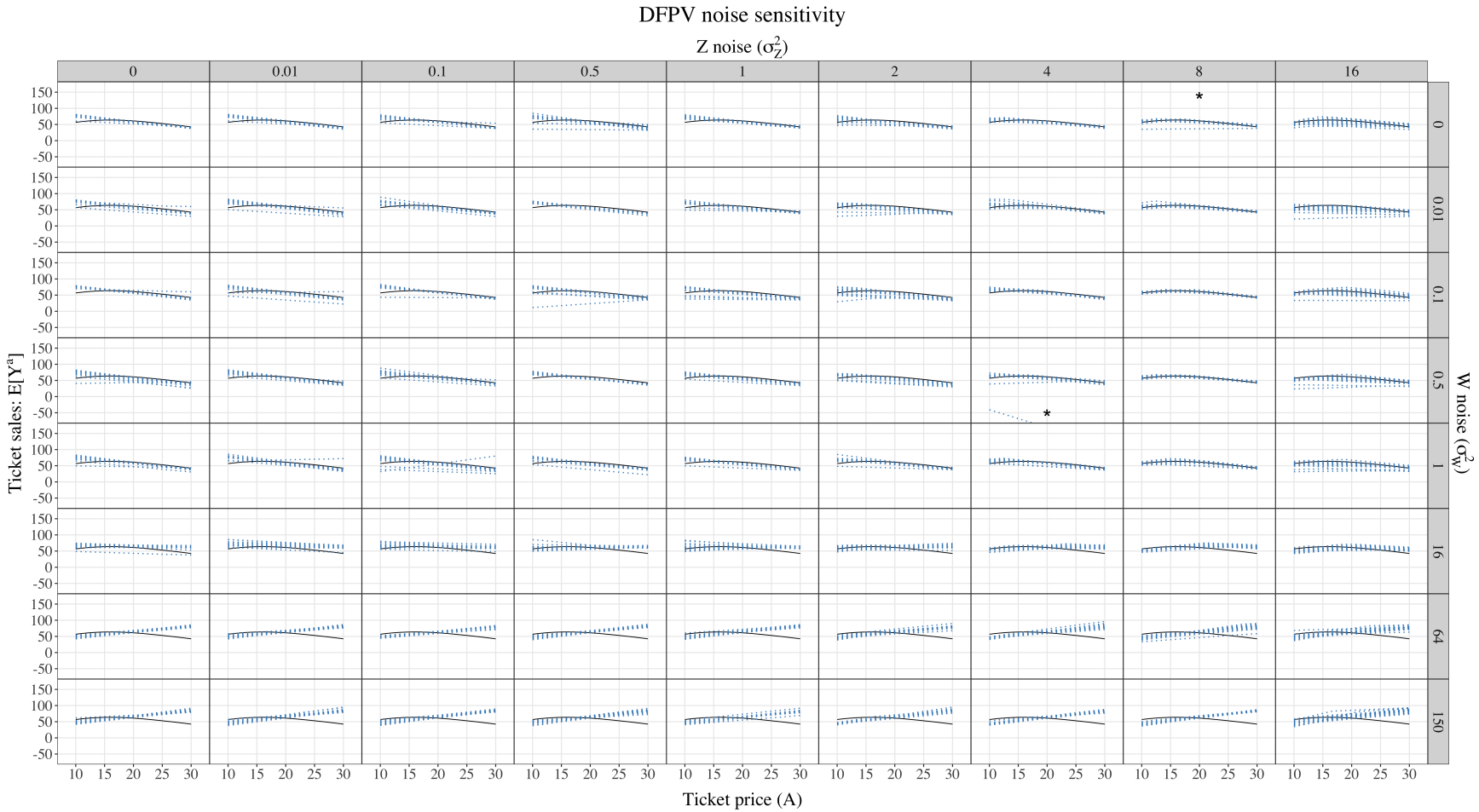


Figure S12: Predicted potential outcome curves for 20 replicates of DFPV. Black curve is the ground truth $\mathbb{E}[Y^a]$. Asterisks each indicate a replicate that lies outside of the plot's range.

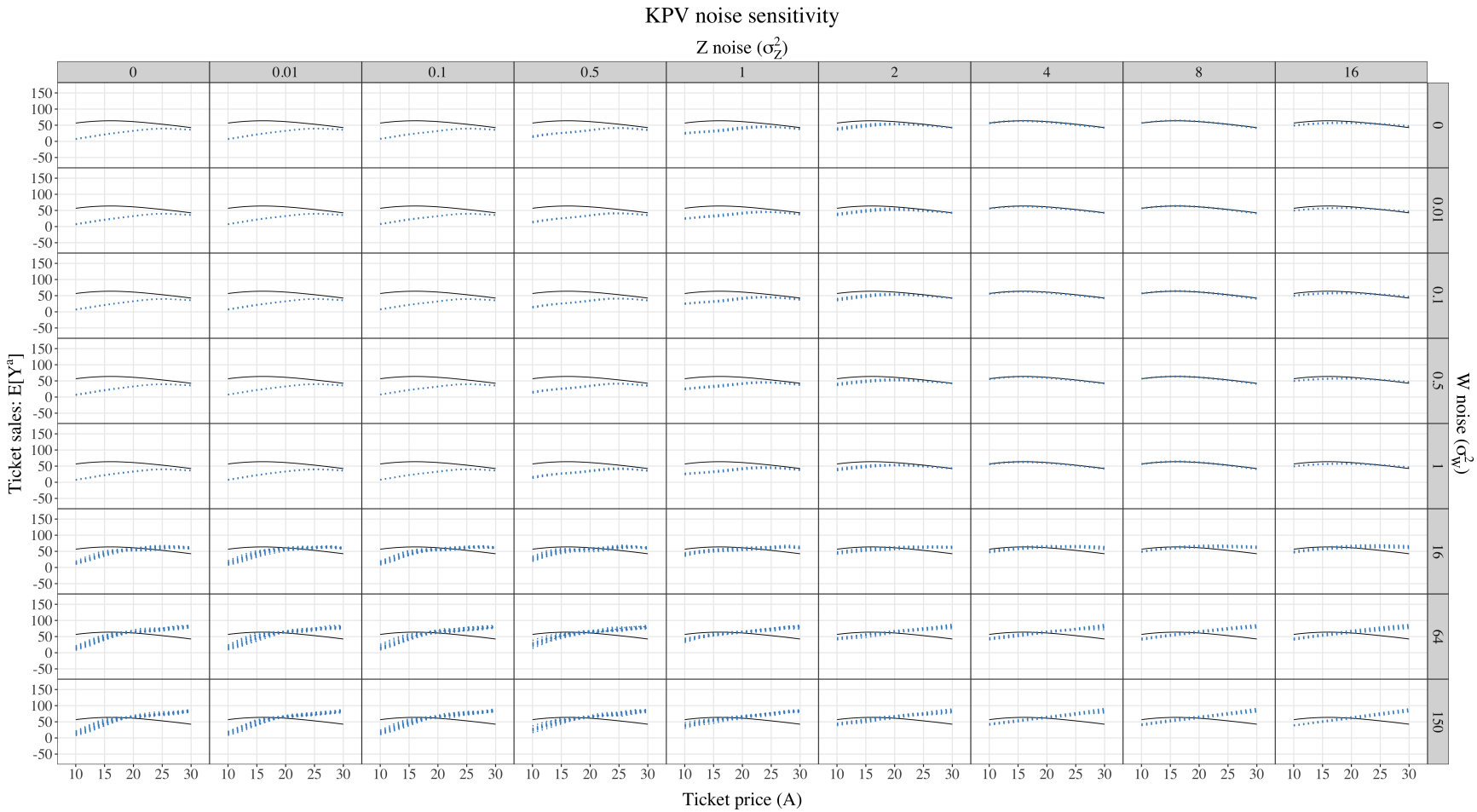


Figure S13: Predicted potential outcome curves for 20 replicates of KPV. Black curve is the ground truth $\mathbb{E}[Y^a]$.

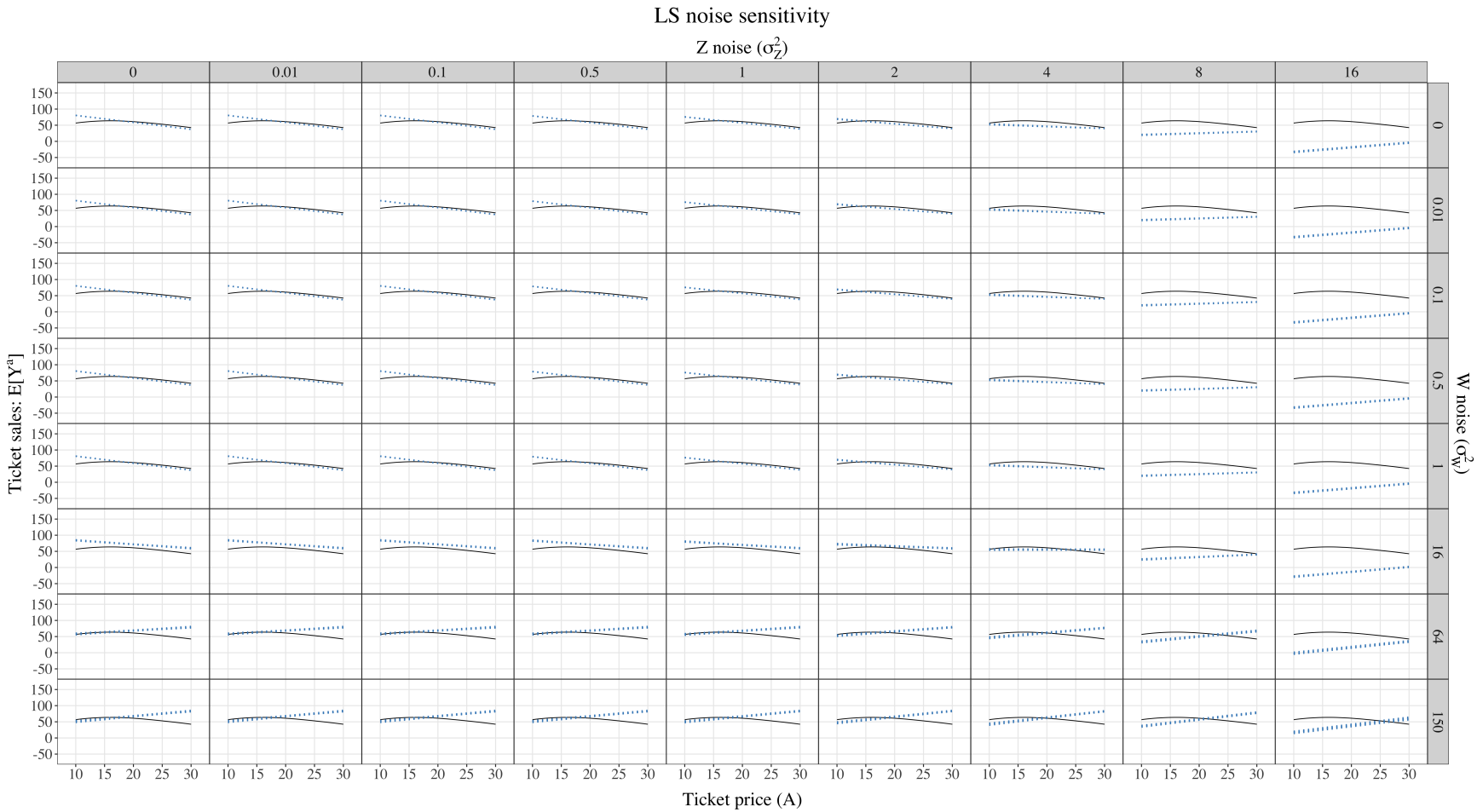


Figure S14: Predicted potential outcome curves for 20 replicates of Least Squares. Black curve is the ground truth $\mathbb{E}[Y^a]$.

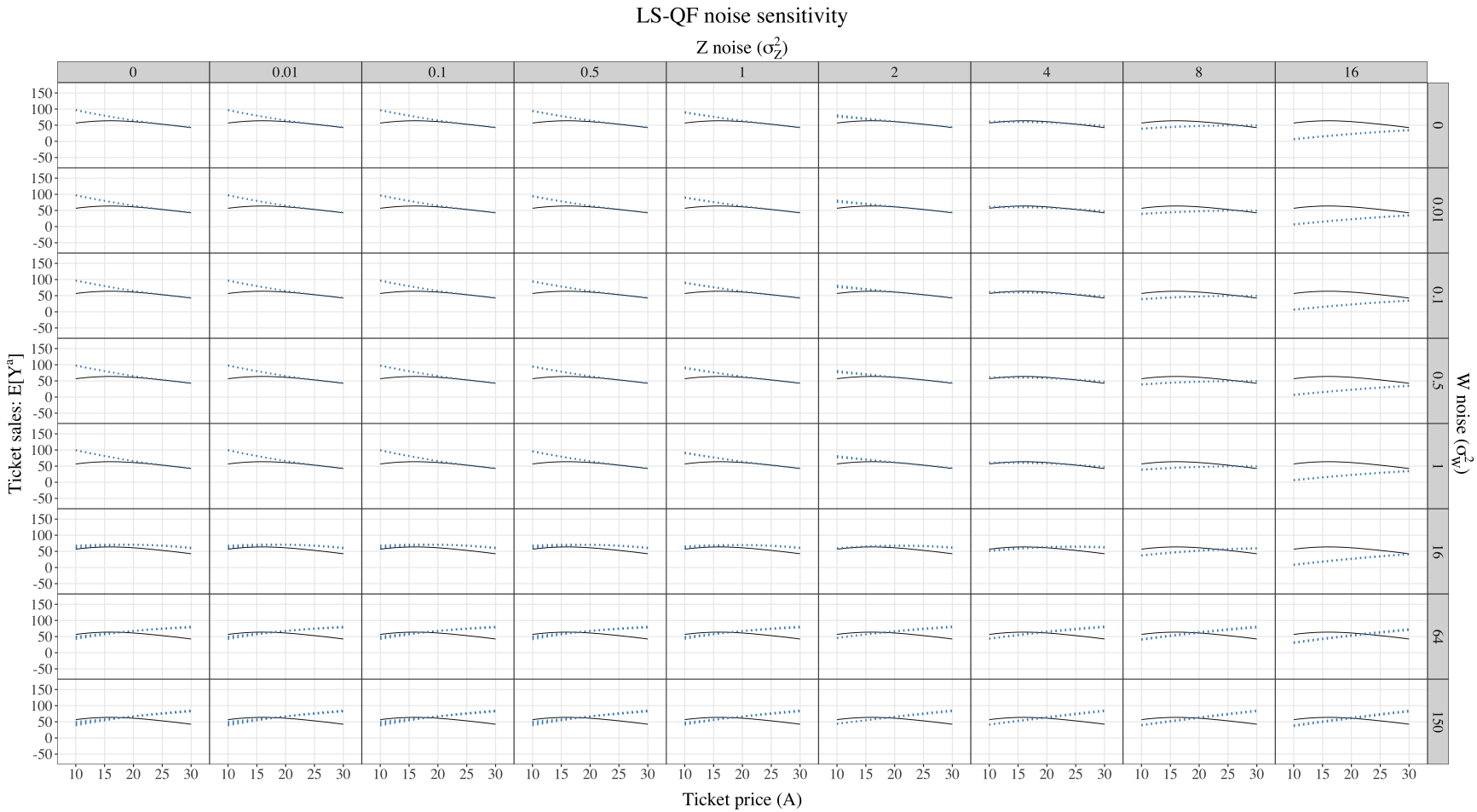


Figure S15: Predicted potential outcome curves for 20 replicates of LS-QF. Black curve is the ground truth $\mathbb{E}[Y^a]$.

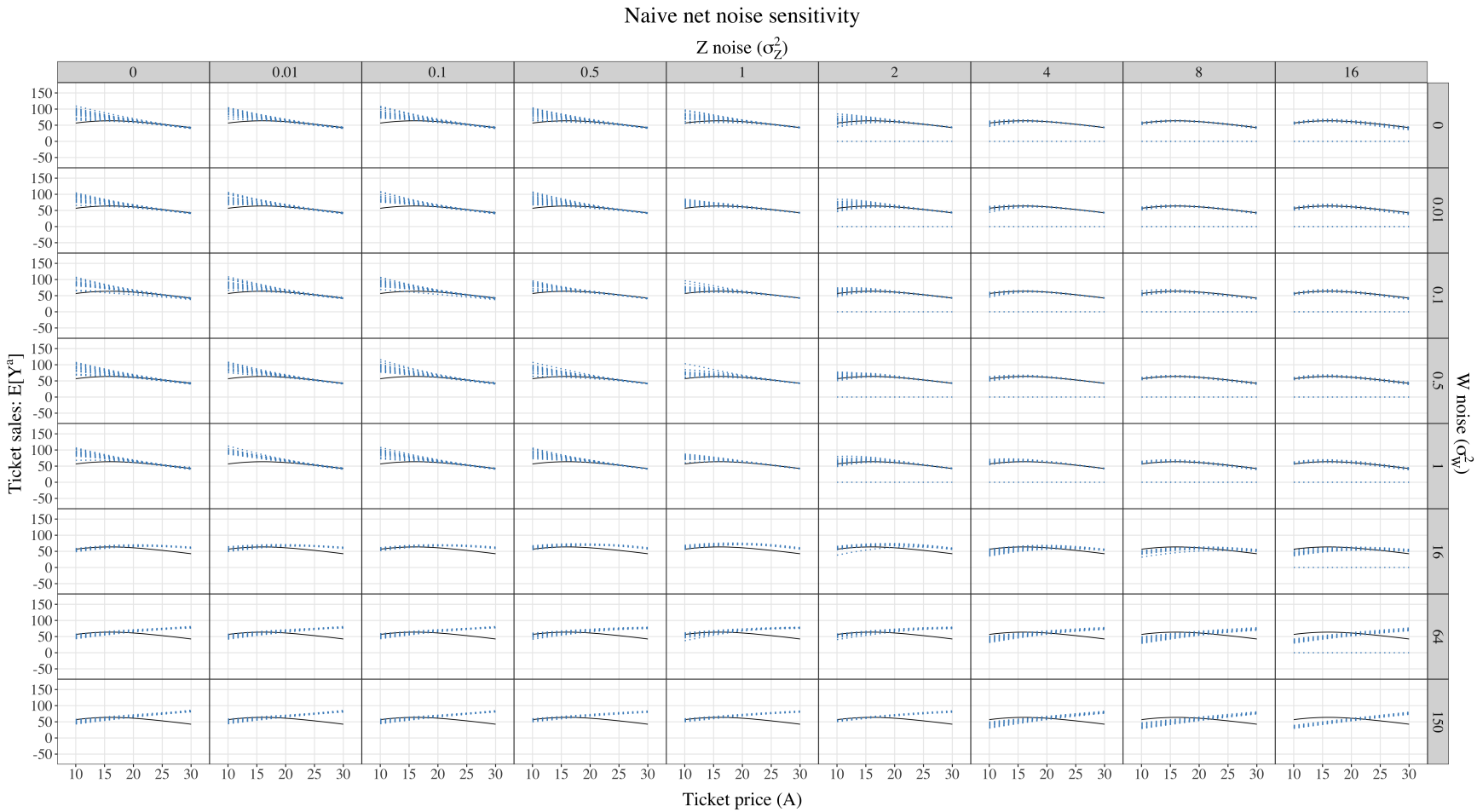


Figure S16: Predicted potential outcome curves for 20 replicates of Naive Net. Black curve is the ground truth $\mathbb{E}[Y^a]$.

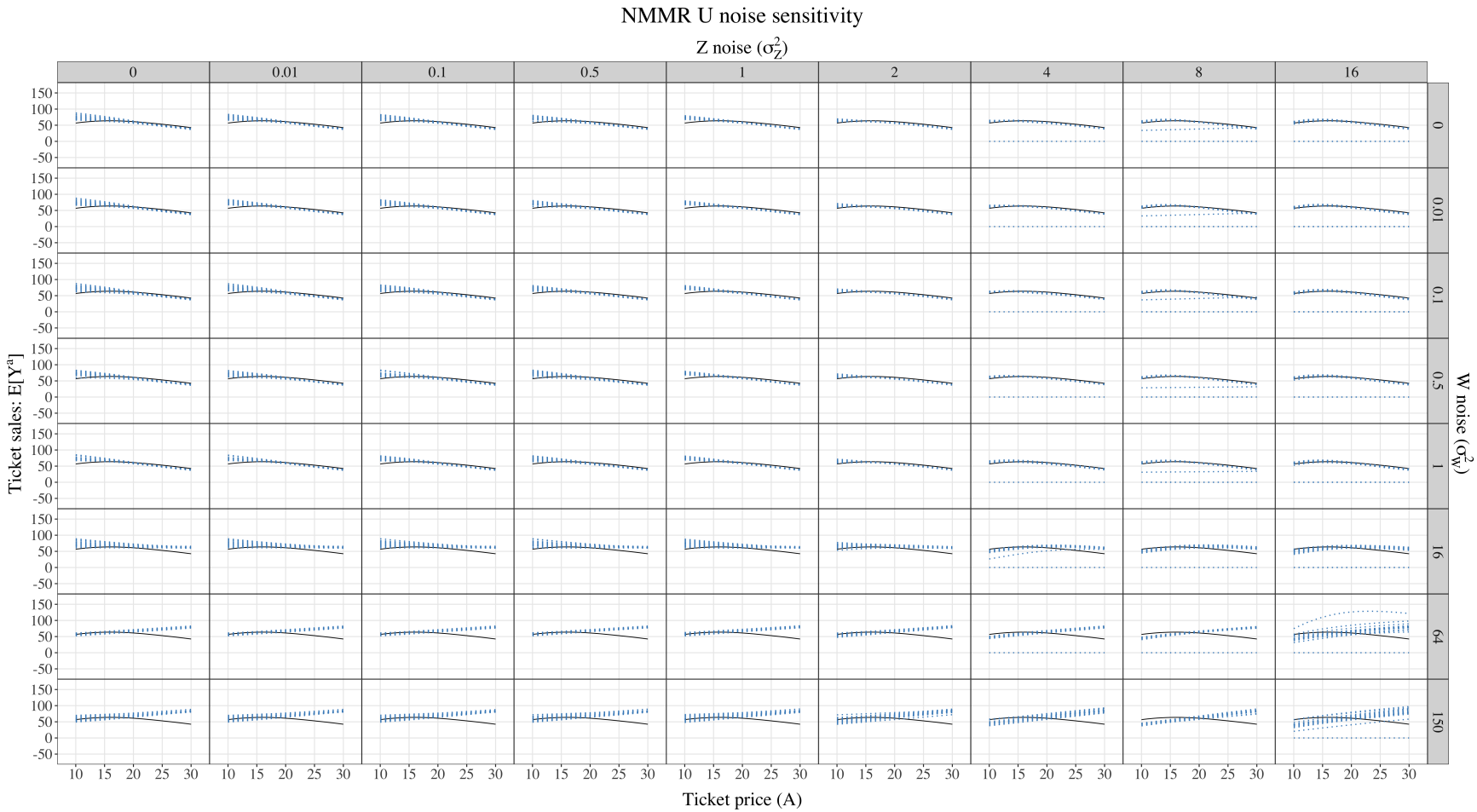


Figure S17: Predicted potential outcome curves for 20 replicates of NMMR-U. Black curve is the ground truth $\mathbb{E}[Y^a]$.

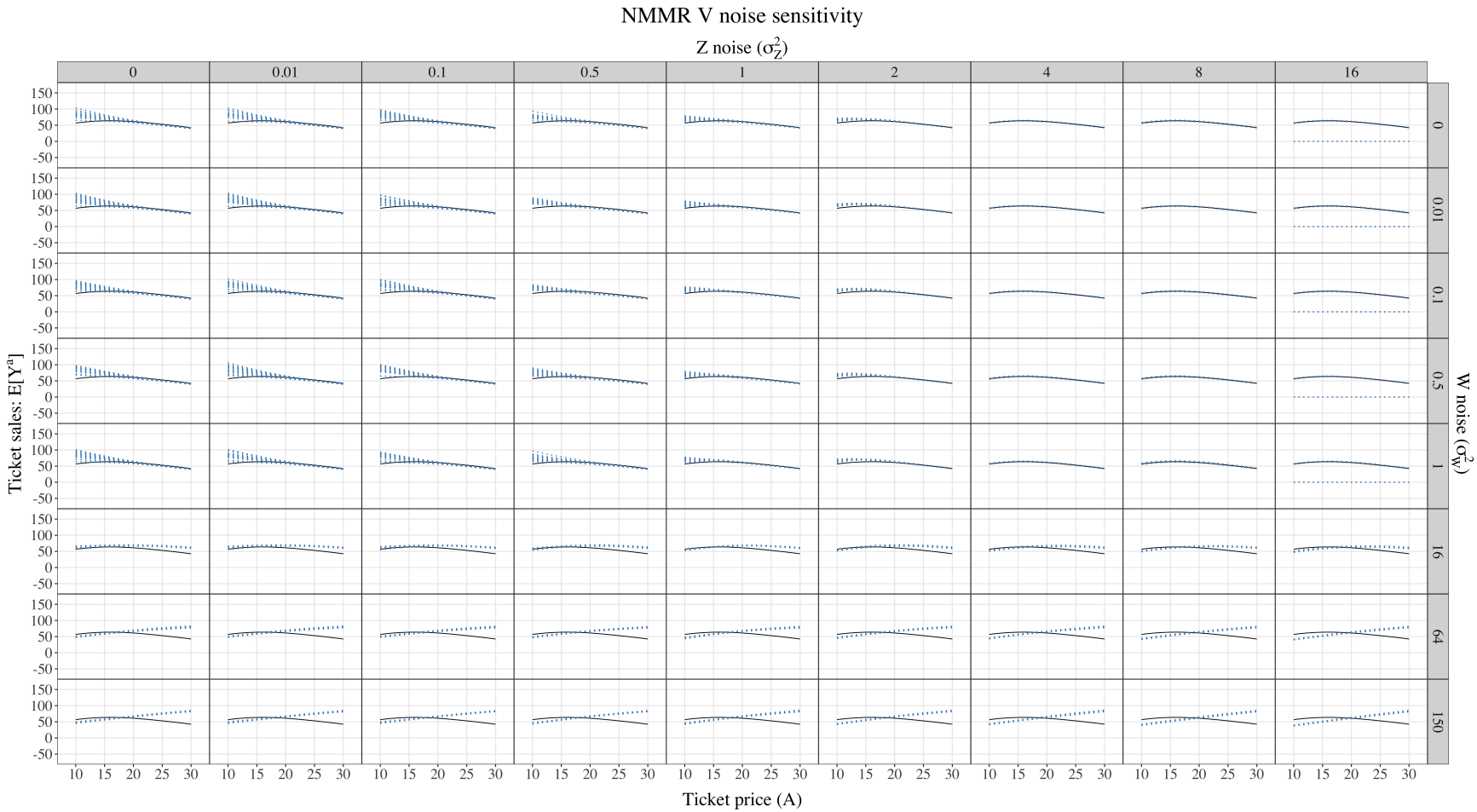


Figure S18: Predicted potential outcome curves for 20 replicates of NMMR-V. Black curve is the ground truth $\mathbb{E}[Y^a]$.